Journal of Official Statistics
vol. 34, i. 1 (marzo 2018)

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 1–25, http://dx.doi.org/10.1515/JOS-2018-0001

Household Classification Using Smart Meter Data

Paula Carroll[1], *Tadhg Murphy*[1], *Michael Hanley*[1], *Daniel Dempsey*[1], *and John Dunne*[2]

This article describes a project conducted in conjunction with the Central Statistics Office of Ireland in response to a planned national rollout of smart electricity metering in Ireland. We investigate how this new data source might be used for the purpose of official statistics production. This study specifically looks at the question of determining household composition from electricity smart meter data using both Neural Networks (a supervised machine learning approach) and Elastic Net Logistic regression. An overview of both classification techniques is given. Results for both approaches are presented with analysis. We find that the smart meter data alone is limited in its capability to distinguish between household categories but that it does provide some useful insights.

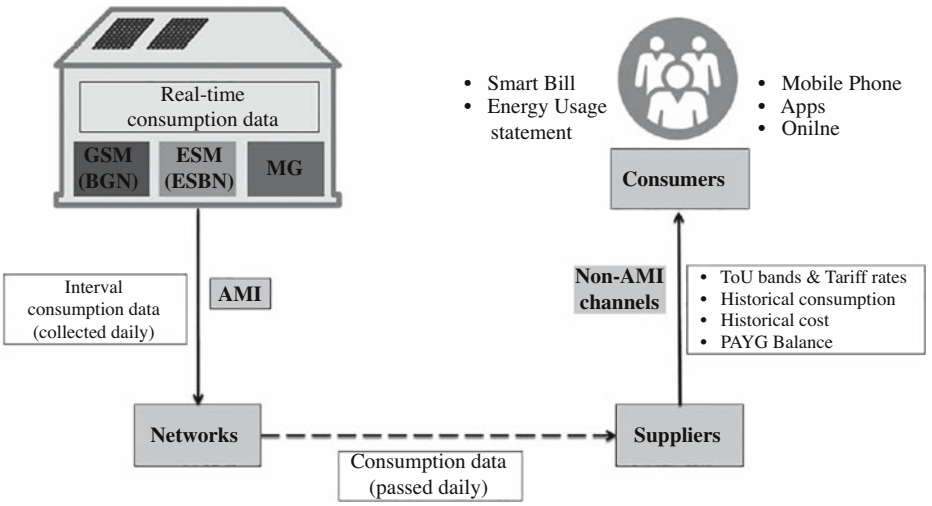
Key words: Neural network; elastic net logistic regression; classification system; household composition; smart meter data.

1. Introduction

Smart Meters (SM) in the residential sector are seen as a key factor in the success of EU targets for reduction in greenhouse gases and increases in the use of renewable energy (European Commission 2014). An SM system has an electronic meter which sends electricity load data to and receives price data from the service provider. The Irish Commission for Energy Regulation (CER) initiated the National Smart Metering Programme (NSMP) in 2007 and Customer Behaviour Trials (CBTs) took place during 2009 and 2010 to assess the performance of SMs and their impact on consumer behaviour. The purpose of the CBTs was to gauge customer response to price incentives. The anonymised data gathered during the trial are available for research purposes (CER 2012).

It is anticipated that a full rollout of SMs in Ireland will commence in 2019. Each consumer will have an individual meter to enable each residential household better manage its electricity usage. More recently, CER have announced the high level design decisions for the NSMP. Figure 1 shows an overview of the proposed architecture, CER (2014). We see that consumption and price data will be exchanged but no information about the building, appliances, or residents will be provided.

[1] Centre for Business Analytics, School of Business, University College Dublin, Belfield, Dublin 4, Dublin, Ireland. Emails: paula.carroll@ucd.ie, tadhg.murphy.1@ucdconnect.ie, michael.hanley@ucdconnect.ie, and daniel.dempsey@ucdconnect.ie
[2] Central Statistics Office, Skehard Road, Mahon, Cork, Ireland. Email: john.dunne@cso.ie
Acknowledgments: We would like to thank the anonymous reviewers and Associate Editor for their constructive suggestions on early drafts of this manuscript.



GSM – Gas Smart Meter provided by Bord Gais Metworks (BGN)
ESM – Electricity Smart Meter provided by Electricity Supply Board Networks (ESBN)
MG – Mico-Generation meter provided by parties yet to be determined
AMI – Automated Meter Infratructure
ToU – Time of Use
PAYG – Pay As Yog Go (enhanced form of PrePayment)

Fig. 1. NSMP High Level Design CER (2014).

Smart Meter Data (SMD) opens up new opportunities for researchers, businesses, and public sector organisations. In particular, the potential role of SMD in the production of official statistics is of interest to national statistical institutes and is the focus of this article. New data sources such as SMD have the potential to provide valuable information and insights about not only energy consumption but also household consumption and possibly, the subject of this study; household composition.

Like most countries, the Central Statistics Office of Ireland (CSO) is exploring ways to modernise how it calculates population estimates, (Dunne 2015). The focus of this research is an exploration of SMD to estimate household composition. Household composition is a classification of households by size and relationship type between the household members. It is currently established in Ireland in a costly census every five years. This involves the distribution of census forms to every household in the state and the subsequent collection of these forms. The cost of the 2011 Census was EUR 55 million. The SMD gathered during the CBT trial are used to attempt to answer our following research question:

Can household composition be estimated from analysis of SM electricity usage?

We evaluate two techniques to classify households; Neural Networks and Elastic Net Logistic Regression. While existing CSO household composition categories cannot be readily identified, useful insights can be gained from SMD analysis. In particular, the models are useful in identifying households of single persons. The model performance worsens as the number of persons in a household increases.

The remainder of this article is structured as follows: Section 2 outlines the challenges and opportunities for national statistics institutes in new data sources such as SMD; Section 3 describes the classification methods and data issues. Section 4 gives results and analysis; Sections 5 and 6 include a discussion and conclusions of the work.

2. Challenges and Opportunities for National Statistical Institutes (NSIs)

The functions of the CSO are spread across many areas with responsibility for the collection, compilation, extraction, and dissemination for statistical purposes of information relating to economic, social, and general activities and conditions in Ireland. Like most NSIs, the CSO is exploring ways to modernise how it operates and are trying to increase and improve the services they offer despite the growing costs of data collection and processing, and ever more challenging fiscal environments. A survey of the evolving National Data Infrastructure in Ireland is given in Dunne (2015). A strategy which focuses on efficient public administration rather than purely the production of official statistics is envisaged. This may be accomplished through the linking of administrative data registers covering persons, business and property. Currently, projections of the population on an annual basis up to 2046 are based on projection forward from the 2011 Census base under a chosen set of assumptions governing births, deaths, and net migration. Dunne (2015) describes some emerging opportunities for future censuses that may exploit administrative data registers either in conjunction with or as a substitute for primary data collection.

Seyb et al. (2013) describe the strategy implemented by Statistics New Zealand to improve and standardise processes in official statistics production. One goal in their change programme is to maximise the use of administrative data as a source wherever possible, with surveys filling gaps in information needs. This is a reversal of traditional survey-based data gathering strategies. Seyb et al. (2013) describe how value can be extracted from a specific administrative data source where the data is well formatted and well defined. They give an example where tax data reference numbers used by Inland Revenue agencies are already mapped to business registers, so matching and coverage issues are easy to resolve. The data items in that instance are well defined financial variables.

Other administrative or new data sources may not be as amenable to adaptation for NSI purposes. The focus of our work is on SMD as a potential new data source for the CSO. Every household uses electricity but the data derives from electricity markets and was not intended for NSI usage. In this article we outline the first steps toward harvesting value from SMD data.

2.1. Evaluating SMD Data for Official Statics Production

The Irish CBT SMD has been explored to identify factors influencing domestic energy consumption. Dwelling characteristics (such as dwelling type, age, and electrical appliances) and occupant characteristics (such as household income, age of household members, household composition) have been used to explain energy consumption. See for example McLoughlin et al. (2012). The reverse, using consumption to predict occupant characteristics, has received little attention (Newing 2016). It should be noted that dwelling and socioeconomic information about the CBT participants were used by

McLoughlin et al. (2012). Such information was available in the CBT but will not be available through the smart meter itself (Van Gerwen et al. 2006 and CER 2014).

2.2. The CBT Data

As noted, SMD data derives from electricity markets and the focus of the CBT trial was consumer responsiveness to pricing structures. However, the CBT data gives an indication of what SMD looks like, its volume and velocity and allows us to attempt to answer our research question.

The CER used a stratified random sampling framework to invite consumers to participate in the CBT. This ensured the sample was broadly representative of the population in terms of household size and other socioeconomic indicators. Over 5,000 consumers were initially recruited, further details are given in Subsection 4.1. Each consumer represents a household, that is, a number of persons sharing a single residential unit.

An incentive of EUR 25 for completing a pre- and posttrial survey was offered. An additional incentive for participation was the possibility of lower electricity bills during the trial depending on the consumer's response to the pricing schemes. The surveys were conducted by computer assisted telephone interviewing and focused on participants' views on attitudes to electricity usage and expectations of the trial, the dwelling, and electrical appliances. Questions on demographics and social relationships between household members were limited as they were not the focus of the CBT study.

The CBT recorded a meter reading of the electricity usage of participating consumers at half hourly intervals over the duration of the trial. Each household meter produced 269 MB of such time series usage data during the trial. There are over 1.6 million households in Ireland. This gives an indication of the type and volume of data associated with electricity consumption per household that will be available after national SM rollout.

The volume of such data presents a significant challenge for NSIs such as the CSO which does not have a history of dealing with high volume data other than its own primary (well structured) sources. The infrastructure required to deal with such data volumes has not been investigated in this study. This study focuses instead on a data processing pipeline and analytics techniques to produce meaningful insights on household composition from a SMD data stream.

3. Classification Techniques

The goal of classification in this article is to assign a household composition category to a household based on its SM electricity usage. The parsimony principle tells us that classification models with a small number of Explanatory Variables (EVs) are preferable. In this article, the Dependent Variable (DV) is the household classification and the EVs are drawn from the SMD data. Further EVs relating to participants' dwelling type and the type of electrical appliances used, are available in the CBT surveys. However, such information will not be available with the SMD after rollout, only the electricity usage data will be available. So, only EVs from the SMD data are used in this proof of concept study.

We use the CBT survey response on household composition to label the SMD. The labelled SMD data are processed through a data reduction pipeline to yield a set of EVs

suitable for model building. The data reduction process is described in Section 4. This labelled data allows us to use a supervised machine learning approach. We train and test a Neural Network to identify household composition based on SMD usage. These results are compared with those from a statistical model, namely Elastic Net Logistic Regression.

3.1. Regression

Regression is often used as a benchmark for classification tasks. Multiple linear regression models the linear relationship between DVs y and a set of EVs x. The general form is $y = \beta_0 + \sum \beta x$ where the coefficients β are calculated so as to minimise some loss function, such as the sum of squared error $\|y - \beta x\|^2$.

A regularisation term may be added to the loss minimisation objective function to achieve parsimony and reduce overfitting. Two popular approaches to regularisation are Ridge regression (Hoerl and Kennard 1988) and LASSO (Tibshirani 1996). Ridge regression adds a squared two-norm penalty term on the coefficients. It is used to reduce the variance inflation due to correlations in the explanatory variables. Least Absolute Shrinkage and Selection Operator (LASSO) adds a one-norm penalty term which has the effect of shrinking coefficients, possibly all the way to zero, thus performing what can be considered a continuous variable selection as opposed to discretely dropping variables outright. Elastic Net harnesses both Ridge and LASSO regularisation approaches by taking a linear combination of both norm penalties (Zou and Hastie 2005). The Elastic Net is fit by minimising $\|y - \beta x\|^2 + \lambda[\![\alpha\|\beta\|_1 + (1 - \alpha)\|\beta\|_2^2]\!]$.

The $\|\beta\|_1$ term is the LASSO penalty. The $\|\beta\|_2^2$ is the ridge penalty. The λ parameter is nonnegative and controls the 'strength' of the regularisation. A larger value of λ corresponds to greater variance reduction in the coefficient estimates but induces stronger bias. A value of $\lambda = 0$ corresponds to standard least squares regression. The α parameter takes values between 0 and 1 and controls the weight of the penalties. An $\alpha > 0.5$ puts more weight on the variable selection properties of the LASSO, while $\alpha < 0.5$ puts more weight on the correlation regularisation properties of the ridge.

Since linear regression models are linear by their nature, they are not well suited where the relationship between the inputs and outputs is not well defined or linear as is the case for electricity consumption and household composition. Generalised Linear Models (GLM) such as logistic regression can be used to overcome this limitation and to attempt to improve the model fit. GLMs extend the ideas behind linear regression. The dependant variables arise from the exponential family and are related to the EVs by a link function $f, f\left(E\left[y\right]\right) = \beta_0 + \sum \beta x$. The logit function can be used as the link function to predict categorical variables in a logistic regression model. This allows binary and multinomial classification, where $\text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$ is the log odds. This forces the output to be a value between 0 and 1 which can be interpreted as a probability that the outcome belongs in a certain class.

The regression coefficients are usually estimated using maximum likelihood estimation. Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximise the likelihood function, so an iterative process such as Newton's method is used instead.

The Elastic Net approach can also be used to reduce overfitting of the logistic regression model.

3.2. Machine Learning and Neural Networks

Standard statistical techniques are based on assumptions that the data items have been sampled independently and from the same distribution. Machine Learning (ML) offers several techniques for intelligent data analysis and knowledge generation through generalisation where such assumptions may not hold (Hand 1998). Techniques include: association rules, decision trees, inductive logic programming, support vector machines, clustering, Bayesian networks, and (artificial) neural networks (A)NNs. NNs are discussed in detail here, a review of the other techniques is beyond the scope of this article. NNs are often preferred when large noisy training samples are available and the relationships may be nonlinear. Disadvantages of NNs include their "black box" nature and the empirical nature of the model development.

The aim of a ML model is similar to that of a regression model. It aims to model how the set of inputs (called *features* in ML parlance) relate to the set of outputs. However, the approach to creating and fitting the model differs from regression. The ML model is learned from a training data set. In supervised learning, the outputs for the training data are known (labelled). A ML learning algorithm adapts the model in response to the training data to improve the fitting of the input/output relationship.

Perhaps the most important concept in ML is that of generalisation. The algorithm should produce sensible outputs for inputs that were not encountered during learning Marsland (2009). Overfitting occurs when a model fits only the training data, meaning that it is not a general function approximation. It has instead begun to learn the noise associated with that specific training data set. To ensure that overfitting does not occur, the data is usually split 60:20:20 into training, validation and testing sets. The learned system is evaluated on the validation data set to assess the ML model fit before being used on unseen test data.

ML can be used to perform classification. We assign the input(s) to discrete output categories. Testing is performed to evaluate the model in terms of the performance of its classification when it is given new data without class labels. The actual class labels of each input are compared with those assigned by the algorithm. *Accuracy* is defined as the percentage of correct matches, that is, the number of correct (or true) household classifications in our case divided by the total number of tests:

$$Accuracy = \frac{Number\ correct\ classifications}{Number\ of\ tests}.$$

The accuracy metric can be misleading when the data are dominated by a high number of entries from a single class. This issue is explored in more detail in Subsection 3.3. A classifier that simply predicts the dominant class will have high accuracy but could not be regarded as a good classifier. Other commonly used error metrics are discussed in more detail in Subsection 4.3.

NNs are ML algorithms modelled on the function and topology of the human brain. NNs have successful applications in diverse areas from credit card fraud detection (Patidar and Sharma 2011) to forestry management (Hickey et al. 2015) and to energy consumption modelling (Aydinalp et al. 2002). Aydinalp et al. (2002) favoured NNs over statistical models due to the simplicity of NN development and the accuracy of the estimate. They

found that NNs were capable of modelling nonlinear electricity consumption relationships outperforming statistical approaches.

In the brain, nerve cells called *neurons* function as simple processing devices. Neurons can be described as simple mathematical functions. A general form for a single neuron is $y = g(w_0 + \sum wx)$ where g is called the transfer function, often a sigmoid or hyperbolic tangent function. The w are weights which are analogous to the coefficients in a regression model and w_0, known as the bias, is analogous to the regression intercept coefficient.

Figure 2 shows a representation of a simple NN with EVs x_i, a single neuron and a single output. NNs consist of an array of neurons that form a connected network (Hopfield 1984; Zhang and Zhang 1999). A Multi-Layer Perceptron (MLP) is a feed-forward NN consisting of an input layer, one or more hidden layers and an output layer.

An iterative approach called the error Back-Propagation (BP) algorithm is used in a MLP to estimate the weights during the training stage. BP consists of two passes through the network: a forward pass and a backward pass. The weights w are fixed in the forward pass. An input vector from the training data propagates through the entire network to produce a set of outputs. The difference between the produced output and target value is calculated as the error. The error is similar to the loss function in a regression model. On the backward pass, the weights w are adjusted according to some error-correction rule (such as a gradient descent function) to reduce the error. In this way, the NN response is moved closer to the desired output. Termination criteria for the iterative model fitting are used to stop the BP algorithm when improvements fall below a threshold. The NN weights are then finalised and the NN is ready for the test phase.

The empirical approach to NN model development means there is no guarantee that the final NN weights are the global optimal weights. They may reflect local optima. Nor is there a guarantee that the selected topology of hidden layers is optimal.

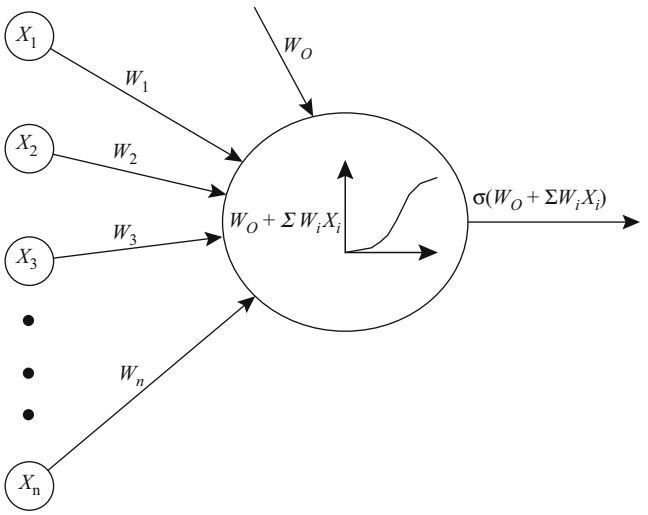


Fig. 2. Model of a simple NN.

3.3. Data Issues

In this section we identify some of the issues that arise for NSIs interested in exploring SM data. When dealing with large volumes of data, analysts have to decide on a data reduction scheme which adequately represents the data. The reduced data representation conveys most of the information while being easier to store and facilitates ease of computation. The choices are to aggregate the data (using representative measures), use samples of the data or to apply more advanced data reduction algorithms.

Recall that the EVs used as inputs to NN models are called *features*. As the number of dimensions (or possible EVs) in the search space increases, the amount of data needed to provide the algorithm with sufficient training examples increases rapidly. This can lead to an explosion in the amount of training data required as well as lengthening the training time for the NN. This issue is known as "the curse of dimensionality". Dimension reduction techniques can be used to address this issue, see, for example Han and Kamber (2006).

Characterisation of Time Series (TS) data such as SMD are discussed in detail in Liao (2005) and Wang et al. (2006). The lower level half-hour granularity per meter gives a better picture of what is happening in each household type than would be apparent by looking at aggregated SMD daily totals. The individual consumer's load profile may offer a unique fingerprint to aid classification. However, it is not desirable to work directly with raw data that are highly noisy. Instead, application dependant extracted features are used. In addition, a choice on the length of the TS is required. In general, larger sample sizes yield better population estimates with lower variability. However, in the case of TS data, longer series may actually increase variability due to any underlying trend. The choice of the TS window length is one of the many design parameters and is an open ML research question. Varying length sequences can be empirically evaluated and/or adaptive windowing (similar to the lag methods in ARIMA models) can be used to weight the contribution of varying length subsequences within the TS data.

In many applications the training data is unbalanced, that is, some categories are under or over represented. It is important to distinguish between imbalance in the training data sets and representativeness of the population of the sample. For example, in the classification of defective products at the end of an assembly line, the majority of products, perhaps 90%, are good since they meet the required standard. The remainder fail and are deemed defective. While the imbalance reflects the distribution of the items in the population, traditional feed forward NNs have difficulty learning from unbalanced data sets. NNs need to see an equal number of defective and good products during the training phase to learn how to distinguish them. Otherwise the NN prioritises the class seen in the majority of samples and treats the minority class as noise (Murphey et al. 2004). In the production example, the NN would classify all products as good and 90% of the time would be correct. This results in misleadingly high accuracy values for the model. Several error metrics are used to interpret the results in conjunction with the accuracy measure. In addition, resampling or oversampling can be used to address issues of unbalanced training data.

Lastly, concerns about the privacy of the individual arise with SMD (Molina-Markham 2010; McKenna 2012). These include that the SM signals may be intercepted for illegal purposes by third parties and that SMs allows surveillance of the individuals' usage rather

than simply tracking usage for billing purposes. The CSO adheres to the UN Fundamental Principles of Official Statistics and seeks to balance the public interest with concerns for privacy of the individual. The CSO was established statutorily under the Statistics Act, 1993 (Statistics Act 1993). This act includes articles on statistical confidentiality so the CSO is well positioned to explore new data sources such as the SMD.

4. Material and Methods

In this article, consumers are classified and assigned to a household category based on their electricity usage. NNs are selected to perform the ML household classification due to their ability to work with high volume noisy data and learn nonlinear relationships. Elastic net logistic regression is selected as a comparative GLM statistical approach. We reduce the individual consumer TS streams to sets of possible features (explanatory variables) and select the most useful subset of features. We evaluate the model performance over varying TS window lengths and compare results from both unbalanced and balanced training data sets.

4.1. Data Pipeline

Some information about the age of household members is available from the CBT pretrial survey. A limiting factor of the CBT survey from the household classification perspective is that detailed age information is given for the Head of Household only. The remaining members of the household are classified as either *under 15 years of age* or *15 years of age and older*. In addition, no information is given on family unit group, for example whether the household consists of a married or cohabiting couple or single parent with children etc. Existing CSO household composition categories distinguish family types, for example *cohabiting couple with children* or *husband and wife with children*. This categorisation is useful in social analysis and understanding changing demographics but we would not anticipate a difference in electricity usage of based on marital status. Indeed, in a sign of changing times, the marriage equality referendum passed in Ireland in 2015, may see the need for the development of new household categories such as *husband and husband with children*.

For the purpose of this smart meter study, an alternative simple household categorisation system was developed according to the numbers of adults and children as shown in Table 1. These 16 household categories were chosen as they match 95% of the existing CSO categories and represent the majority of the CBT data. An even simpler classification based on the number of persons per household was also considered.

Table 1 shows how representative the CBT sample is of the 1.6 million households in Ireland. The final two columns of Table 1 show the similarity of the CBT household distribution to the percentage of households by number of persons according to the 2011 Census (CSO 2011). The minor gap is households with eight or more persons as none participated in the CBT, this category accounts for five per cent of all households in Ireland.

A significant work component of this study was to convert the CBT SMD data to household classifications. Data preprocessing absorbed approximately 65% of the project man hours. Over 150 million data points of usage are included in the SMD trial data in

Table 1. Household category description.

Category	Adults	Children	Meter count	Post-processing count	Num persons	CBT distribution (%)	CSO distribution (%)
A	3	2	41	39	5	1	2
B	3	1	106	105	4	3	3
C	3	0	450	440	3	11	10
D	2	5	9	9	7	0	0
E	2	4	49	48	6	1	1
F	2	3	158	147	5	4	4
G	2	2	338	331	4	9	8
H	2	1	246	244	3	6	7
I	2	0	1,264	1,251	2	32	27
J	1	1	59	59	2	2	2
K	1	0	726	718	1	19	24
L	4	1	64	64	5	2	2
M	4	0	289	283	4	7	5
N	5	1	20	20	6	1	1
O	5	0	92	92	5	2	1
P	6	0	20	20	6	1	0
≥8			0	0	≥8	0	5
Total			3,931	3,870		100	100

multiple CSV files. Each SMD usage file consists of three columns corresponding to a unique household Meter ID, timestamp, and electricity consumed during 30 minute intervals in kWh. In order to allow a valid comparison, SMD from the six month benchmark period from July to December 2009 were considered. Price incentives were evaluated during the later months of the CBT trial. Some work on consumer behaviour and their responsiveness to tariff changes is described in Di Cosmo et al. (2012). Such work could be used to estimate the likely changes in household electricity usage patterns in response to tariff changes.

The data were prepared using the open-source package R (R Core Team 2013). Standard workplace laptops with 8 GB RAM were used for light data preprocessing tasks. Data for households who had not completed the survey were removed leaving 3,931 sets. Meters with missing data were also removed. The data reduction and model building was then carried out using R on the Stokes supercomputer with 7,680 GB of RAM at the Irish Centre for High-End Computing.

Bousquet and Elisseeff (2002) discuss the use of sensitivity analysis to evaluate changes in ML algorithm outcomes to changes in the training set. We were particularly interested in assessing the impact on the model performances of the window length of time series used as the training data. Five different time series window lengths ranging from one day to six months were chosen so that the sensitivity of the classifiers could be empirically assessed.

Feature values for the five different time series windows were calculated on the Stokes supercomputer in the data reduction step. The features are the EVs or inputs for the classification models, further details are given in Subsection 4.2. The prepared data per meter was then labelled with a household classification category. These five files

containing the feature values for the five different time series windows were then ready for use in creating and testing the classification models.

4.2. Data Reduction and Feature Selection

Table 2 shows a summary of the extracted features. Some are suggested in McLoughlin et al. (2012). Others are standard descriptive statistical measures typically used in NN time series modelling (Wang et al. 2006). The remaining features were identified from analysis of the diurnal usage patterns of individual household categories to spot distinctive features which may be unique to a household category. One such example is "morning peak". It was noted that households with children generally had a more pronounced morning peak.

Twenty one features were calculated for each meter to summarise each household's unique load profile over the five time series windows. Detailed descriptions are included in Appendix 1. The raw numeric input SMD data were standardised to between −1 and 1. Standardising is carried out to bring all variables into proportion with one another. Features that demonstrated multicollinearity with high inter-correlation coefficients were removed, leaving 18 input features or explanatory variables.

Outlier analysis was performed on the summarised data to remove any outlying households that might disrupt the performance of the classifiers, leaving 3,870 meters.

*Table 2. Model inputs*Indicates a feature that was not selected.*

Index	Feature (EV)	Short description
1	Mean*	Mean energy consumption
2	Max	Maximum energy consumption
3	ToU Max	Time of day at which maximum consumption occurs
4	TEC	Total energy consumption
5	MDM	Mean daily maximum energy consumption
6	Load factor	Ratio of daily mean to daily maximum energy consumption
7	Variance	How far the energy consumption is spread out
8	SD	Standard deviation from the mean
9	Range	Difference between highest and lowest energy
10	Interquartile range (IQR)	Measure of spread of middle half of data
11	Morning max	Maximum energy use in the morning
12	Morning peak	Height of the morning peak energy consumption
13	Morning range	Morning maximum minus minimum before 10 am
14	Weekday area	Area under the curve for weekday consumption
15	Weekday midpoint*	Area under the curve for weekday consumption divided by 2
16	Weekday centroid	Time of day at weekday midpoint
17	Weekday AM slope	Slope of the morning peak
18	Weekend area	Area under the curve for weekend consumption
19	Weekend midpoint*	Area under the curve for weekend consumption divided by 2
20	Weekend centroid	Time of day at weekend midpoint
21	Weekend AM slope	Slope of the morning peak

Individual data within the meters was not subjected to any outlier analysis, instead this was performed on the aggregated data for each meter. This approach ensured that potentially useful data within individual meters was not removed but that outlying households were removed before the data were input to the classifier. For example, increased usage on a cold day was not deemed outlying. The local outlier factor algorithm which is a density-based outlier detection approach was chosen for this task. It can be computationally expensive as the approach involves the calculation of k-nearest neighbours. Breunig et al. (2000) argue that this approach is more subtle than a simple binary outlier classification and allows the degree of closeness within a neighbourhood to be accounted for.

4.3. Model Development

Two classifier approaches were evaluated. The first was a binomial classifier asking a binary question; whether a particular meter belonged to a particular household category. Classifier output greater than 0.5 was labelled as true (yes). Classifier output less than 0.5 was labelled as false (no). The advantage of a binomial approach is that only a single output is required. It was expected that the classifier would be better able to partition the data set. The disadvantage was that the model had to be run separately for each household category and so involved extra data manipulation.

The second approach was a multinomial classifier asking which household category a meter belonged to. The output produced by the classifier is a vector of values between zero and one. These vector components are interpreted as probabilities that the meter belongs to the household categories. The household category with the highest probability is the most likely category to which the meter belongs. The advantage of the multinomial approach is that only one model is required and less manipulation of the data is needed. However, as the multinomial classifier has multiple outputs, it could potentially lead to a reduction in accuracy. Lower accuracy was anticipated as some overlap of electricity usage between classes was expected.

The "glmnet" package in R was used to implement the elastic net logistic regression models (Friedman et al. 2009). For all models α was set to 0.25. This puts more weight on the ridge penalty which averages correlated groups but still allows for some feature selection. Ten-fold cross validation was used to set λ based on the misclassification error rate. For each Elastic Net model, 70% of the data were used as a training set, the remaining 30% was used for testing the predictive power of the models. The "caret" package in R was used for splitting the data into training and test sets (Kuhn et al. 2014).

The R "nnet" NN package was used to build a single-hidden-layer NN by selecting the number of units in the hidden layer, the initial random weight, and the weight decay (Ripley and Venables 2011). The "neuralnet" package was also used as it allows a choice of training algorithms and the number of hidden layers (Fritsch et al. 2012). Training of the NNs was carried out by back propagation, resilient back propagation with backtracking, resilient back propagation without backtracking and a modified globally convergent approach. The input data for the NNs was split into three subsets in ratios 60:20:20 for training, validation, and testing. The training data set was sampled at random without replacement. From the remaining data set, 50% were sampled at random without replacement to create the validation set with the remaining meters forming the test set.

The performance and suitability of all models were assessed under the headings of accuracy, Sum of Squared Error (SSE), Root Mean Squared Error (RMSE), Sum of Cross Entropy (SCE), Coefficient of Variation and Pseudo R^2. Confusion matrices of actual (row) versus predicted (column) values in each class were also produced. A good classifier exhibits a diagonally dominant matrix. Further details of the error metrics are available in Appendix 2.

For the binomial NN models, the RMSE was computed at each iteration of the NN development for both the training set and the validation set, and the SCE was used for the multinomial model. The training of the network was stopped when the RMSE/SCE error using the validation set registered two consecutive increases. A value of two was chosen as stopping after one increase might be premature and the increase might only be a once-off result in a general trend of decreasing error. More than two consecutive increases was categorised as a trend of increasing error. This check found the point at which the training algorithm had started to overfit the data.

The sensitivity of the binomial and multinomial NN models on both unbalanced data and balanced training data and on the five TS windows described in Subsection 4.1 were evaluated. Computational results are presented in Section 5.

4.3.1. Unbalanced Training Data

As noted in Subsection 3.3, a balanced number of training samples is preferred for ML classification so that one category does not bias the prediction output. Table 1 highlights the imbalance in the CBT which is a concern for training the NN. The number of sample households consisting of two adults and no children (1,264) exceeds any other household type in the trial. It is not a concern for the representativeness of the CBT data.

We used stratified sampling to build the training set for the Elastic Net unbalanced models. We sampled 70% of each category instead of simply taking 70% of the entire data. The value of λ in the elastic net was chosen via ten-fold cross validation where the validation error is the misclassification rate. Separately, the data were split 60:20:20 into training, validation and testing sets for the NN. The models were then applied to the test sets and a full set of error metrics was calculated.

4.3.2. Balanced Training Data

Undersampling (He and Garcia 2009) was used in order to achieve balance in the training data. This technique, for the binomial models, is to only sample enough records from the majority class so that it equals the number of records in the minority class. This is more suited to situations where large data sets are being analysed as it has the advantage of reducing the training time by effectively reducing the size of the training set. Recall the 16 different household composition categories described in Table 1. It shows that following the preprocessing step, the number of households in each of these categories ranged from 9 to 1,251 with a total data set size of 3,870. For the household category with nine meters this meant that the number of entries in the *true* class was nine and the number of entries in the *false* class was 3,861. To perform undersampling on this category required sampling nine records from the majority *false* class of 3,861, meaning that the size of the data set for classification for this category was 18. This under sampling was repeated across each of the household categories to allow classification on balanced data.

For the multinomial model, an equal number of meters in each household category were sampled. As the minimum sample size was nine this required sampling nine from each of the remaining categories and training the classifier with nine instances of each category. This is too small for ML algorithms, so it was decided to only analyse categories where the number of records in the category was greater than 100 meters. This choice ensured that a minimum sample size of 20 was achieved for a 60-20-20 cross validation split when developing the NN model. Eight of the 16 categories met this selection criteria, namely B, C, F, G, H, I, K, and M. These categories accounted for 3,519 (91%) of the CBT meters after preprocessing and is representative of 86% of the population of 1.6 million households in Ireland.

5. Results

Figures 3 and 4 show examples of the weekday and weekend daily usage averaged over a six month period for household categories C and H. Category C is a household with three adults. Category H consists of two adults and one person aged under 15. Weekdays are shown as a solid line, weekends as dashed. These are typical of the diurnal usage pattern showing a peak corresponding to the start of the day, some activity at lunch time and a peak corresponding to preparation of an evening meal.

A box plot of the mean daily usage in Figure 5 highlights the increasing trend in mean values as the number of occupants within the house increases. It was expected that an increase in mean consumption would allow the classifiers to better distinguish between household categories. We also see the variety in the degree of dispersion and shape of the distribution across the household categories. The use of the local outlier factor algorithm means that only households that are relatively extreme were removed during preprocessing. Recall also that variance, standard deviation and IRQ are among the extracted features in Table 2.

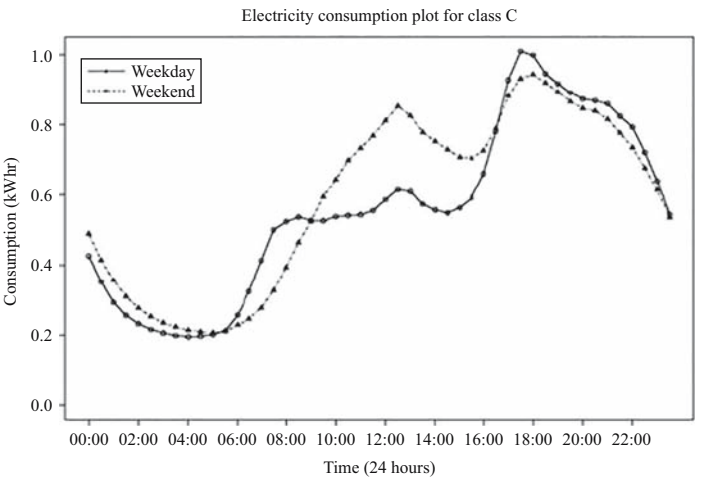


Fig. 3. Sample electricity relative load curves for household category C (three adults).

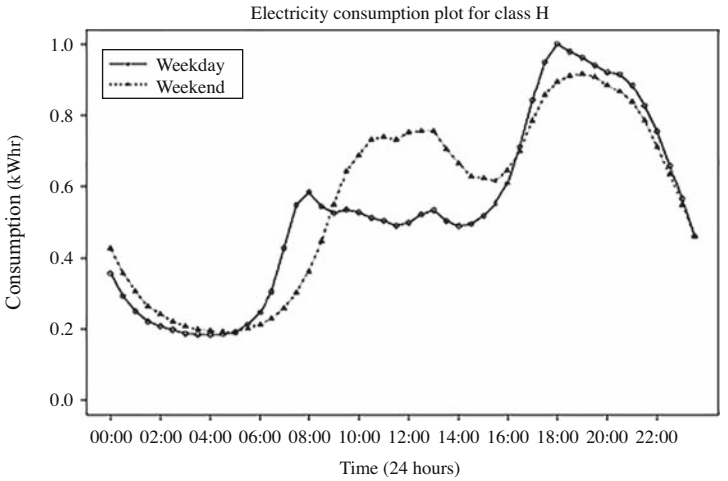


Fig. 4. Household category H (two adults plus one child).

5.1. Classifier Results

The results from the Elastic Net Logistic regression model (EN) and from the Neural Net (NN) models were quite similar, see Table 3. Results for the simpler classification scheme (of numbers of persons) were not significantly better. In some cases it was slightly better, in other cases, it was slightly worse. In the interests of brevity, we present the results for our number of adults/children classification scheme (which is detailed in Table 1).

The six month TS window produced the best performance. The performance benefits in the longer time frame varied in comparison to other time windows. The six month window

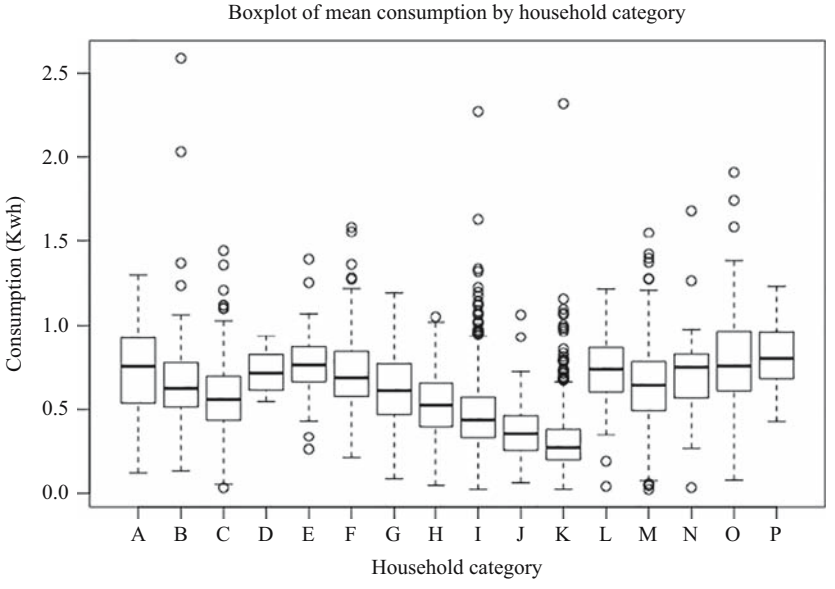


Fig. 5. Boxplot of mean daily usage per household category.

may capture some long distance interactions or seasonality. Again, in the interests of brevity we present the results for the six month window only.

Balanced training data gives the better results for both the EN and NN binomial approaches as shown in Table 3. The unbalanced binomial network has the highest accuracy and lowest SSE and RMSE value. This is misleading however as the classifier could classify everything as zero to yield an accuracy value equal to the proportion of zero or "no" in the actual values which in this case is 93.75%. A similar situation arises with the SSE and RMSE figures. The CV and R^2 terms are the only error metrics presented which can be used to effectively compare the models as they are dimensionless.

For comparison, we note that the corresponding balanced binomial NN using a single work week window produced Accuracy, SSE, RMSE, CV, and R^2 values of (57.017, 41.798, 0.493, 98.676, 0.513). These can be compared with the last line of Table 3 which shows the values of the six month window. Such empirical evidence was used to guide the selection of the design parameters during model evaluation.

Details of the performance of the balanced data binomial models are shown in Table 4.

For brevity, we present just the results of the individual classifiers, that is, asking whether test meters belong to a particular household category. The balanced binomial EN model has the highest R^2 value of 0.55 which signifies that 55% of the variability in the actual values is explained by the model. The best binomial NN was obtained using balanced data with an R^2 value of 0.54. Note the high performance for single adult household category K. This may indicate that category K is more distinctive than the other categories. Recall that category K accounts for 25% of the population, see Table 1. Table 5 shows a sample confusion matrix when testing sample meters for membership of household category K (single adult) using the best binomial NN. The matrix is diagonally dominant but more households are classified as true (163) than as false (123). In this example, the classifier is giving "false positives".

Scatter plots such as Figure 6 are useful to visualise the partitioning ability of the classifier. The *y*-axis refers to predicted probability (equivalent to the probability that meter belongs to a particular class). The *x*-axis labelled as "index" refers to the *i*th test object. The data are evenly distributed data between the upper and lower halves of the plot area for both the EN and NN. Dark coloured dots represent households that are *true*, that is,

Table 3. Testing results – binomial, balanced versus unbalanced data.

	Testing data				
Classifier	Accuracy	SSE	RMSE	CV	R^2
Unbalanced binomial EN[1]	93.79	57.01	0.19	719.52	0.08
Unbalanced binomial EN[2]	88.75	100.79	0.28	343.34	0.15
Balanced binomial EN[2]	60.52	60.05	0.48	94.16	0.55
Unbalanced binomial NN[1]	93.750	38.850	0.193	837.141	0.053
Unbalanced binomial NN[2]	88.792	67.799	0.285	349.754	0.109
Balanced binomial NN[2]	63.264	38.235	0.476	95.169	0.544

[1]Results show the mean values from the 16 individual binomial models for household categories A-P.
[2]Results show the mean values from the eight binomial models for household categories B, C, F, G, H, I, K, and M, that is, the households used in the balanced data analysis. The six month time frame is used.

Table 4. Testing results – household category binomial models using balanced data.

		Test data				
Classifier	Household category	Accuracy	SSE	RMSE	CV	R^2
Bal. Bin. EN	B	58.73	14.19	0.47	87.94	0.58
Bal. Bin. EN	C	44.70	66.56	0.50	90.14	0.55
Bal. Bin. EN	F	62.92	22.17	0.50	105.77	0.47
Bal. Bin. EN	G	70.35	42.03	0.46	101.62	0.53
Bal. Bin. EN	H	57.14	37.86	0.51	105.07	0.47
Bal. Bin. EN	I	55.94	180.27	0.49	99.05	0.51
Bal. Bin. EN	K	76.1	73.37	0.41	87.17	0.64
Bal. Bin. EN	M	61.76	38.89	0.48	82.13	0.61
Bal. Bin. EN	**Mean**	**60.96**	**59.42**	**0.48**	**94.87**	**0.55**
Bal. Bin. NN	B	50.00	10.46	0.51	102.29	0.48
Bal. Bin. NN	C	63.79	39.99	0.48	95.89	0.54
Bal. Bin. NN	F	70.69	13.45	0.48	96.30	0.54
Bal. Bin. NN	G	64.39	30.49	0.48	96.13	0.54
Bal. Bin. NN	H	54.17	25.41	0.52	102.90	0.47
Bal. Bin. NN	I	57.63	119.81	0.49	98.10	0.52
Bal. Bin. NN	K	79.37	41.98	0.38	76.62	0.71
Bal. Bin. NN	M	66.07	24.28	0.47	93.13	0.57
Bal. Bin. NN	**Mean**	**63.26**	**38.24**	**0.48**	**95.17**	**0.54**

test households that are in category K. Any dark dot above 0.5 is correctly classified. We see some dark dots below the 0.5 threshold. These are test objects that are incorrectly classified as not being in category K.

The grey dots represent test objects that are *false*, that is, not in category K. Again we see that some are correctly classified (below the 0.5 line) and some are incorrectly classified (above the 0.5 line). These plots show that the classifiers have similar prediction accuracy for both *true* and *false*. The majority of the predictions are concentrated in the top and bottom quarters of the NN plot area as expected from a good classifier.

We had very limited success using multinomial classifiers using balanced data. Again the EN and NN had similar performances. The unbalanced data has the better R^2 value, but the balanced approach has the lower SCE and CV values as shown in Table 6 for the

Table 5. Sample confusion matrix, household category K, binomial NN using balanced data.

		Predicted		
Test		False	True	Σ
Actual	False	100	43	143
	True	23	120	143
	Σ	123	163	286

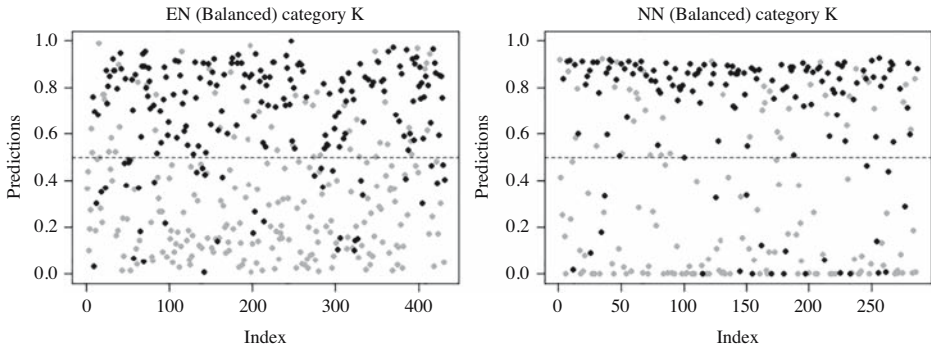


Fig. 6. Scatter plot for household category K (single adult), binomial EN and NN using balanced data.

six month time frame. Recall that R^2 measures how well the variability in the actual values is captured by the model. If the model is distorted or biased towards a particular category then the R^2 value can also be distorted by the unbalance. The CV value is not affected by the imbalance as it only evaluates the relative closeness of the predictions to the actual values. It is useful when comparing models which use either balanced or unbalanced data.

A sample confusion matrix for a multinomial NN is shown in Table 7. Category K can be predicted with 75% accuracy by the NN but category B displays an accuracy of 0%. The R^2 for this model was 0.16.

5.2. *Results Summary*

In summary, the binomial approaches trained on the six month time series using balanced training data achieved the best performance. They are of less practical value than a multinomial classifier as they have to be tested against each household category and a weighted average calculated to yield an equivalent multinomial response. There was no significant difference between the EN and NN classifiers or simpler number of persons classification scheme. Some household categories were easier to identify than others. The R^2 value of the balance binomial NN for single person households was 0.71 (Table 4).

6. Discussion and Conclusion

This novel study describes an approach to household classification using smart meter data. The study presents a proof of concept for the use of ML and GLM models on new data

Table 6. Testing results – multinomial approach.

	Testing data					
Classifier	Accuracy	SCE	SSE	RMSE	CV	R^2
Unbal. Multi EN	39.00	751.91	781.76	0.86	698.15	0.26
Bal. Multi EN	21.55	212.92	209.92	0.92	733.64	0.16
Unbal. Multi. NN	35.27	655.05	600.43	0.88	1409.23	0.22
Bal. Multi. NN	21.25	138.86	134.73	0.92	734.10	0.16

Table 7. Sample confusion matrix for multinomial NN using balanced data.

Test		Predicted by household category									
		B	C	F	G	H	I	K	M	Σ	% Acc
Actual by household category	B	0	0	6	6	0	6	2	0	20	0
	C	0	0	4	10	1	3	1	1	20	0
	F	0	0	8	6	0	4	2	0	20	40
	G	0	0	5	2	1	8	4	0	20	10
	H	0	0	4	4	1	7	4	0	20	5
	I	1	0	1	2	0	8	8	0	20	40
	K	0	0	0	0	0	5	15	0	20	75
	M	0	0	10	4	0	3	3	0	20	0
		1	0	38	34	3	44	39	1	160	

sources such as SMD for use in the production of official statistics and by the public sector. The binomial approach proved more useful to gain insight to household composition. The multinomial models have greater potential for practical applications but were less able to distinguish between the categories.

This study focused on exploring a specific consumer behaviour trial smart meter data set with a view to learning a little more about it in the context of official statistics. The data were anonymised so could not be linked to other data sources but may yet provide a set of auxiliary information to allow NSIs determine whether a house is occupied or not, provide estimates of the number of persons living in a house. Rich data on households, giving a good indicator of a person or the number of persons living in a household, is highly valuable to developing small area statistics or census like statistics on small areas.

The CSO is exploring additional data sources, (Dunne 2015). A building energy rating system has been in operation in Ireland since 2009. The CSO has access to this data but currently only one third of households have been rated. As this system evolves, it may be a potential administrative source that could be linked to live (unanonymised) SM data to improve the classification performance. This project has not yet been costed, but is one of a number of possible data sources being considered for inclusion as a piece of the jigsaw in the developing National Data Infrastructure. While the CSO has access under the Statistics Act to access utility data such as SMD, it needs to evaluate how accessing such data can be socially justified and that any such access is proportionate and protects the privacy of the individual.

The insights gained during this study highlight some of the challenges and problems associated with classification schemes. The aim was to evaluate SMD to identify *existing* CSO household composition categories. As noted in Subsection 4.1, this smart meter study uses a simpler household categorisation system based on the numbers of adults and

children sharing a dwelling unit. Some households, such as single person households (Category K) appear to be more easily identified. However, the usage patterns for most existing CSO household categories are not sufficiently unique to be identified by their electricity consumption alone. There is a potential role for SMD driven models to estimate household composition in nonresponse or hard to reach households. It is also likely the classifiers could identify an empty (zero occupants) household. No such households were included in the CBT.

There is significant interest in NSI communities to identify and harness new data sources which may offer the opportunity for new insights. These sources may not be well structured, may have corrupt or missing segments, and may require considerable preprocessing to be manipulated into a useful format for analysis. Furthermore, the data may come from a domain not familiar to NSI staff and will involve a significant learning curve.

NNs and ML techniques have become more widely used for classification tasks, offering alternatives to traditional statistical methods for organisations intent on exploring new, possibly noisy, data sources. The NN and EN models had similar performance. There was no distinct advantage in favour of either the machine learning or generalised linear modelling approach. Neither approach was able to classify households with high reliability. The confusion matrices give some insight into how households can be misclassified based on the similarity of usage patterns. Statistical models such as ENs may be more familiar to NSI communities in comparison to ML techniques so may be a more suitable approach.

Finally, in response to our research question whether CSO household composition can be estimated from analysis of SM electricity usage, we report only limited success in identifying households in general, but suggest that future studies linking SMD to supplementary information about the dwelling/building or other properties of the household could be beneficial.

Appendix 1

Features

The 21 features created for development of the models are described below. l is the total number of half hourly intervals over the particular time frame, n is the total number of intervals in a day, m is the total number of days in the time frame and E is the electrical demand in kWh:

1. Mean: mean consumption over the time frame l. $E_{mean} = \frac{1}{l}\sum_{i=1}^{l} E_i$
2. Max: maximum consumption during the time frame l. $E_{max} = \max\left(\{E_i\}\right)$ where $1 \leq i \leq l$
3. ToUmax: the time slot i when Max occurs, $1 \leq i \leq n$. Note $n = l \pmod{48}$ as we cycle through the days in a time frame.
4. TEC: total electricity consumed over the time frame l. $E_{TEC} = \sum_{i=1}^{l} E_i$
5. MDM: Mean daily max is the average of the Max values for each of the m days. $E_{MDM} = \frac{1}{m}\sum_{j=1}^{m} E_j$ where $E_j = max(\{E_i\})$ for each m days: $1 + n(m-1) \leq i \leq nm$

6. Load Factor: This is the average of the ratios of the daily mean to daily maximum consumption. It is a measure of the peak of a household's load profile. A larger load factor indicates a household who uses electricity more evenly across the day while a low load factor indicates small periods of large consumption. For example, the load factor for the first day is $E_{LF_1} = \frac{(1/n)\sum_{i=1}^{n} E_i}{\max(\{E_i, 1 \leq i \leq n\})}$.
7. Variance: a measure of how far the electricity readings are spread out from the mean reading. $E_{VAR} = \frac{1}{l}\sum_{i=1}^{l}(E_i - E_{mean})^2$
8. Standard Deviation: a measure of the variation or dispersion around the mean reading, it is the square root of the variance. $E_{SD} = \sqrt{E_{VAR}}$
9. Range: the difference between the biggest and smallest electricity consumption readings. $E_{Range} = \max\left(\{E_i,\ 1 \leq i \leq l\}\right) - min(\{E_i,\ 1 \leq i \leq l\})$
10. Interquartile range (IQR): measures the difference between the third quartile and first quartile values of the data.
11. Morning max: the mean daily maximum electricity demand prior to 10 am on a weekday. $E_{Mornmax} = \frac{1}{m}\sum_{j=1}^{m} E_j$ where $E_j = max(E_i)$ and i is within the first 20 time slots of each day.
12. Morning peak: the morning max minus the mean value between 10 am and 12 am on a weekday. This feature measures the size of a morning spike if one exists. It was observed that households with children were more likely to have a defined peak in the morning time on a weekday.
13. Morning range: the Morning max minus the minimum value before 10 am.
14. Weekday Area: The area under the curve was approximated using the trapezoid rule.
15. Weekday Midpoint: The midpoint of the function was defined as half the total weekday area, the value returned was the time of day where the midpoint occurred.
16. Weekday Centroid: Analogous to the geometric centroid, the centroid of a function is the "centre of mass" of that function.
17. Weekday AM Slope: The slope of the early morning peak (up to 10 am) was taken as the rate of increase of energy consumption over time during the early morning period.
18. –21. The procedures for features 15–17 were repeated to produce the equivalent features derived from the weekend energy consumption. These features were observed to differ to those during the working week, possibly due to behavioural changes at weekends.

Appendix 2

Error Metrics

The models were assessed using the following error metrics where $y_i =$ predicted value of the ith meter, $t_i =$ true value of the ith meter, $N =$ Number of meters and $\bar{t} =$ mean of the true values.

1. Percentage of Correct Predictions (Accuracy): For the binomial model, the values predicted by the classifier are rounded to the nearest integer. For example, a

prediction of 0.364 is rounded to zero, which indicates *false*. This says the meter does not belong to the category being tested. A predicted value of 0.759 is interpreted as *true* and means that the meter is assigned to that particular category. If the predicted category matches the true category, then the prediction is correct.

For a multinomial classifier, a "winner-takes-all" approach assigns the category with the largest value to 1 and sets the remaining categories to 0. For example, a model concerned with four household categories produces output (0.25, 0.48, 0.10, 0.17). This is interpreted as (0, 1, 0, 0). This is a correct prediction if this was the true category of the meter.

2. Sum of Squared Error (SSE): The sum of the squared differences between the actual and predicted value. SSE $= \sum_{i=1}^{N} (y_i - t_i)^2$.
3. Root Mean Squared Error (RMSE): RMSE is an extension of SSE. The SSE is divided by the total number of meters to find the mean squared error (MSE). $RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - t_i)^2}{N}}$.
4. Sum of Cross Entropy Error (SCE): The SCE is computed for each of the meters in the data set and is summed over the entire data set to get the SCE for the data set. RMSE was used to compute the training error in the binomial classifier while SCE is used in the multinomial classifiers. SCE $= -1 \sum_{i=1}^{N} (t_i \log_{10} y_i)$.

   The reason for choosing SCE over RMSE for the multinomial classifier is demonstrated in the following example: a classifier concerned with four household categories produces (0.12, 0.57, 0.16, 0.15) and the true classification is (0, 1, 0, 0). The RMSE is 0.25.

   Now suppose the predicted output was (0.33, 0.57, 0.0, 0.1), the RMSE is then 0.28. Although the probability of the meter being classified as the second category is the same in both cases, the RMSE differs. Using the SCE all but one of the error terms is zero and the SCE for both cases is $-\log(0.57) = 0.24$.
5. Coefficient of Variation (*CV*): *CV* evaluates the relative closeness of the predictions to the actual values. The *CV* for a model describes the accuracy of the model in terms of the relative sizes of the residuals and the actual values. A high *CV* represents a large dispersion in the variables. An advantage to using this error term is that it is unitless and therefore it can be used to compare model performance. For balanced data, the *CV* value is just a multiple of the RMSE term but for unbalanced data it is particularly useful. $\frac{\sqrt{\frac{\sum_{i=1}^{N} (y_i - t_i)^2}{N}}}{\bar{t}} 100$.
6. Pseudo R^2: R^2 quantifies how much of the variability is explained by the model. It indicates how well the data points fit some model representation of the data. Like *CV*, *R*-squared is unitless. In this study pseudo R^2 is defined as: $1 - \frac{\sum_{i=1}^{N} (y_i - t_i)^2}{\sum_{i=1}^{N} (t_i)^2}$

   The values of R^2 lie between zero and one. An R^2 value of 1 represents a perfect fit while a value of 0 represents inappropriate model fit.

7. References

Aydinalp, M., V.I. Ugursal, and A.S. Fung. 2002. "Modeling of the Appliance, Lighting, and Space-cooling Energy Consumptions in the Residential Sector Using Neural Networks." *Applied Energy* 71(2): 87–110. Doi: http://dx.doi.org/10.1016/S0306-2619(01)00049-6.

Bousquet, O. and A. Elisseeff. 2002. "Stability and Generalization." *Journal of Machine Learning Research* 2(3): 499–526.

Breunig, M.M., H.P. Kriegel, R.T. Ng, and J. Sander. 2000. "LOF: Identifying Density-Based Local Outliers." In *ACM sigmod record* 29(2): 93–104. Doi: http://doi.acm.org/10.1145/335191.335388.

Commission for Energy Regulation (CER). 2012. CER Smart Metering Project – Electricity Customer Behaviour Trial, 2009–2010 [dataset]. 1st Edition. Irish Social Science Data Archive. SN: 0012-00. www.ucd.ie/issda/CER-electricity (accessed January 15, 2018).

CER. 2014. *Commission for Energy Regulation National Smart Metering Programme Smart Metering High Level Design*. Decision Paper CER/14/046. Available at: http://www.cer.ie/docs/000699/CER14046%20High%20Level%20Design.pdf (accessed March 2017).

European, Commission. 2014. *A Policy Framework for Climate and Energy in the Period from 2020 to 2030. COM (2014), 15*. Available at: http://ec.europa.eu/smart-regulation/impact/ia_carried_out/docs/ia_2014/swd_2014_0015_en.pdf (accessed March 2017).

CSO. 2011. "Census 2011 Reports." Available at: http://www.cso.ie/en/census/census2011reports/ (accessed September 2017).

Di Cosmo, V., S. Lyons, and A. Nolan. 2012. "Estimating the Impact of Time-of-Use Pricing on Irish Electricity Demand." *The Energy Journal* 35(2): 117–136. Doi: http://dx.doi.org/10.5547/01956574.35.2.6.

Dunne, J. 2015. "The Irish Statistical System and the Emerging Census Opportunity." *Statistical Journal of the IAOS* 31(3): 391–400. Doi: http://dx.doi.org/10.3233/SJI-150915.

Friedman, J., T. Hastie, and R. Tibshirani. 2009. "glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. R package version, 1.4". CRAN: Wien, Austria. Available at: https://CRAN.R-project.org//src/contrib/Archive/glmnet (accessed January 15, 2018).

Fritsch, S., F. Guenther, and M. Suling. 2012. "neuralnet: Training of Neural Networks. R package version 1.32. 2012." CRANWien, Austria. Available at: https://CRAN.R-project.org//src/contrib/Archive/neuralnet (accessed January 15, 2018).

Han, J. and M. Kamber. 2006. *Data Mining: Concepts and Techniques*. 2nd Edition. Morgan Kaufmann. ISBN 13: 978-1-55860-901-3.

Hand, D.J. 1998. "Data Mining: Statistics and More?" *The American Statistician* 52(2): 112–118. Doi: http://dx.doi.org/10.1080/00031305.1998.10480549.

He, H. and A.W. Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 29(9): 1263–1284. Doi: http://dx.doi.org/10.1109/TKDE.2008.239.

Hickey, C., S. Kelly, P. Carroll, and J. O'Connor. 2015. "Prediction of Forestry Planned End Products Using Dirichlet Regression and Neural Networks." *Forest Science* 61(2): 289–297. Doi: https://doi.org/10.5849/forsci.14-023.

Hopfield, J.J. 1984. "Neurons with Graded Response have Collective Computational Properties Like Those of Two-State Neurons." In proceedings of the National Academy of Sciences of the United States of America 81(10): 3088–3092.

Hoerl, A. and R. Kennard. 1988. "Ridge Regression." *Encyclopedia of Statistical Sciences* 8: 129–136. Doi: http://dx.doi.org/10.1002/0471667196.ess2280.pub2.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, RC Team, and M. Benesty. 2014. "caret: Classification and Regression Training. R package version 6.0–21." Wien, Austria: CRAN. Available at: https://CRAN.R-project.org/package=caret (accessed January 15, 2018).

Liao, W. 2005. "Clustering of Time Series Data - a Survey." *Pattern Recognition* 38(11): 1857–1874. Doi: http://dx.doi.org/10.1016/j.patcog.2005.01.025.

Marsland, S. 2009. *Machine Learning: An Algorithmic Perspective*. Chapman and Hall/CRC. ISBN:978-1-4200-6718-7.

McKenna, E., I. Richardson, and M. Thomson. 2012. "Smart Meter Data: Balancing Consumer Privacy Concerns with Legitimate Applications." *Energy Policy* 41: 807–814. Doi: https://doi.org/10.1016/j.enpol.2011.11.049.

McLoughlin, F., A. Duffy, and M. Conlon. 2012. "Characterising Domestic Electricity Consumption Patterns by Dwelling and Occupant Socio-Economic Variables: An Irish Case Study." *Energy and Buildings* 48: 240–248. Doi: http://dx.doi.org/10.1016/j.enbuild.2012.01.037.

Molina-Markham, A., P. Shenoy, K. Fu, E. Cecchet, and D. Irwin. 2010. "Private Memoirs of a Smart Meter." In *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-efficiency in Building*: 61–66. Doi: http://dl.acm.org/citation.cfm?id=1878446.

Murphey, Y.L., H. Guo, and L.A. Feldkamp. 2004. "Neural Learning from Unbalanced Data." *Applied Intelligence* 21(2): 117–128. Doi: http://dx.doi.org.ucd.idm.oclc.org/10.1023/B:APIN.0000033632.42843.17.

Newing, A., B. Anderson, A. Bahaj, and P. James. 2016. "The Role of Digital Trace Data in Supporting the Collection of Population Statistics–the Case for Smart Metered Electricity Consumption Data." *Population, Space and Place* 22(8): 849–863. Doi: http://dx.doi.org/10.1002/psp.1972.

Patidar, R. and L. Sharma. 2011. "Credit Card Fraud Detection Using Neural Network." *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, 2231–2307. Doi: http://dx.doi.org/10.1007/978-3-319-46675-0_53.

R Core Team. 2013. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. ISBN 3-900051-07-0, Available at: http://www.R-project.org/ (accessed March 2017).

Ripley, B. and W. Venables. 2011. "nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models. R package version, 7(5)." CRAN: Wien, Austria. Available at: https://CRAN.R-project.org/package=nnet (accessed January 15, 2018).

Seyb, A., R. McKenzie, and A. Skerrett. 2013. "Innovative Production Systems at Statistics New Zealand: Overcoming the Design and Build Bottleneck." *Journal of Official Statistics* 29(1): 73–97. Doi: https://doi.org/10.2478/jos-2013-0005.

Statistics Act. 1993. *An Act to Provide for the Collection, Compilation, Extraction and Dissemination of Official Statistics and for Related Matters*. Available at: http://www.irishstatutebook.ie/1993/en/act/pub/0021/print.html (accessed March 2017).

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society* (Series B) 58: 267–288. Available at: http://www.jstor.org/stable/2346178 (accessed January 15, 2018).

Van Gerwen, R., S. Jaarsma, and R. Wilhite. 2006. *Smart Metering, Technical Report*. Available at: https://idc-online.com/technical_references/pdfs/electrical_engineering/Smart_Metering.pdf (accessed March 14, 2017).

Wang, X., K. Smith, and R. Hyndman. 2006. "Characteristic-Based Clustering for Time Series Data." *Data Mining and Knowledge Discovery* 13: 335–364. Doi: http://dx.doi.org/10.1007/s10618-005-0039-x.

Zhang, L. and B. Zhang. 1999. "A Geometrical Representation of McCulloch–Pitts Neural Model and Its Applications." *IEEE Transactions on Neural Networks* 10(4): 925–929. Doi: http://dx.doi.org/10.1109/72.774263.

Zou, H. and T. Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society* (Series B) 68: 301–320. Doi: http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.

Received May 2015
Revised March 2017
Accepted September 2017

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 27–39, http://dx.doi.org/10.1515/JOS-2018-0002

Constraint Simplification for Data Editing of Numerical Variables

Jacco Daalmans[1]

Data editing is the process of checking and correcting data. In practise, these processes are often automated. A large number of constraints needs to be handled in many applications. This article shows that data editing can benefit from automated constraint simplification techniques. Performance can be improved, which broadens the scope of applicability of automatic data editing. Flaws in edit rule formulation may be detected, which improves the quality of automatic edited data.

Key words: Data editing; constraint simplification; conditional constraints; optimization.

1. Introduction

Collected micro data usually contain errors, for example, pregnant men, average salary of five million euro and components of a total that do not add up to that total. Correction of such errors is often necessary to prevent flaws and inconsistencies in statistics to be published. The process of checking and correcting is called data editing, see, for example, De Waal et al. (2011) and Pannekoek et al. (2013). A common approach for data editing is based on the paradigm of Fellegi and Holt (1976), stating that the data in a record should be adapted to satisfy all edit rules by changing the fewest possible values.

Error localization according to the Fellegi and Holt paradigm can be formulated as a Mixed Integer Linear Program (MILP) problem, see, for example, De Waal et al. (2011). Although, in general, a solution to a data editing problem can be found in reasonable time – typically a few seconds – the worst-case performance of a MILP problem is known to be exponential in the problem size. Even when using modern computers, it can take hours or even days to obtain a solution for a single record.

From Operations Research and Artificial Intelligence it is well known that performance of a mathematical optimization problem can be improved by a constraint simplification step, see, for example, Paulraj and Sumathi (2010), Telgen (1983), Chmeiss et al. (2008), and Piette (2008). This means eliminating redundant constraints and simplifying unnecessary complicated constraints, before optimization. Nevertheless, remarkably few applications of constraint simplification are known in the context of data editing. Bruni (2005) explains how redundant edit rules can be detected. Also, Statistics Canada developed a software tool with

[1] Statistics Netherlands, PO Box 24500, 2490 HA Den Haag, The Netherlands. Email: j.daalmans@cbs.nl
Acknowledgments: The authors would like to thank three reviewers and the associate editor for useful comments that greatly contributed to improving the article. The author is also grateful for useful advice from Edwin de Jonge, Mark van der Loo, Jeroen Pannekoek, Sander Scholtus, and Ton de Waal.

simplification features (BANFF support team 2008, Chap. 2). These applications do however not allow for conditional ("IF-THEN") rules, where the variables involved in the IF and THEN statements may contain errors. Such rules frequently occur in official statistics and are especially problematic for computational performance, due to the integer variables that arise in the corresponding MILP problem.

This article contributes to fill the gap for constraint simplification techniques for error localisation of numerical data. Special attention will be given to conditional rules. We present automated methods that work at the formal level through solving MILP problems.

An advantage of automated procedures is that removal of redundant constraints can be done out of sight, so that users can still specify all possible rules without ending up with an inefficient edit set. Working at the formal level means that the methods can be applied to a generic class of rules, regardless of their semantic meaning.

Since edit rule simplification improves computational performance, it has the potential of further extending the application possibilities of automated data editing. Besides this, expert's feedback on automatically detected redundant edit rules might help to increase the understanding of the joint effects of a set of rules. Due to the complex interdependence and misspecification, a set of rules may have different implications than intended. Correction of erroneous rules improves the quality of automatic edited data and avoids the need for manual adjustment of results. For example, the following redundant rule was found in an edit set, actually used by Statistics Netherlands:

$$\text{IF}(\text{Questionaire\_ID} \neq 1 \text{ OR } \text{Questionaire\_ID} \neq 2) \text{THEN} (\text{VariableX} = \text{VariableY}) \quad (1)$$

Manual inspection might reveal that the OR-operator was meant to be an AND-operator.

The structure of this article is as follows. Section 2 describes the MILP formulation for data editing of numerical variables. Sections 3–5 present formal, mathematical algorithms for simplifying edit sets: eliminating single variables is described in Section 3, eliminating redundant parts from conditional rules is discussed in Section 4 and the redundancy of rules as a whole is considered in Section 5. Section 6 presents real-life applications of constraint simplification and data editing. Finally, Section 7 finishes this article with a discussion.

2. Outline of Basic Approach

We introduce the basic idea of MILP problems first. Then, it is explained how edit rules can be translated into MILP constraints.

2.1. *Definition of a MILP Problem*

A MILP problem consists of a loss function to be minimized and a set of inequality constraints involving both real and integer variables. A general form is given by

$$\begin{aligned} &\text{Minimize } f(\mathbf{x}, \mathbf{z}) = \mathbf{c}^{\mathrm{T}} \begin{pmatrix} \mathbf{x} \\ \mathbf{z} \end{pmatrix}, \\ &\text{s.t. } \mathbf{A} \begin{pmatrix} \mathbf{x} \\ \mathbf{z} \end{pmatrix} \leq \mathbf{b}, \\ &\mathbf{x} \in \mathbb{R}^p \text{ and } \mathbf{z} \in \mathbb{Z}^q \end{aligned} \quad (2)$$

where $\mathbf{x}$ and $\mathbf{z}$ are vectors of real and integer decision variables, $\mathbf{c}$ is a constant vector ($\mathbf{c} \in \mathbb{R}^{p+q}$), $\mathbf{A}$ is a coefficient matrix and $\mathbf{b}$ a vector of upper bounds, see, for example, Bertsimas and Tsitsiklis (1997).

In the remainder of this article several algorithms are proposed that make use of the feasibility of a set constraints. This can be checked by a MILP solver by using a trivial loss function with $\mathbf{c} = \mathbf{0}$. Of course, if a solution exists the optimum value will be zero, but if the set of constraints is infeasible, most MILP solvers return an error message.

2.2. Edit Rules as MILP Constraints

This subsection introduces the edit rules that are considered in this article and explains how these rules can be transformed into MILP constraints. The edit rules in this article can be subdivided into unconditional and conditional rules.

We consider linear unconditional rules, like

$$\begin{aligned} &\text{Total turnover} = \text{Domestic turnover} + \text{Foreign Turnover},\\ &\text{Total turnover} \geq 0, \end{aligned}$$

that can be straightforwardly formulated as MILP contraints. One could note that the constraints in (2) do not allow for "larger than" and "equality" signs, but it is well-known that these rules can be reformulated into the required form. For example, an equality can be written as two inequalities and a constraint $x > 0$ can be approximated by $-\mathrm{x} \leq -\epsilon$, where ϵ is a sufficiently small value.

We also consider 'simple' and ' compound' conditional rules. A 'simple' conditional edit has the following form

$$\text{IF} < \text{Statement 1} > \text{THEN} < \text{Statement 2} >,$$

where each "statement" is a linear equality or inequality. Compound rules may also contain:

- AND-operators in the IF-clause and/or
- OR-operators in the THEN-clause.

An example is:

$$\begin{aligned} &\text{IF}\,(\text{Number of employees} > 0\ \text{AND Turnover} > 0)\ \text{THEN}\\ &\quad(\text{Wages} > 0\ \text{OR Labour costs} > 0). \end{aligned} \tag{3}$$

Note that above we did not consider rules with:

- OR-operators in the IF-clause and/or
- AND-operators in the THEN-clause,

but these rules can be rewritten as a number of simple conditional rules. For example, the edit:

$$\begin{aligned} &\text{IF}\,(\text{Number of employees} > 0\ \text{OR Turnover} > 0)\ \text{THEN}\\ &\quad(\text{Wages} > 0\ \text{AND Labour costs} > 0) \end{aligned}$$

is equivalent to the combination of the following four "simple" conditional rules:

IF Number of employees > 0 THEN Wages > 0,

IF Number of employees > 0 THEN Labour costs > 0,

IF Turnover > 0 THEN Wages > 0,

IF Turnover > 0 THEN Labour costs > 0.

As mentioned by Chen et al. (2010), conditional rules need to be expressed in Disjunctive Normal Form (DNF), before these can be further converted into the required MILP format. A DNF is a disjunction of assignments (a sequence of OR's) that makes a rule *True*, see, for example, Hooker (2000).

To explain the transformation to DNF, note that a conditional rule is satisfied, if either the IF-clause is violated, or if the THEN-clause is fulfilled. Thus, a condition rule can be put in DNF, by joining the negation (i.e., opposite) of the "IF"-clause with the original "THEN"-clause. For example, the rule: "If Turnover > 0 THEN Wages > 0" can be stated as "Turnover ≤ 0 OR Wages > 0".

For compound edits, the IF-clause is assumed to be a conjunction (sequence of AND's). According to Morgan's law, the negation of a conjunction is a disjunction of negations. To illustrate this, the example in (3) can be written in DNF as

Number of employees ≤ 0 OR Turnover ≤ 0 OR

Wages > 0 OR Labour costs > 0,

where the first two statements are negations of the original IF-clause statements.

An expression for n_C edit rules in DNF is given by

$$\bigcup_{j=1}^{D_i} \left(\left(\boldsymbol{a}_{ij}^C \right)^T \mathbf{x} \leq b_{ij}^C \right) \quad i = 1, \ldots, n_C. \tag{4}$$

where an edit rule i is stated as a disjunction with D_i disjunctive terms. The coefficients and upper bounds for the jth term are denoted by $\boldsymbol{a}_{ij}^C$ and b_{ij}^C respectively. Again, 'equality', 'larger than' or 'smaller than' constraints can be reformulated into the form (4).

To express the constraints in (4) as MILP constraints, the following formulation can be used, based on the so-called Big M method.

$$\begin{aligned} &\left(\boldsymbol{a}_{ij}^C \right)^T \mathbf{x} \leq b_{ij}^C + M(1 - z_{ij}), \quad i = 1, \ldots, n_C, \quad j = 1, \ldots, D_i, \\ &\sum_{j=1}^{D_i} z_{ij} = 1 \quad i = 1, \ldots, n_C, \\ &-z_{ij} \leq 0 \quad i = 1, \ldots, n_C, \quad j = 1, \ldots, D_i. \end{aligned} \tag{5}$$

where z_{ij} are integer variables and M is a sufficiently large constant.

The equation $\sum_{j=1}^{D_i} z_{ij} = 1$ guarantees that only one disjunctive term is selected per disjunction. For each selected term (i, j with $z_{ij} = 1$), it is enforced that $\left(\boldsymbol{a}_{ij}^C \right)^T \mathbf{x} \leq b_{ij}^C$. For each non-selected term (i, j with $z_{ij} = 0$), the first constraint in (5) becomes redundant.

As shown in (5) integer variables are needed for the formulation of conditional rules. Because integer variables are much less efficiently handled than continuous variables, conditional rules can be less efficiently processed than unconditional ones. Therefore, it is very beneficial to replace conditional rules by unconditional ones.

3. Fixed Value Elimination

The aim of this technique is to shorten edit rules by elimination of 'fixed' variables, that is, variables with only one admissible value. As a result, an edit set may become simpler, possibly giving rise to a better performance of data editing software. Moreover, misspecification of edit rules might be detected by manual inspection of fixed values. Consider the following example:

$$\begin{aligned} &\text{Edit } 1\text{: } x_1 + x_2 + x_3 = 10, \\ &\text{Edit } 2\text{: } x_1 + x_2 \geq 10, \\ &\text{Edit } 3\text{: } x_3 \geq 0. \end{aligned} \tag{6}$$

It is immediately clear that x_3 necessarily has to be zero. In other words, x_3 is a fixed variable. Fixed values can be identified by solving two MILP programming problems for each continuous variable. Each variable is minimized and maximized once, subject to the MILP representation of the edit rules. If the minimum and maximum value turn out to be the same, the variable at hand is a fixed variable. Its value can be substituted in all edits in which it appears and a constraint is added stating that the fixed variable can only attain the fixed value.

Besides fixed values, any finite minimum or maximum is a candidate for content-wise analysis, because these bounds may be different than intended.

In our example, we can add the rule $x_3 = 0$ to our edit set and substitute x_3 in all other rules. We obtain

$$\begin{aligned} &\text{Edit } 1'\text{: } x_1 + x_2 = 10, \\ &\text{Edit } 2'\text{: } x_1 + x_2 \geq 10, \\ &\text{Edit } 3'\text{: } 0 \geq 0. \\ &\text{Edit } 4'\text{: } x_3 = 0. \end{aligned} \tag{7}$$

Of course, these edits can be further simplified, Edits $2'$ and $3'$ are obviously redundant. The further simplification for redundant rules will be explained in Section 5.

4. Simplification of Compound Rules

This section deals with the simplification of compound rules by elimination of unnecessary statements. Two new MILP algorithms are presented, based on existing methods from Dillig et al. (2010). The aims are again to improve computational performance and to detect misspecification of edit rules. A possible outcome, especially beneficial to computation performance, is that a conditional rule can be replaced with an unconditional one.

4.1. Implicitly Unsatisfiable Statements

In this subsection compound edit statements of the form (*A* OR *B* OR . . .) are simplified by deletion of statements that cannot be satisfied, given the available set of edit rules. Dillig et al. (2010) call these statements "non-relaxing", since these do not enhance the feasible area of a MILP problem. If, after simplification, only one component remains, a conditional rule has been converted into an unconditional one. An example is

$$\text{Edit 1: } x_1 > 0 \text{ OR } x_2 > 0,$$
$$\text{Edit 2: } x_2 < 0.$$

It is immediately clear that the statement $x_2 > 0$, within Edit 1, cannot possibly be satisfied, because of Edit 2. This statement can be removed from Edit 1, since it is redundant. Consequently, Edit 1 can be formulated as an unconditional rule. A more formal definition is given below:

Definition:
A statement e_{ij} of a compound edit e_i within a feasible edit set $\mathbf{E}$ is implicitly unsatisfiable, if $\mathbf{E} \cup e_{ij}$ is infeasible.

Here, $\mathbf{E} \cup e_{ij}$ stands for the edit set that is obtained by extracting a compound edit's component e_{ij} from e_i and adding it to the set $\mathbf{E}$, as if it were a single edit.

An algorithm for removal of implicitly unsatisfiable statements is stated below

Algorithm 1: Identification and removal of implicitly unsatisfiable statements

Input: Feasible edit set $\mathbf{E}$
Output: Feasible edit set $\mathbf{E}$, without implicitly unsatisfiable components.
1 FOR each compound edit $e_i \in \mathbf{E}$ do
2 FOR each statement $e_{ij} \in e_i$ do
3 $\mathbf{E}^* \leftarrow \mathbf{E} \cup e_{ij}$;
4 IF isFeasible($\mathbf{E}^*$) = FALSE THEN $e_i \leftarrow e_i \setminus e_{ij}$
5 NEXT
6 NEXT

In each step one statement of a compound edit is added to a feasible edit set. Subsequently, the feasibility of the extended edit set is checked by isFeasible(), a function that can be implemented by a MILP solver, see Section 2. If the extended edit set is infeasible, the compound edit's statement is implicitly unsatisfiable and therefore redundant.

When applied to our previous example, the algorithm means that the constraints $x_1 > 0$ and $x_2 > 0$ are added to Edits 1 and 2 one by one and that the feasibility is verified for both resulting edit sets. Because the addition of $x_2 > 0$ renders Edits 1 and 2 infeasible, $x_2 > 0$ is an implicitly unsatisfiable statement. It can be deleted from Edit 1 accordingly.

4.2. Implicitly Satisfied Statements

This subsection aims at replacing compound edit rules (*A* or *B* or. . .) with single, unconditional rules. The main idea is that if a compound rule contains a statement (say A)

that is necessarily *True*, the compound rule can be replaced with that single statement. Implicitly satisfied statements are called non-constraining by Dillig et al. (2010), since these do not reduce the feasible region of a MILP problem. Consider the following example:

$$\text{Edit 1: } x_1 < 50 \text{ OR } x_2 > 100,$$
$$\text{Edit 2: } x_1 > 100 \text{ OR } x_2 > 0.$$

For all possible x_1 values, at least one of the statements $x_1 < 50$ and $x_1 > 100$ is not satisfied. Thus, Edits 1 and 2 imply that either $x_2 > 0$, or the even stronger condition $x_2 > 100$, needs to be true. As a result, we obtain that $x_2 > 0$ always needs to hold, in other words $x_2 > 0$ is implicitly satisfied. Consequently, Edit 2 can be replaced with this single statement. A more formal definition is stated below:

Definition:
A component e_{ij} of a compound edit e_i within a feasible edit set $\mathbf{E}$ is implicitly satisfied if $\mathbf{E} \cup \neg\, e_{ij}$ is infeasible (where $\neg$ stands for negation).

This definition makes use of the equivalence of the statements that a compound edit's component is implicitly satisfied and that the opposite of that component cannot occur. An algorithm for identifying implicitly satisfied statements is as follows

Algorithm 2: Identification and replacement of implicitly satisfied statements

Input: Feasible edit set $\mathbf{E}$
Output: Feasible edit set **E,** without implicitly satisfied statements.
1 FOR each compound edit $e_i \in \mathbf{E}$ DO
2 FOR each statement $e_{ij} \in e_i$ DO
3 $\mathbf{E}^* \leftarrow \mathbf{E} \cup \neg\, e_{ij}$;
4 IF isFeasible($\mathbf{E}^*$) = FALSE THEN $\mathbf{E} \leftarrow \{\mathbf{E} \setminus e_i\} \cup e_{ij}$
5 NEXT
6 NEXT

This algorithm has a similar structure as Algorithm 1. Each step of the algorithm checks the feasibility of an extended edit set that is obtained by adding the negation of a statement of a compound rule to the given edits in $\mathbf{E}$. If the resulting edit set turns out to be infeasible, the added statement is "implicitly satisfied". The statement is added to the edit set as an unconditional rule and the conditional rule from which the statement is obtained is deleted.

When applied to our previous example, the constraints $x_1 \geq 50, x_2 \leq 100, x_1 \leq 100$ and $x_2 \leq 0$ are added to Edits 1 and 2 one by one, which are the negations of the original edit components. Feasibility is checked for all resulting edit sets. Because the addition of $x_2 \leq 0$ renders Edits 1 and 2 infeasible, $x_2 > 0$ is implicitly satisfied. Hence, Edit 2 can be replaced with the unconditional rule $x_2 > 0$.

5. Redundant Edit Removal

This section's aim is to simplify edit sets by removal of redundant edits, that is, rules that can be left out of an edit set, without affecting the set of feasible records. The removal of

redundant constraints may speed up the error correction process without losing power of correction. Because redundant edits may emerge as a result of fixed value substitution and simplification of conditional edits, it is important that redundant edit removal is conducted after these other steps. Consider the following example:

$$\begin{aligned}
&\text{Edit 1: } x_1 + x_2 \leq T_1,\\
&\text{Edit 2: } x_3 + x_4 \leq T_2,\\
&\text{Edit 3: } T_1 + T_2 = T_3,\\
&\text{Edit 4: } x_1 + x_2 + x_3 + x_4 \leq T_3.
\end{aligned}$$

Edit 4 can be omitted because it is implied by Edits 1, 2, and 3.

An edit is redundant if other edits imply that the edit is 'always satisfied'. As mentioned in Subsection 4.2, this is equivalent to the statement that the negation of the edit cannot occur. This leads to the following definition

Definition:
An Edit e_i from an edit set E is redundant, if $\{\boldsymbol{E} \setminus e_i\} \cup \neg\, e_i$ is infeasible.

The edit set $\{\boldsymbol{E} \setminus e_i\} \cup \neg\, e_i$ is obtained from **E**, by replacing Edit e_i by its negation.

In literature many methods have been mentioned for detection of redundant constraints. Paulraj and Sumathi (2010) performed a comparative study. Below we describe a method mentioned by for example Felfernig et al. (2011), Chmeiss et al. (2008) and Bruni (2005). The reason for choosing this method is its simplicity and the possibility of implementing it by a MILP solver.

Algorithm 3: Identification and removal of redundant edits

Input: Feasible edit set **E**
Output: Feasible edit set **E**, without redundant edits
1 FOR each Edit $e_i \in \mathbf{E}$ DO
2 $\quad \mathbf{E}^* \leftarrow \{\mathbf{E} \setminus e_i\} \cup \neg\, e_i$;
3 $\quad$ IF isFeasible($\mathbf{E}^*$) = FALSE THEN $\mathbf{E} \leftarrow \mathbf{E} \setminus e_i$
4 NEXT

When applied to previous example, the algorithm means that the negations of Edits 1, 2, 3, and 4 are added to the edit set one by one and that the feasibility is verified for all of the resulting set of rules. In this way, the redundancy of Edit 4 can be easily demonstrated.

Below a few words on the computation of negations. The negation of an equality constraints can be expressed as combination of two inequality constraints. For example, in previous example the negation of Edit 3, can be expressed as $T_1 + T_2 < T_3$ OR $T_1 + T_2 > T_3$. These two constraints are added to the three original rules one by one. Only if both additions lead to infeasible edit sets, one could conclude that Edit 3 is redundant. In our example, Edit 3 is however clearly not redundant.

The negation of a compound edit rule e_i, expressed as the disjunction $\cup_{j=1}^{D_i}\left(\left(\boldsymbol{a}_{ij}^{C}\right)^{T}\mathbf{x} \leq b_{ij}^{C}\right)$, is given by,

$$\bigcap_{j=1}^{D_i}\left(\left(\boldsymbol{a}_{ij}^{C}\right)^{T}\mathbf{x} > b_{ij}^{C}\right) \quad j = 1, \ldots, D_i,$$

a combination of D_i linear, unconditional constraints that all have to be satisfied.

6. Applications

Aim of this section is to apply constraints simplification methods on 'real-life' edit sets. We would like to show that these edit sets can actually be simplified. Moreover, we demonstrate that constraint simplification improves data editing's performance.

All applications were performed on a 32-bit Windows 7 laptop with a 2.80 GHz CPU and 3 Gigabyte of RAM memory. The methods from Sections 3–5, were implemented by R. The free LpSolveAPI was used as a MILP solver (Konis 2016) and the Editrules package (De Jonge and Van der Loo 2015) was implemented for automatic data editing. The following edit sets were considered:

1. Sales: Real-life edit set used for the 2012 Dutch Structural Business Statistics for sale of motor vehicles for businesses with fewer than ten employed persons;
2. Maintenance: Real-life edit set used for the 2012 Dutch Structural Business Statistics for maintenance of motor vehicles for businesses with fewer than ten employed persons;
3. Health-care: Edit set under development, meant to be used for a Dutch survey among welfare and childcare institutions;

All methods for constraint simplification in Sections 3–5 were applied to these three data sets. Automatic data editing was applied to the first two edit sets only, because of the lack of data for the third application.

Table 1. Results of three real-life applications.

	Sales	Maintenance	Health-care
Original edits			
Number of edits	115	119	196
-of which conditional:	26	29	114
Number of variables in edits	74	74	75
Simplification			
Fixed values	3	7	2
Conditional edits			
-Implicitly unsatisfiable components	1	1	4
-Implicitly satisfied components	1	1	3
Redundant edits	22	29	10
-of which conditional	7	13	3
Cleaned edits			
Number of edits	93	90	186
-of which conditional:	19	16	104
Computation Time (in seconds)	5	6	2,465

Table 2. Automatic data editing, original and simplified edit sets.

	Sales ($N = 614$ records)		Maintenance ($N = 197$ records)	
	Original edits	Simplified edits	Original edits	Simplified edits
Processed records*	613	613	197	197
Total time (in seconds)	2,639	2,039	479	217
Number records $<$ 10 sec	592	598	191	194

* = given a maximum computation time of ten seconds per record.

Firstly, Table 1 shows the feasibility of constraint simplification on a regular computer. One could note that computation time for the third application is relatively large, about 40 minutes, which can be explained from the many conditional rules. Large computation time is however not a problem, because edit rules simplification only needs to be conducted once, after designing or revising an edit set.

Secondly, Table 1 demonstrates that all simplification features in Sections 3–5 are useful, as each feature actually simplifies all three edit sets. The total number of edit rules is reduced by 5–20%; the number of conditional edits by 10–45%.

Table 2 shows that total computation time for automatic data editing is reduced by 23% for the Sales application and even by 55% for the Maintenance data set. The latter reduction can be largely attributed to only one record, whose computation times are 297 and 109 seconds for the original and simplified edit sets. This actually points out that the worst-case performance is important in data editing, but also shows that worst-case performance can be noticeably improved by rule simplification.

A practical solution to the possibly long computation time is to limit the available time for each record. The last row in Table 2 shows that edit rule simplification slightly increases the amount of records that are processed within ten seconds.

7. Discussion

Many works from the literature present automatic constraint simplification techniques that are able to greatly improve computational performances of large optimization problems. But, to the best knowledge of the author, these techniques are not often applied in the field of data editing.

This article shows that automated data editing can benefit from constraint simplification. A number of methods was presented for numerical data, based on MILP programming. Much attention was given to conditional IF-THEN rules that often occur in official statistics and that are particularly important for computational performance.

The feasibility of constraint simplification was demonstrated on a regular computer using freely available MILP solvers. It was shown that real-life edit sets can actually be simplified. As a result, the total computation time for localising erroneous values was reduced up to 55%; a reduction that can be mainly attributed to a few records with the

largest computation time. Hence, constraint simplification is an important step in further enhancing the practicality of automatic data editing.

Another benefit is that constraint simplification provides insight in the joint consequences of a set of rules. Manual inspection of automatically determined redundant rules and variables with a fixed value or finite bounds might reveal errors in rule formulation. Correction of these errors increases the quality of automated data editing and reduces the need for manual correction of automatically edited data.

A practical merit of the proposed methods is that simplification can be automated, out of sight of users, so that practitioners in the field do not have to bother about specifying constraints in a compact way.

This article implicitly assumed that edits are interconnected. However, if this is not the case, it is advisable to split an edit set $\mathbf{E}$ into disjunct sets, $\oplus_i \mathbf{E}_i$, such that $e_i \in \mathbf{E}_i$ and $e_j \in \mathbf{E}_j$ $(i \neq j)$ do not have any variable in common. Disjunct edit sets can be treated independently, which may improve performance of both data editing and edit rule simplification.

The simplification methods in this article have been designed for feasible edit sets. Despite that infeasible edit rules are useless for practical application, infeasible rules may occur in practise, for instance due to misspecification. In general, it can be hard to find the cause of a contradiction, especially if the number of edit rules is large. Therefore, most methods for dealing with inconsistency concentrate on isolating a smallest possible subset of inconsistent edit rules: a so-called irreducible inconsistent subset (IIS). Several algorithms for detecting IIS's are available from literature. The so-called "Deletion Filter" by Chinneck (1997) can be advised for many applications as it is easily understood, suitable for conditional "IF-THEN" edits and applicable for MILP programming. In a recent publication, Bruni and Bianchi (2012) proposed another, innovative approach, based on Farka's lemma. Their method however relies on an assumption, the so-called Integral Point property, that is unknown to be true for general applications.

A direction for further research is to introduce more constraint simplification techniques for data editing. In this article we considered numerical data. Methods for categorical data could be developed in the future.

8. References

Banff Support Team. 2008. *Functional Description of the BANFF System for Edit and Imputation*. Ottawa: Statistics Canada (Technical report).

Bertsimas, D. and J.N. Tsitsiklis. 1997. *Introduction to Linear Optimization*. Nashua: Athena Scientific.

Bruni, R. 2005. "Error Correction for Massive Data Sets." *Optimization Methods and Software* 20: 291–310. Doi: http://dx.doi.org/10.1080/10556780512331318281.

Bruni, R. and G. Bianchi. 2012. "A Formal Procedure for Finding Contradictions into a Set of Rules." *Applied Mathematical Sciences* 6: 6253–6271.

Chen, D., R.G. Batson, and Y. Dang. 2010. *Applied Integer Programming; Modelling and Solution*. Hoboken: John Wiley & Sons. Doi: http://dx.doi.org/10.1002/9781118166000.

Chinneck, J.W. 1997. "Finding a Useful Subset of Constraints for Analysis in an Infeasible Linear Program." *INFORMS Journal on Computing* 9: 164–174. Doi: http://dx.doi.org/10.1287/ijoc.9.2.164. Available at: http://www.sce.carleton.ca/faculty/chinneck/docs/UsefulSubset.pdf (accessed January 2017).

Chmeiss, A., V. Krawczyk, and L. Sais. 2008. "Redundancy in CSPs." In Proceedings of the 18th European Conference on Artificial Intelligence (ECAI 2008), August 21–25, 2008. Patras, Greece. Amsterdam: IOS Press. Doi: http://dx.doi.org/10.3233/978-1-58603-891-5-907.

De Jonge, E. and M. van der Loo. 2015. *Editrules: R Package for Parsing and Manipulating of Edit Rules and Error Localization*. R Package Version 2.9-0. Available at: http://cran.r-project.org/package=editrules (accessed May 2017).

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. New York: John Wiley & Sons. Doi: http://dx.doi.org/10.1002/9780470904848.

Dillig, I., T. Dillig, and A. Aiken. 2010. "Small Formulas for Large Programs: On-Line Constraint Simplification in Scalable Static Analysis." In Proceedings of the 17th international conference on Static analysis (SAS'10), September 14–16, 2010 Perpignan, France. Berlin Heidelberg: Springer-Verlag. Available at: http://theory.stanford.edu/~aiken/publications/papers/sas10.pdf (accessed January 2017).

Felfernig, A., C. Zehentner, and P. Blazek. 2011. "CoreDiag: Eliminating Redundancy in Constraint Sets." Proceedings of 22nd International Workshop on Principles of Diagnosis, October 4–7, 2011, Murnau, Germany. Available at: http://www.ist.tugraz.at/felfernig/images/stories/home/dx_corediag.pdf (accessed March 2017).

Fellegi, I.P. and D. Holt. 1976. "A Systematic Approach to Automatic Edit and Imputation." *Journal of the American Statistical Association* 71: 17–35. Doi: http://dx.doi.org/10.1080/01621459.1976.10481472.

Hooker, J. 2000. *Logic-Based Methods for Optimization: Combining Optimization and Constraint Satisfaction*. New York: John Wiley & Sons. Doi: http://dx.doi.org/10.1002/9781118033036.

Konis, K. 2016. *lpSolveAPI: R Interface for lpsolve. Version 5.5.2.0-17* R package version 5.5.2.0. Available at: https://cran.r-project.org/web/packages/lpSolveAPI/index.html (accessed January 2017).

Pannekoek, J., S. Scholtus, and M. van der Loo. 2013. "Automated and Manual Data Editing: a View on Process Design and Methodology." *Journal of Official Statistics* 29: 511–537. Doi: http://dx.doi.org/10.2478/jos-2013-0038.

Paulraj, S. and P. Sumathi. 2010. "A Comparative Study of Redundant Constraints Identification Methods in Linear Programming Problems." *Mathematical Problems in Engineering*. Article ID 723402. Doi: http://dx.doi.org/10.1155/2010/723402.

Piette, C. 2008. "Let the Solver Deal with Redundancy." In Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'08), November 3–5, 2008, Dayton, Ohio. Washington DC: IEEE Computer Society. Doi: http://dx.doi.org/10.1109/ICTAI.2008.38. Available at: https://hal.archives-ouvertes.fr/hal-00865304/document (accessed March 2017).

Telgen, J. 1983. "Identifying Redundant Constraints and Implicit Equalities in Systems of Linear Constraints." *Management Science* 29: 1209–1222. Doi: http://dx.doi.org/10.1287/mnsc.29.10.1209.

Received March 2016
Revised August 2017
Accepted September 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 41–54, http://dx.doi.org/10.1515/JOS-2018-0003

Design-Based Estimation with Record-Linked Administrative Files and a Clerical Review Sample

Abel Dasylva[1]

This article looks at the estimation of an association parameter between two variables in a finite population, when the variables are separately recorded in two population registers that are also imperfectly linked. The main problem is the occurrence of linkage errors that include bad links and missing links. A methodology is proposed when clerical-reviews may reliably determine the match status of a record-pair, for example using names, demographic and address information. It features clerical-reviews on a probability sample of pairs and regression estimators that are assisted by a statistical model of comparison outcomes in a pair. Like other regression estimators, this estimator is design-consistent regardless of the model validity. It is also more efficient when the model holds.

Key words: Probabilistic record-linkage; administrative data; clerical-review; mixture-model; probability sample.

1. Introduction

Computerized record-linkage aims at linking records that relate to the same individual or entity, with minimal human intervention. Such records are called matched records. In many cases, this is a challenging task because it must be based on pseudo-identifiers such as names and demographic characteristics, which are non-unique and possibly recorded with spelling variations or typographical errors. These limitations lead to errors that include bad links and missing links.

In general the computerized linkage of two large files comprise of five major steps. First, the linkage variables are parsed and standardized. Second, records in the two files are compared using blocking keys. Only the pairs that agree on some blocking key are subsequently compared more extensively. Third, the linkage variables are extensively compared to produce comparison outcomes. Fourth, a decision is made for each pair. Finally, conflicting linkage decisions are dealt with, such as when linking the same census record to two death records. Linkage methodologies differ according to how linkage decisions are made in the fourth step. In a deterministic linkage, the decision may be based on arbitrary criteria, according to subject matter knowledge. In probabilistic linkage, the decision is based on a linkage weight which is a measure of the similarity between two records. This weight is typically the sum of outcome weights that correspond to the similarity of the different linkage variables. For the final decision, a pair linkage weight is

[1] Statistics Canada, SRID, 100 Tunney's Pasture, Ottawa, Ontario K1A0T6, Canada. Email: abel.dasylva@statcan.gc.ca

compared to one or two thresholds to determine whether it should be linked, reviewed clerically or rejected (Fellegi and Sunter 1969). The overall linkage performance is characterized by the rates of linkage errors, which are determined by the linkage weights and thresholds.

Two files may be linked to study the association between variables that have been recorded separately in each file. For example, consecutive censuses may be linked to create a longitudinal dataset. In this case, the variables of interest measure the same characteristic at different time points. Estimation with an imperfectly linked dataset is challenging because linkage errors must be accurately measured and accounted for (Lahiri and Larsen 2005; Chipperfield et al. 2011; Chambers 2009). The measurement may be based on a statistical model, clerical-reviews or both.

In theory, linkage errors may be estimated from a model, without any human intervention. On one hand, Fellegi and Sunter (1969) have suggested models based on the assumption that the linkage variables are conditionally independent given the match status. However, these models have been quite inaccurate (Belin and Rubin 1995). On the other hand, models that incorporate interactions may lack the identification property, see Kim (1984) and more recently Fienberg et al. (2009). The above difficulties justify the continued use of clerical-reviews or training samples (Belin and Rubin 1995; Howe 1981; Heasman 2014; Gill 2001; Guiver 2011), possibly in conjunction with a statistical model (Larsen and Rubin 2001).

In this work, the problem that consists in estimating an association parameter from an imperfectly linked dataset is framed as a survey sampling problem. In general, survey sampling aims at estimating a finite population parameter without bias, by taking a probability sample, where each population unit has a known and positive inclusion probability. Using such a sample, a population total is estimated without bias with an Horwitz-Thompson (HT) estimator; the sum of sample values weighted by the corresponding reciprocal selection probability. However, the HT estimator may have a large variance, especially when the inclusion probability is not correlated with the variable of interest. A popular alternative is a regression estimator when some auxiliary variables are observed for all population units. The regression estimator is not unbiased but design-consistent, that is, with a bias that is negligibly small in large samples. This estimator also has a smaller variance than the HT estimator, when the variable of interest is a nearly linear function of the auxiliary variables. Regression estimators offer examples of generalized regression estimators (GREG) and calibration estimators that have been thoroughly studied by Särndal et al. (1992) and by Deville and Särndal (1992). These estimators are also referred to as model-assisted estimators because they are inspired by some implicit statistical model; typically a linear model relating the auxiliary variables to the variables of interest. They are efficient when the model holds and less so otherwise. However they remain design-consistent regardless of the model validity (see Särndal et al. 1992, section 6.7, pp. 239).

The proposed problem formulation brings questions that have been already addressed by Särndal et al. (1992) and others, about optimal sampling designs, design-consistent estimators and the efficient use of auxiliary information through statistical models. This body of work is applied to our problem with some adaptation. The resulting estimators are regression estimators, that are built in two steps. First, all record-pairs that satisfy blocking

criteria are used to fit a model for predicting the match status of pairs within the blocks, irrespective of whether they are part of the clerical sample. Second, a regression estimator is fitted based on the clerical data. The described framework also applies when the match status is determined by other means than clerical reviews, for example through limited access to unique identifiers or additional information from a third party.

The following sections are organized as follows. Section 2 presents the notation and background. Section 3 describes model-based estimators in the record-linkage context. Section 4 discusses sampling designs. Section 5 presents simulation results. Section 6 presents the conclusions and future work.

2. Notation and Background

Consider two duplicate-free registers A and B, which contain records about N individuals. Register A contains K linkage variables and the variable of interest x_i for the ith record in A. Register B contains the same linkage variables as A and the variable of interest y_j for the jth record in B. Let U denote the finite population of all N^2 record-pairs in the cartesian product of the two files, that is, of all pairs (i,j) where $1 \leq i,j \leq N$.

For the record-pair (i,j) in the Cartesian product of the two registers, the linkage variables may be compared to produce a K-tuple $\boldsymbol{\gamma}_{ij} = \left(\gamma_{ij}^{(1)}, \ldots, \gamma_{ij}^{(K)}\right)$ of comparison outcomes, also called vector of comparison outcomes. In large files, some linkage variables are also coarsely compared to define blocks that altogether represent a small subset U^* of U and yet contain most matched pairs. The subset U^* of blocked pairs is the union of B disjoint subsets, $U_1^* \ldots U_B^*$, where each subset represents a distinct block. For each pair, this blocking information is also included in the comparison vector $\boldsymbol{\gamma}_{ij}$. The comparison vector $\boldsymbol{\gamma}_{ij}$ provides the basis for linking the records, for example using Fellegi and Sunter (1969) optimal linkage rule. However such a linkage is not required in the proposed estimation methodology.

Let M_{ij} denote the indicator variable that is set to 1 if the pair (i,j) is matched, that is, associated with the same individual. The variable M_{ij} is also called the match status of the pair (i,j). The comparison vector $\boldsymbol{\gamma}_{ij}$ is crucial for making an inference $\widehat{M}_{ij}$ about the unknown match status M_{ij}. The inferred match status $\widehat{M}_{ij}$ can take many forms. For example, it can be set to the conditional or posterior match probability $P\left(M_{ij} = 1 \middle| \boldsymbol{\gamma}_{ij}\right)$ given the comparison vector. It can also be interpreted as the "weight-share" of the pair (i,j), with the meaning of the Generalized Weight Share Method. See Lavallée (2002, chap. 9) for applications of this method to record-linkage.

For finite population inference, the goal is estimating a total of the following form:

$$Z = \sum_{(i,j) \in U} M_{ij} z_{ij} \tag{1}$$

In the above expression, $z_{ij} = f(x_i, y_j)$ and f is some known function.

For model-based inference, assume that the record-generating individuals represent an Independent Identically Distributed (IID) sample according to some distribution or superpopulation depending on a parameter θ. Inference about this parameter may be based on an equation of the form $E\left[S(\theta; x, y)\right] = 0$, where S is a score function (e.g., a log-likelihood), while (x,y) is the observation associated with an individual from the

superpopulation. The parameter θ may be estimated through the following unbiased estimating equation where $z_{ij}(\theta) = S(\theta; x_i, y_j)$.

$$\sum_{(i,j)\in U} M_{ij} z_{ij}(\hat{\theta}) = 0 \tag{2}$$

In both cases, the inferences use the recorded values of the variables in matched pairs, regardless of whether these values are free of nonsampling errors such as typographical errors, measurement errors, etc.

Resources for error-free clerical reviews are available to measure the match status. However they are costly and must be minimized. The clerical sample s has a fixed size. It is split into a blocking stratum U^* and a nonblocking stratum $U \setminus U^*$. Let s^* denote the sample of blocked pairs in the clerical sample. The samples in the different strata are selected independently and their sampling designs are arbitrary. Let π_{ij} denote the first-order sample inclusion probability for the record-pair (i,j).

3. Model-Assisted Estimators

The proposed estimators are regression estimators (Särndal et al. 1992, chap. 6) that have the following general difference form:

$$\hat{Z} = \underbrace{\sum_{(i,j)\in U^*} \widehat{M}_{ij} z_{ij} + \sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} \left(M_{ij} - \widehat{M}_{ij} \right)}_{(1)} + \underbrace{\sum_{(i,j)\in s\setminus s^*} \pi_{ij}^{-1} M_{ij} z_{ij}}_{(2)} \tag{3}$$

This estimator is the sum of contributions from the two strata. The first contribution exploits the inferred match status to estimate the total over the blocking stratum with a greater precision. The second contribution is simply a Horwitz-Thompson estimator for the total over the nonblocking stratum. The above estimator may be viewed as a calibration estimator (Deville and Särndal 1992), where the estimated total is calibrated to the corresponding total based on inferred match status. It estimates the total with no sampling error and no bias when the following two conditions are met:

i. Perfect blocking criteria selecting all matched pairs.
ii. Perfect inference of the match status, that is, $M_{ij} = \widehat{M}_{ij}$.

The estimator is also unbiased if the inferred status ignores the information of the clerical sample:

$$E\left[\hat{Z}|U\right] = \sum_{(i,j)\in U} M_{ij} z_{ij} = Z \tag{4}$$

This is the case if $\widehat{M}_{ij}$ is only a function of z_{ij} and γ_{ij}. The inferred status may be set to the conditional match probability given the vector of comparison outcomes and the variables x_i, y_j, that is,

$$\widehat{M}_{ij} = P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}\right) \tag{5}$$

This particular inference strategy would minimize the mean squared error (over the super population) between the predicted total $\sum_{(i,j)\in U^*} \widehat{\boldsymbol{M}}_{ij} z_{ij}$ and the actual total $\sum_{(i,j)\in U^*} \boldsymbol{M}_{ij} z_{ij}$ over the blocking stratum, among all inference strategies where $\widehat{\boldsymbol{M}}_{ij}$ is only a function of x_i, y_j and $\boldsymbol{\gamma}_{ij}$, if the record-pairs were IID. Under a Simple Random Sampling (SRS) design in the blocking stratum, the resulting estimator would also be more efficient than the Horwitz-Thompson estimator, if the pairs were IID.

The conditional match probability may be estimated under the assumption of IID pairs according to a two-component mixture distribution, where the different comparison outcomes and the variables x_i, y_j are assumed conditionally independent given the match status, where $\tau = 0, 1$:

$$P\left(x_i, y_j, \boldsymbol{\gamma}_{ij} \middle| M_{ij} = \tau\right) \;=\; P\left(x_i, y_j \middle| M_{ij} = \tau\right) \prod_{k=1}^{K} P\left(\boldsymbol{\gamma}_{ij}^{(k)} \middle| M_{ij} = \tau\right) \tag{6}$$

The parameters $\boldsymbol{\psi}$ of this mixture include the mixing proportion $\lambda = P\left(M_{ij} = 1\right)$, the marginal m-probabilities $P\left(x_i, y_j \middle| M_{ij} = 1\right)$ and $P(\boldsymbol{\gamma}_{ij}^{(k)} \middle| M_{ij} = 1)$, and the marginal u-probabilities $P\left(x_i, y_j \middle| M_{ij} = 0\right)$ and $P(\boldsymbol{\gamma}_{ij}^{(k)} \middle| M_{ij} = 0)$, under the assumption of IID pairs. They may be estimated with an Expectation-Maximization (E-M) algorithm. See Jaro (1989) or Winkler (1988) for applications of E-M to record-linkage, and Dempster et al. (1977) for a general reference on E-M. An important feature of this mixture model is the use of x_i and y_j as additional linkage variables. The mixture model becomes simpler when the variables x_i and y_j are highly correlated with the linkage variables. In this case x_i and y_j bring no new information about the match status, given $\boldsymbol{\gamma}_{ij}$. Mathematically, this is expressed by the conditional independence of $\left(x_i, y_j\right)$ and the match status given the comparison outcomes:

$$P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}\right) \;=\; P\left(M_{ij} = 1 \middle| \boldsymbol{\gamma}_{ij}\right) \tag{7}$$

The inference strategy may be inefficient if the assumed mixture model does not hold. For example, this problem may occur if the couple $\left(x_i, y_j\right)$ contains additional information about the match status, but the inference $\widehat{\boldsymbol{M}}_{ij} = P\left(M_{ij} = 1 \middle| \boldsymbol{\gamma}_{ij}\right)$ is used instead. The estimator is also less efficient if the linkage variables are correlated but their conditional independence is assumed.

Let $P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)$ denote a preliminary estimate of the conditional match probability according to the mixture model. This estimate is computed in the E-Step of the E-M algorithm and it does not use the clerical results. In most cases, this mixture model will estimate the conditional match probability with some bias even if it accounts for some of the interactions among the different variables. To adjust for this bias, the match status may be inferred using a linear function $\beta_0 + \beta_1 P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)$ of the estimated conditional probability, where the regression coefficients β_0 and β_1 are estimated from the clerical sample. In this case, the inferred match status is computed as follows:

$$\widehat{\boldsymbol{M}}_{ij} \;=\; \hat{\beta}_0 + \hat{\beta}_1 P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right) \tag{8}$$

A special case is when a ratio estimator estimates the total over the blocking stratum. That is,

$$\hat{Z} = \frac{\sum_{(i,j)\in U^*} z_{ij} P\left(M_{ij}=1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)}{\sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} P\left(M_{ij}=1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)} \sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} M_{ij} + \sum_{(i,j)\in s\setminus s^*} \pi_{ij}^{-1} M_{ij} z_{ij} \tag{9}$$

In this case $\hat{\beta}_0 = 0$ and $\hat{\beta}_1$ is computed as follows:

$$\hat{\beta}_1 = \frac{\sum_{(i,j)\in U^*} z_{ij} P\left(M_{ij}=1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)}{\sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} P\left(M_{ij}=1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)} \tag{10}$$

The estimator can also be written in terms of uniform g-weights $\left[g_{ij}\right]_{ij}$, where $g_{ij} = \hat{\beta}_1$:

$$\hat{Z} = \sum_{(i,j)\in s^*} g_{ij} \pi_{ij}^{-1} z_{ij} M_{ij} + \sum_{(i,j)\in s\setminus s^*} \pi_{ij}^{-1} M_{ij} z_{ij} \tag{11}$$

The following model provides the basis for better weighted least squares estimators:

$$E\left[M_{ij} \middle| x_i, y_j, \gamma_{ij}\right] = \beta_0 + \beta_1 P\left(M_{ij}=1 \middle| x_i, y_j, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right) \tag{12}$$

$$var\left(M_{ij} \middle| x_i, y_j, \gamma_{ij}\right) \propto P\left(M_{ij}=1 \middle| z_{ij}, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right)\left[1 - P\left(M_{ij}=1 \middle| x_i, y_j, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right)\right] \tag{13}$$

In this case, the estimated regression coefficients minimize the following quadratic function:

$$Q\left(\beta_0, \beta_1; \hat{\boldsymbol{\psi}}\right) = \sum_{(i,j)\in s^*} \frac{\pi_{ij}^{-1}\left[M_{ij} - \beta_0 + \beta_1 P\left(M_{ij}=1 \middle| x_i, y_j, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right)\right]^2}{P\left(M_{ij}=1 \middle| x_i, y_j, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right)\left[1 - P\left(M_{ij}=1 \middle| x_i, y_j, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right)\right]} \tag{14}$$

The resulting estimator may be written in terms of nonuniform g-weights incorporating the inferred match status. This estimator is improved by fine tuning the variance structure with Generalized Estimating Equations (Jiang 2007).

The proposed estimators are no longer unbiased because the clerical review results are used to make inferences about the pairs match status. However, like other regression estimators (Särndal et al. 1992, Result 6.6.1, pp. 235, section, 6.7, pp. 238), they are design-consistent regardless of the assumed models.

4. Sampling Design

Model-based stratified sampling has been used to approximately minimize the variance of regression estimators (Särndal et al. 1992). In this design, the strata are defined by the variance of the error in the assumed linear model. This strategy also applies to the current

context where a single total is estimated. To be specific, the design-based variance $var\left(\hat{Z}|U\right)$ of the model-assisted estimator is the sum of two terms:

$$\begin{aligned} var(\hat{Z}|U) \;=\; & var\left(\sum_{(i,j)\in s^*}\pi_{ij}^{-1}z_{ij}\left(M_{ij}-\widehat{M}_{ij}\right)\middle|U\right) \\ & + var\left(\sum_{(i,j)\in s\setminus s^*}\pi_{ij}^{-1}M_{ij}z_{ij}\middle|U\right) \end{aligned} \tag{15}$$

The first term is approximately minimized by a Neyman allocation where the pairs are stratified according to the model-based conditional variance of the error $e_{ij}=z_{ij}\left(M_{ij}-\widehat{M}_{ij}\right)$, that is $var\left(e_{ij}\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right)$. This conditional variance is given by the following expression.

$$\begin{aligned} var\left(e_{ij}\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right) \;&=\; var\left(z_{ij}\left(M_{ij}-\widehat{M}_{ij}\right)\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right) \\ &=\; z_{ij}^2 var\left(M_{ij}-\widehat{M}_{ij}\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right) \\ &=\; z_{ij}^2\Big(var\left(M_{ij}\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right) \\ &\quad +\left[\widehat{M}_{ij}-P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right)\right]^2\Big) \\ &=\; z_{ij}^2\Big(P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right)\left[1-P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right)\right] \\ &\quad +\left[\widehat{M}_{ij}-P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right)\right]^2\Big) \end{aligned} \tag{16}$$

With known conditional match probabilities $P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right)$ and the best possible inference $\widehat{M}_{ij}=P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right)$ we have

$$var\left(e_{ij}\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right) \;=\; z_{ij}^2\widehat{M}_{ij}\left(1-\widehat{M}_{ij}\right) \tag{17}$$

Suppose that the pairs are stratified based on $\boldsymbol{\gamma}_{ij}$ and $\left(x_i,y_j\right)$ or some fine discrete approximation if these variables are continuous. Note that by design, in such as stratum, the pairs have the same $z_{ij}=z$ value and an identical conditional match probability $P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right)=p$. Thus they are identically distributed according to $zBernoulli\left(p\right)$. If these pairs were independent, the variance of the errors e_{ij} would be well approximated by the common variance $var\left(e_{ij}\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right)=z^2p\left(1-p\right)$, based on the Law of Large Numbers (LLN). In the corresponding Neyman allocation, the sample size is proportional to the stratum variance. An estimator with the same minimum variance is obtained via a Neyman allocation, where the strata are based on $z_{ij}^2\widehat{M}_{ij}\left(1-\widehat{M}_{ij}\right)$ the estimated conditional error variance. The resulting allocation is no longer optimal when the conditional match probability $P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij}\right)$ is estimated with some bias. Let $\hat{p}$ denote the corresponding stratum estimate. In this case the stratum variance is increased to $z^2p\left(1-p\right)+\left(\hat{p}-p\right)^2$.

5. Simulations

The proposed estimators are evaluated in six scenarios that consider different factors, including the discriminating power of the linkage variables, the sample size, the model for the distribution of linkage errors, clerical errors, and the correlation among the pairs. All the scenarios consider a one-to-one linkage between two registers. In each register the records are partitioned into perfect blocks of equal sizes. Consequently two matched records always fall within the same block.

The different scenarios account for different features of practical linkages.

Scenario 1 emulates the process by which administrative records may be generated from a finite population of individuals, with correlations among pairs that are within the same block. It considers seven binary linkage variables that have conditionally independent typographical errors with a common distribution. This distribution is given by the following transition probabilities:

$$P\left(c_i^{(k)}, c_j^{(k)} \middle| \zeta_i^{(k)}, M_{ij} = 1\right) = P\left(c_i^{(k)} \middle| \zeta_i^{(k)}\right) P\left(c_j^{(k)} \middle| \zeta_i^{(k)}\right) \tag{18}$$

$$P\left(c_i^{(k)} \middle| \zeta_i^{(k)}\right) = (1-\alpha) I\left(c_i^{(k)} = \zeta_i^{(k)}\right) + \alpha I\left(c_i^{(k)} \neq \zeta_i^{(k)}\right) \tag{19}$$

where α is the probability of a recording error.

In the above expressions, $c_i^{(k)}$ is the k-th linkage variable for record i in register A, $\zeta_i^{(k)}$ is the latent true (i.e., free of recording errors) value of the variable for the associated individual, with $c_j^{(k)}$ and $\zeta_j^{(k)}$ denoting the corresponding variables in register B. Note that, by definition $\zeta_i^{(k)} = \zeta_j^{(k)}$ in a matched pair (i,j). For each record i, the latent variables $\zeta_i^{(k)}$ are IID. The comparison outcomes are based on exact comparisons with $\gamma_{ij}^{(k)} = I\left(c_i^{(k)} = c_j^{(k)}\right)$.

The variables of interest x_i and y_j are also binary and mutually independent of the linkage variables in each register, and each matched pair. The files are linked to study the joint distribution of these two variables, that is, to estimate the frequencies of the different cells in a two-way contingency table. In this case $z_{ij} = I\left(\left(x_i, y_j\right) = \left(x, y\right)\right)$ where $x, y = 0, 1$. This setup is similar to that described by Chipperfield et al. (2011). However, the goal here is finite population inference on a single finite population.

From the finite population, different IID samples are drawn using one of two designs. For each resulting sample, three estimators are computed for the number of matched pairs in each cell of the two-way contingency table. They include the H-T estimator, a second model-assisted estimator (hereafter simply referred to as 2nd estimator) using the inference $\widehat{\boldsymbol{M}}_{ij} = P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)$ and a third estimator (hereafter simply referred to as 3rd estimator) using the inference $\widehat{\boldsymbol{M}}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)$.

The first sample design is stratified according to the x-y value pairs. In each stratum, a fixed size SRS sample is drawn. The second sample design is also stratified based on the x-y value pairs, but it uses substrata, which are based on the conditional variance of the prediction error. Each x-y stratum has the same number of substrata but the boundaries are selected to obtain nearly equal substrata sizes, after the pairs are sorted according to their conditional variance in each stratum. Consequently, substrata boundaries may differ

from an x-y stratum to the next. The same x-y stratum sample size is used as in the first design. However in the second sample design, this sample size is allocated optimally among the substrata using a Neyman allocation, where the estimated variance of a substratum is estimated as the mean conditional error variance over all the corresponding pairs. A substratum sample allocation is further constrained to have at least two units and not to exceed the substratum size.

Scenario 1 is the baseline scenario. It evaluates the two model-assisted estimators in the best case, with the correct model for the comparison outcomes. This situation maximizes their relative advantage over the naïve H-T estimator. Scenarios 2 through 5 are built after Scenario 1, that is, with correlated pairs. However they each incorporate a slight modification. Scenario 2 considers linkage variables with more typographical errors and hence less discriminating power than in Scenario 1. Scenario 3 considers a smaller (1,000 pairs instead of 4,000 pairs) clerical-review sample. Scenario 4 considers linkage variables that are not conditionally independent by correlating the latent variables $\zeta_i^{(k)}$. This correlation is produced by generating the $\zeta_i^{(k)}$'s according to a mixture model with conditional independence based on a binary latent class ξ_i. However the estimated conditional match probability $P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)$ is estimated under the assumption of conditional independence among all linkage variables. Scenario 5 considers clerical errors.

Scenario 6 considers agreement frequencies for variables such as names and birthdate that have been used for linking high quality person files. The specific frequencies are based on an example provided by Newcombe (1988, Table 5.1). Unlike the other scenarios, Scenario 6 considers pairs with IID and conditionally independent comparison vectors.

The simulation parameters are as follows. All scenarios are based on $N = 10,000$ individuals, 1,000 blocks, a block size of 10, $K = 7$ linkage variables, $P(x = 1) = 0.5$, $P\left(y = 1|x = 0\right) = 0.4$, $P\left(y = 1|x = 1\right) = 0.7$, 10 substrata per x-y stratum, 100 E-M iteration and 100 repetitions.

The x-y stratum sample size is set to 1,000 for all scenarios except for Scenario 3 (smaller clerical sample), where it is set to 100. The conditional agreement probabilities are uniform across the linkage variables in Scenarios 1 through 5. However, they vary across these scenarios. For Scenarios 1 and 3 through 5, the conditional probability of agreement is 0.98 for a matched pair and 0.5 for an unmatched pair. For Scenario 2, these conditional probabilities are respectively 0.82 and 0.5. For Scenario 6, the conditional agreement probabilities are given in Table 1. The remaining parameters only apply to Scenarios 1 through 5 and are set as follows. The parameter α is set to 0.1 for Scenarios 1 through 5. For the intrinsic variables, the probability $P\left(\zeta_i^{(k)} = 1\right)$ is uniformly set to 0.5. For the recording errors, the probability $P\left(c_i^{(k)} = 1 \middle| \zeta_i^{(k)} = 0\right)$ is set to 0.01 except for Scenario 2 (weaker linkage variables), where it is set to 0.1. As for the probability $P\left(c_i^{(k)} = 1 \middle| \zeta_i^{(k)} = 1\right)$ is set to 0.99 except for Scenario 2, where it is set to 0.9. Scenario 4 (misspecified case) involves the additional parameters that are set as follows. The probability $P(\xi_i = 1)$ is set to 0.5, while the conditional probabilities $P\left(\zeta_i^{(k)} = 1|\xi_i = 0\right)$ and $P\left(\zeta_i^{(k)} = 1|\xi_i = 1\right)$ are respectively set to 0.3 and 0.7.

For cell (0,0), the results for the H-T estimator and the second estimator are shown in the box plots of Figures 1 and 2, and in Tables 2 and 3. The box plots show the five-number summary of the relative bias for the estimated cell count. In these figures, the horizontal axis

Table 1. Agreement frequencies for Scenario 6 based on Newcombe (1988, Table 5.1).

	Agreement probability	
Linkage variable	Matched	Unmatched
Surname	0.965	0.001
First name	0.79	0.009
Middle initial	0.888	0.075
Year of birth	0.773	0.011
Month of birth	0.933	0.083
Day of birth	0.851	0.033
Province/country of birth	0.981	0.117

indicates the estimator (1 for the H-T estimator, or 2 for the second estimator), the sampling design (1 or 2) and the scenario (1 through 3 in Figure 1, and 4 through 6 in Figure 2). For example, in Figure 1, 2.1.1 corresponds to the box plot for the second estimator under the first scenario and the first design. As for Tables 2 and 3 they show the average bias and CV of the estimated count for cell (0,0). The results for the other cells are not shown because they are similar to those of cell (0,0). As for the third estimator, the corresponding results are not shown because they are similar to those of the second estimator.

For Scenario 1 (our baseline), all three estimators have a very small relative bias, with no clear advantage for the H-T estimator under either sampling design. However the model-assisted estimator halves the CV of the H-T estimator, under the first sampling design. The gain in precision becomes negligible under the second sampling design. This is expected because the model information is already exploited through the stratification, which also benefits the H-T estimator.

The results for Scenario 2 show a worse performance for the model-assisted estimator, when the linkage variables are less discriminating. Indeed, the corresponding absolute relative bias is larger than that of the H-T estimator, under either sampling design. As for

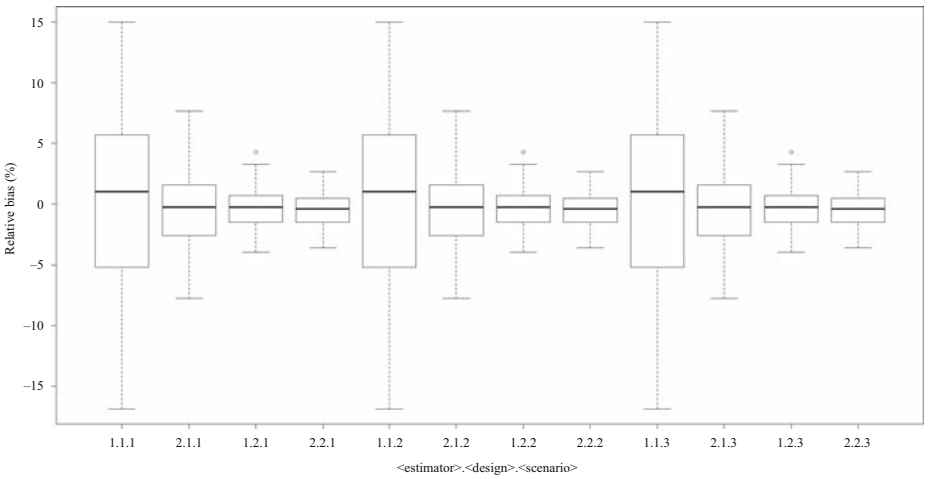


Fig. 1. Box plots of the relative bias for cell (0,0) in Scenarios 1 through 3. Estimator 1 is the HT estimator. Estimator 2 is the 2nd estimator.

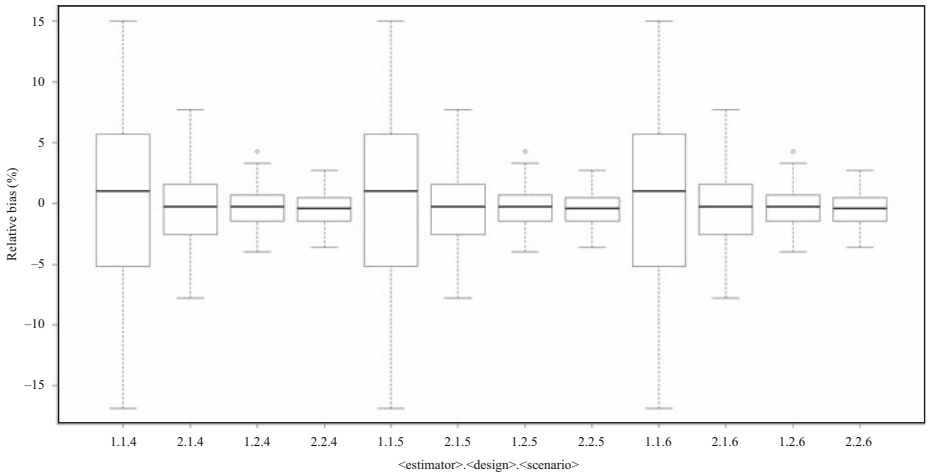


Fig. 2. Box plots of the relative bias for cell (0,0) in Scenarios 4 through 6. Estimator 1 is the HT estimator. Estimator 2 is the 2nd estimator.

the expected gain in precision under the first sampling design, it is dramatically smaller than in Scenario 1. Under the second design, the gain is negligible.

The results for Scenario 3 show the same trends as in Scenario 1, with similar gains in precision for the model-assisted estimator. Intuitively the use of a model partially makes up for the reduced sample size.

For Scenario 4, where the model is misspecified, both the H-T estimator and the model-assisted estimator have a small relative bias, under either design. For the model-assisted estimator, the gain in precision is slightly reduced compared to Scenario 1.

In Scenario 5, with clerical errors, Table 2 shows that the relative bias of all the estimators is significantly increased compared to Scenario 1. However, under the first sampling design, the model-assisted estimators offer a significant advantage over the HT estimator, even if this advantage is smaller than in Scenario 1. Under the second design, this gain in precision vanishes and all the estimators have much less precision than in the first sampling design.

Table 2. Relative bias and CV for cell (0,0) for Scenarios 1 through 3.

Scenario	Design	Estimator	Relative bias (%)	CV (%)
1	1	1	−0.12	7.52
		2	0.45	3.33
	2	1	0.34	1.52
		2	0.48	1.36
2	1	1	0.77	7.62
		2	0.94	6.43
	2	1	−0.17	5.67
		2	−0.29	5.44
3	1	1	0.18	25.18
		2	0.11	12.57
	2	1	0.32	6.79
		2	−0.04	6.37

Table 3. Relative bias and CV for cell (0,0) for Scenarios 4 through 6.

Scenario	Design	Estimator	Relative bias (%)	CV (%)
4	1	1	1.21	7.71
		2	0.62	4.22
	2	1	0.25	2.40
		2	0.21	2.29
5	1	1	−4.94	8.25
		2	−5.25	3.66
	2	1	−6.31	14.84
		2	−6.23	14.79
6	1	1	−0.79	7.40
		2	−0.10	0.48
	2	1	−0.01	0.82
		2	0.01	0.12

In Scenario 6, the model-assisted estimator greatly outperforms the H-T estimator both regarding the bias and the precision, under either sampling design. The gain in precision is also dramatically larger than in the other scenarios. This is because in Scenario 6, the linkage variables collectively provide much more discrimination than in the previous scenarios. The combination of this discrimination with a correct statistical model produces the observed gains.

Overall, the model-assisted estimators offer the best performance when the following three conditions are met:

i. The linkage variables provide a high discrimination.
ii. The clerical-reviews are very reliable.
iii. The assumed statistical model is correct.

Of the above three conditions, the reliability of the clerical-review is the most critical one as it may be expected.

The simulation results also shed some light on the choice of the sampling design. In all scenarios without clerical errors, the precision is much greater under the second sampling design, where the pairs are stratified according to the estimated conditional match probability. This result further underscores the importance of using auxiliary variables that leverage the comparison outcomes.

Although this work considers a one-to-one linkage, this assumption does not play a major role in the estimation procedure. Hence the proposed methodology also applies to an incomplete linkage so long as the clerical reviews remain error-free. However the resulting model-assisted estimators may be less efficient if the unmatchable records greatly differ in distribution from the other records. Then the pairs outcomes are better modeled by a three component mixture including two classes of unmatched pairs. In this case, specifying a good model may be more challenging.

6. Conclusions and Future Work

This study casts the problem of design-based estimation with linked administrative files in the classical survey methodology framework. It also proposes a new estimation

methodology based on model-assisted estimators and sampling-designs that are evaluated through simulations. The simulations clearly demonstrate the equal importance of auxiliary variables based on the linking variables and high quality clerical reviews. Specifying good models is also important for the efficiency of the resulting estimators. However using the correct model is not required, because, like previous model-assisted estimators (Särndal et al. 1992), the proposed estimators remain design consistent even when the model is misspecified.

There are two potential issues with clerical reviews including the quality of the supporting information and the quality of the review process. Meaningful clerical reviews are obviously impossible unless the supporting information is sufficient and reliable. Even when it is the case, many questions remain about the quality of the review process and ways to objectively measure it. Indeed there are few studies on this subject, beyond that by Newcombe et al. (1983). Furthermore, such studies may be hard to replicate, either because they have not disclosed important methodological details, or because their results are heavily dependent on the used datasets that are unavailable. A second challenge is the development of anonymization techniques. They prevent clerical reviews and adversely impact the linking efficacy. Solutions based on privacy-preserving record linkage are being actively researched to address these problems (Schnell et al. 2009). However, in situations where clerical reviews have been effective (e.g., with available names, birthdates and addresses in the original files), it is still unclear whether these solutions offer competitive privacy-preserving alternatives to clerical reviews. A third challenge concerns missing values in the linked files. The problem arises because clerical reviews are expensive, such that it is desirable to avoid sampling pairs where some variables of interest are missing. Such missing variables represent an unusual form of item nonresponse, because it is known prior to sample selection. Devising solutions for an optimal sample selection represents a new and promising avenue of research.

7. References

Belin, T.R. and D.B. Rubin. 1995. "A Method for Calibrating False-Match Rates in Record Linkage." *Journal of the American Statistical Association* 90: 694–707. Doi: http://dx.doi.org/10.2307/2291082.

Chambers, R. 2009. "Regression Analysis of Probability-Linked Data." *Official Statistics Research Series*, vol. 4.

Chipperfield, J.O., G.R. Bishop, and P. Campbell. 2011. "Maximum Likelihood Estimation for Contingency Tables and Logistic Regression with Incorrectly Linked Data." *Survey Methodology* 37: 13–24.

Dempster, A., N. Laird, and D. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society Series B* 39: 1–38. Available at: http://www.jstor.org/stable/2984875 (accessed November 2017).

Deville, J.-C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association* 37: 376–382.

Fellegi, I.P. and A.B. Sunter. 1969. "A Theory of Record Linkage." *Journal of the American Statistical Association* 64: 1183–1210.

Fienberg, S., P. Hersh, A. Rinaldo, and Y. Zhou. 2009. "Maximum Likelihood in Latent Class Models for Contingency Table Data." In *Algebraic and Geometric Methods in Statistics*, edited by Paolo Giblisco, Eva Riccomagno, Maria Piera Rogantin, and Henry P. Wynn, 7–62. New York: Cambridge University Press.

Gill, L. 2001. *Methods for Automatic Record Matching and Linkage and their Use in National Statistics*. London: Office of National Statistics.

Guiver, T. 2011. *Sampling-Based Clerical Review Methods in Probabilistic Linking*. Canberra: Australia Bureau of Statistics.

Heasman, D. 2014. "Sampling a Matching Project to Establish the Linking Quality." *Survey Methodology Bulletin* 72: 1–16.

Howe, G.R. 1981. "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-Up Studies." *Computers and Biomedical Research* 14: 327–340.

Jaro, M.A. 1989. "Advances in Record Linkage Methodology to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association* 84: 414–420.

Jiang, J. 2007. *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.

Kim, B.S. 1984. *Studies of Multinomial Mixture Models*. PhD thesis, Chapel Hill: University of North Carolina.

Lahiri, P. and D. Larsen. 2005. "Regression Analysis with Linked Data." *Journal of the American Statistical Association* 100: 222–227. Available at: http://www.jstor.org/stable/27590532 (accessed November 14, 2017).

Larsen, M. and D. Rubin. 2001. "Iterated Automated Record Linkage Using Mixture Models." *Journal of the American Statistical Association* 96: 32–41.

Lavallée, P. 2002. *Le Sondage indirect ou la méthode du partage des poids*. Bruxelles: Éditions de l'Université de Bruxelles.

Newcombe, H.B., M.E. Smith, and G.R. Howe. 1983. "Reliability of Computerized Versus Manual Death Searches in a Study of the Health of Eldorado Uranium Workers." *Computers in Biology and Medicine* 13: 157–169.

Newcombe, H. 1988. *Handbook of Record Linkage*. New-York: Oxford Medical Publications.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New-York: Springer.

Schnell, R., T. Bachteler, and J. Reiher. 2009. "Privacy-Preserving Record Linkage using Bloom Filters". *BioMed Central Medical Informatics and Decision Making*, 9.

Winkler, W.E. 1988. "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage". In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 22–25, 1988, New Orleans, Louisiana. 667–671.

Received July 2015
Revised October 2017
Accepted October 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 55–81, http://dx.doi.org/10.1515/JOS-2018-0004

Administrative Data Quality: Investigating Record-Level Address Accuracy in the Northern Ireland Health Register

Brian Foley[1,2], *Ian Shuttleworth*[1], *and David Martin*[3]

Many national statistical institutes (NSIs) are seeking to supplement or replace their traditional population census with a methodology underpinned by administrative sources. Health service register data are key in this regard owing to their high population coverage; it is therefore important to improve understanding of data quality in this administrative source. This study investigated the factors associated with record-level address data mismatch between the Northern Ireland (NI) Health Card Registration System (HCRS) and the 2011 Census, using the NI Longitudinal Study (NILS). Address information in the form of anonymised Unique Property Reference Number (XUPRN) was available for circa 334,000 NILS members with census returns in 2001 and 2011, which provided a benchmark to assess XUPRN accuracy in their linked HCRS record for comparable time points. Multinomial logistic regression revealed a significantly greater likelihood of address mismatch in the HCRS for: males; young adults; individuals with no limiting long-term illness; migrants in the year prior to each census; and residents of communal establishments. Identification of population groups affected by poor quality address information in administrative sources can assist NSIs with the development and implementation of methodological improvements to ensure that official population statistics generated from these sources are fit for purpose.

Key words: Address data quality; census; population statistics; longitudinal data.

1. Introduction

The use of administrative data sources in official statistical systems is well established (Eurostat 2003; Karr 2012; Wallgren and Wallgren 2014; Agafiţei et al. 2015). Many national statistical institutes (NSIs) draw upon these data, routinely collected by government departments, state agencies and other organisations via the operation of a service, transaction or registration to inform statistics on the economy and society. Challenges facing NSIs including financial constraints and the high cost of data collection via sample surveys (United Nations Economic Commission for Europe (UNECE) 2011)

[1] School of Natural and Built Environment, Queens University Belfast, Belfast, BT7 1NN, UK. Emails: brian.foley@nisra.gov.uk and i.shuttleworth@qub.ac.uk
[2] Northern Ireland Statistics and Research Agency, Belfast, BT9 5BF, UK.
[3] Geography and Environment, University of Southampton, Southampton, SO17 1BJ, UK Email: d.j.martin@soton.ac.uk

Acknowledgments: Author B.F. is supported by funding from the Economic and Social Research Council (ESRC) for the Administrative Data Research Centre in Northern Ireland. The help provided by the staff of the Northern Ireland Longitudinal Study (NILS) and the NILS Research Support Unit is acknowledged. The NILS is funded by the Health and Social Care Research and Development Division of the Public Health Agency (HSC R&D Division) and NISRA. The NILS-RSU is funded by the ESRC and the Northern Ireland Government. The authors alone are responsible for the interpretation of the data and any views or opinions presented are solely those of the author and do not necessarily represent those of NISRA/NILS.

are resulting in the increasing integration of administrative sources in statistical production processes.

Among the range of administrative sources in existence, one of the most informative in terms of providing basic demographic and address information is a health service register, referred to hereafter as a health register. These data are based on individual registrations to access primary care services, mainly via general practitioners. Such services are free at the point of use in many countries, for example the United Kingdom (UK), Italy and Denmark (Roland et al. 2012; Lo Scalzo et al. 2009; Pedersen et al. 2012), which is a strong incentive to register, hence the high population coverage generally provided by these data. Accordingly, health registers have utility from a statistical perspective, with many NSIs using these data in their methodology for producing annual subnational population estimates. Furthermore, a number of NSIs that conduct a traditional population census are investigating the potential of administrative sources to supplement or replace this approach; health registers are likely to be key in this regard given their value in a statistical context.

It is therefore important to improve understanding of quality issues in this administrative source given its use in producing population statistics; of specific interest for this study is record-level address data. Among various applications, these data inform the estimation of internal migration flows by some NSIs and they are one of the main matching variables for record linkage between administrative sources integrated in population statistics production processes. Our focus is the health register in Northern Ireland (NI), which is termed the Health Card Registration System (HCRS). Previous studies have investigated the quality of address information in this administrative source (Shuttleworth and Barr 2011; Barr and Shuttleworth 2012; Shuttleworth and Martin 2016) and the England and Wales equivalent, the Patient Register (Smallwood and Lynch 2010), using 2001 Census data as the reference. This study builds upon the existing research by investigating address data quality from both a cross-sectional and longitudinal perspective using data from the 2001 and 2011 censuses of NI. The first aim was to investigate the extent of address mismatch between the HCRS and census in 2011 and to identify some of the associated characteristics using a univariate approach. Secondly, combining results from the 2001- and 2011-based analyses, the aim was to identify the individual-, household-, and area-level factors associated with address mismatch from a longitudinal perspective, employing a multivariate methodology. These aims were met by the use of the NI Longitudinal Study (NILS), one of three UK census-based longitudinal studies (along with the Office for National Statistics (ONS) and Scottish longitudinal studies), but unique in its high sample fraction.

This article is presented in five sections. In Section 2, the use of administrative data to inform the production of population statistics is outlined in the context of the different methodologies employed by NSIs internationally; this is followed by specific reference to health register data and associated quality issues. A detailed description of the NILS is provided in section three, highlighting its value in facilitating a detailed assessment of address data quality in the NI HCRS. We also outline the methods underlying the analysis. The results in Section 4 reveal the factors associated with address mismatch in the HCRS in a cross-sectional and longitudinal context. In Section 5, the findings are interpreted from various perspectives including the wider implications for the statistical application of

health register data before, finally, the conclusions from the study are presented in Section 6.

2. Administrative Data and Population Statistics: National and International Perspectives

A census is fundamental to enumerating a country's population and producing statistics on its main demographic and socioeconomic characteristics from national to small area level. Most countries employ a traditional approach when conducting a census and while the principle of collecting information from all individuals and households remains, the methodology has evolved. This includes the introduction of online data collection in recent census rounds (Moore et al. 2008). The internet is becoming the primary mode of data collection in many of the traditional census-taking countries; the 2016 Canadian Census had an internet collection response rate of 68% (UNECE 2016), the target for the 2021 UK Census is a 65% online response rate (UNECE 2015), while the United States Census Bureau (USCB) are planning for the internet to be the main response mode in their 2020 Census (USCB 2015a). One of the main alternative census methods is a register-based system underpinned by administrative sources, which is well-established in many of the Nordic states (Martin 2006; UNECE 2007; Lange 2014). Owing to the advantages of using administrative data such as reduced respondent burden and the capacity for more frequent statistical outputs, other countries are turning increasingly to these sources. For example, Sweden and Austria made the transition to a completely register-based census in 2011 (Andersson et al. 2013; Kukutai et al. 2015), while Switzerland and Germany introduced integrated census systems combining administrative registers and sample surveys in 2010 and 2011, respectively (UNECE 2012a, 2012b). Although the UK conducts a traditional census, various administrative sources were used to quality assure population estimates from their 2011 event (National Records of Scotland (NRS) 2012; Northern Ireland Statistics and Research Agency (NISRA) 2012; ONS 2012a). Looking ahead, the ONS has committed to greater use of administrative data to enhance the quality of statistics from the 2021 Census of England and Wales (ONS 2014a); this approach has also been adopted by the other UK NSIs in Scotland (NRS 2014) and NI (NISRA 2014). It is clear therefore that administrative data are becoming increasingly relevant internationally in the context of the population census.

While the traditional census provides accurate data for a point in time, there is a requirement for high quality population statistics throughout the intercensal period. Many NSIs in the countries concerned use administrative sources to produce annual population statistics. Internal migration, the movement of individuals within a country, greatly influences the size and composition of populations at subnational level and its estimation is underpinned by administrative data in the UK (ONS 2016), Canada (Statistics Canada (SC) 2015), United States (USCB 2016), and Australia (ABS 2014); health register data are one of the main sources drawn upon in this context. Furthermore, these data are used in the methodology for estimating the subnational distribution of immigrants throughout England and Wales (ONS 2011). Looking forward, the NSIs of New Zealand and England and Wales are assessing the feasibility of replacing their traditional census with a model based largely on linked administrative sources that is less costly and can produce more

frequent population statistics (Bycroft 2015; ONS 2015a). In this regard, the ONS are releasing experimental population estimates at national and subnational level on an annual basis, which are based on linked administrative data sources including the Patient Register (ONS 2015b).

Like many administrative sources, health register data are affected by particular quality issues. Overcoverage is an acknowledged feature for reasons such as emigrants not de-registering prior to moving abroad and multiple area registrations; 'list inflation' of between four per cent and five per cent above the population estimate has been reported for the Patient Register in England and Wales (ONS 2012b) and the HCRS in NI (O'Reilly et al. 2012). Another issue concerns the lag in updating of residential moves, which is especially prevalent among males and in urban and deprived areas (Shuttleworth and Barr 2011; Statistics New Zealand 2013). From a temporal perspective, Barr and Shuttleworth (2012) found that 44% of an internal migrant study cohort in the NILS lagged in reporting a change of address to the HCRS by more than one year, while in a London-based study by Millett et al. (2005), seven per cent of participants took longer than three years to re-register with a general practitioner after a change of address.

There is a need to build upon the existing evidence base on address data quality in administrative sources such as a national health register. To provide context for this study, Figure 1 outlines, conceptually, major contributors to inaccurate address information in administrative data, drawing on the illustrative approach used by Raymer et al. (2015). The delayed or failed reporting of a residential move to the relevant administrative system is considered one of the main statistical challenges associated with the use of administrative data to produce population statistics (ONS 2012c).

The NILS facilitated a novel approach for this study, providing address information for a large sample from consecutive censuses and for regular intervals from the HCRS over a thirteen-year period to conduct a detailed longitudinal analysis. Although NI is the smallest country in the UK, its varied settlement pattern with a large urban centre (Belfast) and rural countryside interspersed with towns is similar to that observed in many other

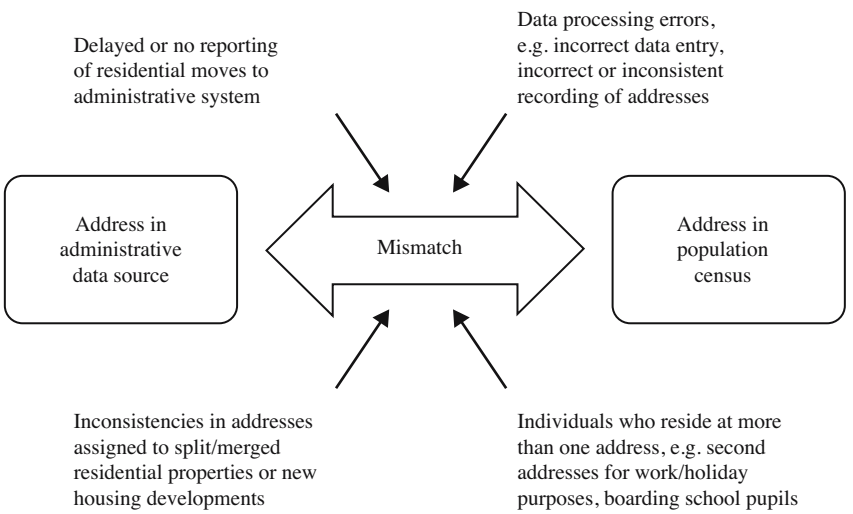


Fig. 1. Overview of contributors to address inaccuracy in an administrative data source.

countries, which broadens the applicability of the findings. Our study may be of interest to organisations and NSIs that use health register data for statistical or research purposes as it provides an insight on record-level inaccuracies in geographical referencing and the associated factors.

3. Data and Methods

3.1. The NILS

The analysis was based on the NILS, which is a representative circa 28% sample of the 1.84 million population of NI. The spine of the NILS is records from the NI HCRS, which is administered by the Health and Social Care Business Services Organisation (BSO) and includes almost 100% of the NI population (O'Reilly et al. 2012). The HCRS data are obtained from the National Health Authority Information Registration System, which is linked to all general practitioner practices in NI. Although each person registered to access primary care services is assigned a unique Health and Care Number, duplicate registrations are identified as one of the factors contributing to overcoverage in the HCRS; however, quality assurance within NISRA sees the removal of duplicate records prior to the use of the data for statistical and research purposes (NISRA 2016a). Once an individual is registered on the HCRS, they can voluntarily update their address information via their general practitioner practice or the BSO website (ONS 2017).

Membership of the NILS is based on having one of 104 predesignated birth dates, with records linked to decennial census returns from 1981 to 2011. There is no complete household structure in the NILS as it is a sample, unlike census data, with some NILS members being a single representative from a household and sometimes multiple members from the same household. The linkage of HCRS and census records in the NILS employs a sequential match-key approach; this methodology enables records for the same individual to be linked where their address differs across both sources by assigning highest matching weights to name, gender and date of birth and using part of the address information in one of the match-keys. Manual verification of linked records and thorough quality checks ensure a high match rate. Further information on the NILS record linkage processing is available in a data matching methodology working paper (Northern Ireland Longitudinal Study 2015). Census data provide a rich source of information on demographic and socioeconomic characteristics of NILS members at individual- and household-level. In addition, residential moves of NILS members from 2001 onwards can be determined from routine HCRS updates provided to the NILS at six-month intervals. Note that the HCRS and census are separate data sources with neither used to update the other; the aforementioned data linkage is for the purpose of the NILS. A more detailed description of the NILS is available in O'Reilly et al. (2012).

Census data in the NILS is based on the enumerated population and therefore excludes wholly imputed individuals and households from the standard coverage adjustment process. Some NILS members do not have census information; this arises where individuals emigrated or died yet remained on the HCRS or where they simply did not provide a census return. This study was based on NILS members with a census record, which facilitated a comparison of address information with that in the HCRS. The lowest

level of geography available for NILS research is Super Output Area (SOA), which was introduced by NISRA for the 2001 Census; there are 890 SOAs in NI, with an average population size of 2,000 (NI Neighbourhood Information Service 2013).

Pointer, maintained by Land and Property Services, is the main address database for NI, in which each property has a Unique Property Reference Number (UPRN). In the HCRS and NI census data, address information includes the UPRN. Record-level address data in the NILS are provided in the form of anonymised UPRN, termed XUPRN. For this study, the availability of XUPRN in the HCRS and census data permitted an indirect comparison of the record-level address recorded in both sources. Owing to issues such as address formats not recognised by Pointer or incomplete address information, it is not possible to assign a UPRN in all cases. Accordingly, in the NILS, XUPRN can be missing in the census or HCRS data, missing in both or present in both; in the case of the latter, XUPRN can be a match or mismatch. This categorisation was used in presenting the cross-sectional results for 2011. The incidence of unassigned XUPRN was particularly high among addresses in Fermanagh, a predominantly rural area in the south-west of NI. This was mainly due to the use of a non-standard addressing system based on geographical units called townlands, which was not entirely incorporated in Pointer.

3.2. Methods

3.2.1. Research Approach

The main element of this study was a longitudinal assessment of address data accuracy in the HCRS in terms of identifying the individual-, household-, and area-level factors associated with XUPRN mismatch. The census in 2001 and 2011 provided benchmark address information for NILS members as this was the most accurate record of where individuals resided at the time of enumeration. In addition, the routine updates of HCRS data gave a detailed address history for NILS members. The reference for one of the twice-yearly updates of HCRS data in the NILS was April, making it possible to obtain XUPRN data for a time point close to the 2001 and 2011 censuses conducted on April 29th and March 27th, respectively. A cross-sectional comparison of record-level XUPRN between the census and HCRS was undertaken for 2001 and 2011. This facilitated classification of NILS members with a census record and valid XUPRN recorded in both sources over time into four categories, namely (i) matching XUPRN between the census and HCRS in 2001 and 2011, (ii) matching XUPRN in 2001 only, (iii) matching XUPRN in 2011 only and (iv) mismatching XUPRN in 2001 and 2011. The primary focus of this study is the longitudinal assessment of address data accuracy, with descriptive results from the 2011 cross-sectional analysis also provided. In addition, the Supplemental data include further regression model output from the longitudinal analysis and 2001 and 2011 cross-sectional analyses. For clarification, *address* and *XUPRN* are used interchangeably, with *mismatch* and *match* referring to NILS members having an address recorded in the HCRS that was different or the same, respectively, to that from which their census questionnaire was returned.

3.2.2. Explanatory Variables

Most of the person- and household-level explanatory variables used in the analysis were obtained from the 2001 and 2011 censuses of NI. Their selection was informed by similar

studies undertaken by Shuttleworth and Barr (2011), Barr and Shuttleworth (2012), and Shuttleworth and Martin (2016), which point to similarities between the hard-to-enumerate population from a census perspective and those with a greater likelihood of having inaccurate address information recorded in administrative systems. Accordingly, age and gender were included alongside marital status, level of education, socioeconomic status, and health status. Country of birth was chosen to investigate if XUPRN mismatch was prevalent among the immigrant population in NI. Migration, based on the 'One year ago, what was your usual address' census question was considered very relevant as it reflected residential moves, while housing tenure and household accommodation type and composition were informative in the context of residential mobility. Transition variables were derived for the longitudinal analysis based on the characteristics of NILS members in both the 2001 and 2011 censuses, for example, single in 2001 and married in 2011. The variables in question were marital status, limiting long-term illness (LLTI), National Statistics Socio-economic Classification (NSSEC), migration, and housing tenure. In addition, a derived variable capturing the frequency of record level address changes reported in the HCRS at six-month intervals between 2001 and 2011 was included. For the variables describing educational qualifications and religion, 2001 Census data were used on account of their antecedent status.

To investigate the geography of XUPRN mismatch, a number of area-level variables were included in the analysis. The 2005 NI Multiple Deprivation Measure (MDM) and component Proximity to Services domain score at SOA level (NISRA 2005a) provided a measure of spatial deprivation and the extent to which people had poor geographical access to key services, respectively, close to the mid-point of the 2001 to 2011 period. Based on the official statistical classification of settlements in NI (NISRA 2005b), the SOA of the census-recorded address was assigned urban or rural status. A variable based on whether the census-recorded address was in Fermanagh or elsewhere in NI was created to assess the effect of the addressing problems associated with the former. Explanatory variables that did not follow an approximate normal distribution were log transformed and expressed as quartiles where necessary.

3.2.3. Analysis

An initial descriptive analysis provided the percentage distributions of selected variables across the various XUPRN status categories for the 2011 cross-sectional analysis initially and, subsequently, the longitudinal assessment of address data quality in the HCRS. To further explore the multivariate relationships in the latter, multinomial logistic regression analysis was conducted using NILS records in the 16 to 74 age group; this restricted age category was relevant in the context of the education and socioeconomic status explanatory variables. For the four-category dependent variable, matching XUPRN in 2001 and 2011 was selected as the reference and the coefficients were expressed as relative risk ratios. Regarding the individual explanatory variables, the reference category was chosen on the basis of having the most records with a status of XUPRN match in both years. Data analysis was undertaken at the secure research environment of the NILS Research Support Unit, primarily using STATA version 14 (StataCorp 2016).

4. Results

4.1. Cross-Sectional Analysis (2011): Descriptive Results

The following descriptive results apply to 485,185 NILS members with a 2011 Census return. Regarding age, the XUPRN match rate between the census and HCRS fluctuated slightly around 85% for the 16 and under and 50 plus age groups (Figure 2). The trough in the XUPRN match rate for the 20 to 40 age group was reflected in the XUPRN mismatch rate, which peaked at 26% for those in their late twenties. The proportions of those with an unassigned XUPRN in the census, HCRS or both remained below ten per cent across all ages; some data were unavailable for the very old on account of aggregation for disclosure control purposes.

There was a reasonably large range in the XUPRN match and mismatch rates across other selected variables (Table 1). Individuals with the following characteristics were associated with a noticeably higher rate of XUPRN mismatch: male; marital status of single; without an LLTI; migrated within NI during the previous twelve months; private renter; living in a flat, apartment or communal establishment; living in a one person (under 65 age group), cohabiting couple or student household. In terms of area effect, the XUPRN match rate of 76% for rural SOAs was around seven per cent lower compared to SOAs classified as urban. Another finding of note was the relatively high proportion in the 'XUPRN not assigned' category of those who moved to NI from outside the country

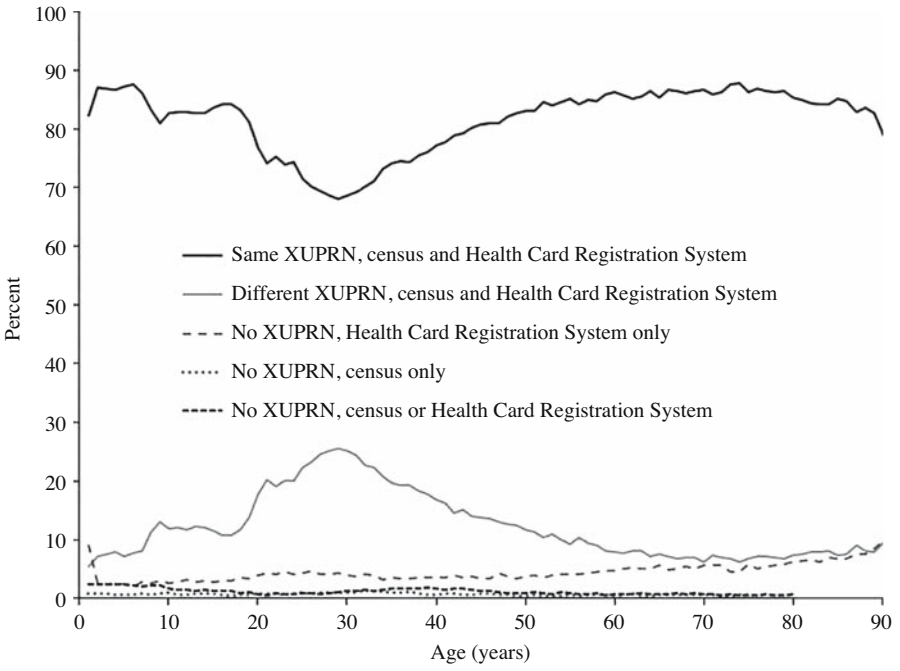


Fig. 2. Percentage distribution of Northern Ireland Longitudinal Study members by age according to XUPRN status category, based on a comparison of XUPRN in the census and Health Card Registration System in 2011; XUPRN is the anonymised Unique Property Reference Number.

Table 1. Distribution of individual-, household- and area-level characteristics based on a comparison of XUPRN in the census and Health Card Registration System (HCRS) in 2011 (figures are percentages based on the count numerator in the final column); XUPRN is the anonymised Unique Property Reference Number.

Variable	No XUPRN: census or HCRS	No XUPRN: census only	No XUPRN: HCRS only	Same XUPRN	Different XUPRN	Total
Sex						
Male	1.22	0.76	4.45	78.46	15.12	234,910
Female	1.21	0.66	3.80	83.01	11.32	250,275
Marital status						
Single	1.28	0.73	3.89	79.07	15.03	235,626
Married	1.34	0.77	4.22	82.68	10.99	186,524
Limiting long-term illness						
Yes	0.79	0.53	4.83	83.69	10.16	102,362
No	1.33	0.76	3.91	80.06	13.95	379,468
Address one year ago						
Same	1.19	0.63	3.93	82.66	11.60	450,766
Moved from within NI	1.59	2.14	5.10	51.58	39.59	25,538
Moved from outside NI	1.45	1.13	15.68	64.88	16.86	3,368
Housing tenure						
Owned	1.38	0.72	3.97	83.14	10.80	353,354
Social rented	0.47	0.45	2.57	82.97	13.54	57,643
Private rented	0.94	0.84	4.78	68.91	24.52	60,968
Accommodation type						
Detached	2.26	1.10	5.64	79.58	11.42	210,150
Semi-detached/terraced	0.28	0.23	1.96	84.71	12.82	245,546
Flat/apartment	1.27	1.72	8.98	61.19	26.83	23,835
Communal establishment	1.84	1.80	21.80	38.39	36.15	5,045

Table 1. Continued.

Variable	No XUPRN: census or HCRS	No XUPRN: census only	No XUPRN: HCRS only	Same XUPRN	Different XUPRN	Total
Household composition						
Couple, dependent children	1.75	0.78	3.78	83.62	10.07	217,390
Lone parent family	0.60	0.41	2.60	82.69	13.69	71,411
One person (over 65 years)	0.49	0.45	5.27	86.36	7.44	21,893
One person (other)	0.66	0.82	4.43	71.75	22.35	32,607
Cohabiting couple	0.80	0.84	3.45	70.48	24.43	27,094
Couple, no dependent children	1.12	0.75	4.25	81.51	12.37	40,649
Students/Communal establishment	1.53	1.79	19.34	34.93	42.41	6,204
Other	0.85	0.66	5.01	80.28	13.20	67,874
Urban/rural classification						
Urban	0.44	0.38	2.55	83.51	13.11	311,508
Rural	2.61	1.31	6.91	75.94	13.24	173,677

within the previous twelve months (15.7%) and residents of communal establishments (21.8%). The minor inconsistencies in the variable totals were due to the 'No code required' category in many of the census-based variables.

4.2. *Longitudinal Analysis (2001 and 2011)*

4.2.1. Descriptive Results

Descriptive results in Table 2 from the 2001- and 2011-based longitudinal analysis were based on 334,670 NILS members with returns from both censuses. The distribution across the four XUPRN status categories was: 246,506 individuals (73.7%) with a matching XUPRN in the census and HCRS in 2001 and 2011; 28,661 (8.6%) with a matching XUPRN in 2001 only; 41,338 (12.4%) with a matching XUPRN in 2011 only; and 18,165 (5.4%) with a mismatching XUPRN in 2001 and 2011. A matching XUPRN in both years was more common among females and those born in the UK or Republic of Ireland (ROI). Consistency over time in the following transition variables was associated with the highest proportions in the category of XUPRN match in both years: being married; having an LLTI; not having migrated in the twelve months prior to the census; and residing in housing that was owned outright. For the category of XUPRN mismatch at both time points, there were noticeably high proportions among individuals who migrated in the twelve months before one or both of the censuses and those whose housing tenure changed from rented in 2001 to owner-occupied in 2011. A large proportion of those who migrated in 2001 only were in the category of XUPRN mismatch in 2001 only, as was the case with reference to 2011. As with Table 1, the minor differences in the variable totals were due to the 'No code required' category.

4.2.2. Multivariate Results

Results from the multinomial logistic regression model in Table 3 detail the multivariate relationships with longitudinal XUPRN status for 243,088 NILS members in the 16 to 74 age group. Statistical significance of variables was widespread on account of the large sample size so most attention is given to the size of effects. Address mismatch in the HCRS was more prevalent among males, who were 2.2 times more likely to have a mismatched XUPRN in 2001 and 2011 relative to matching XUPRN in both years. There was a greater likelihood of address mismatch in both years for those aged 25 to 34 in 2001 compared to the 35 to 44 age group, with those aged 45 to 74 in 2001 less likely to have a mismatched XUPRN between the HCRS and census. While the level of educational qualifications and country of birth were not strongly associated with longitudinal XUPRN status, Catholics had a greater likelihood of address mismatch in one or both years relative to Protestants and other Christians. The variable capturing the frequency of address changes in the HCRS between 2001 and 2011 was influential; compared to non-movers, those who reported at least one address change were less likely to have a mismatched XUPRN in 2011 only or both years relative to individuals with a matching XUPRN over time. Furthermore, the category of XUPRN mismatch in 2001 only had extremely high relative risk ratios, which is discussed further below.

Table 2. Distribution of individual-, household- and area-level characteristics by longitudinal XUPRN status, 2001 to 2011 (figures are percentages based on the count numerator in the final column); XUPRN is the anonymised Unique Property Reference Number.

	XUPRN match, 2001 and 2011	XUPRN match 2001, mismatch 2011	XUPRN mismatch 2001, match 2011	XUPRN mismatch, 2001 and 2011	Total
Sex					
Female	76.51	7.61	11.66	4.22	176,883
Male	70.46	9.63	13.12	6.78	157,787
Country of birth					
UK/Republic of Ireland	73.92	8.45	12.23	5.39	327,988
Outside UK/Republic of Ireland	69.65	8.91	15.81	5.63	3,782
Marital status (2001–2011)					
Single – Single	70.20	11.95	11.93	5.93	133,688
Married – Married	81.04	3.76	10.36	4.85	120,964
Other	73.94	6.36	14.48	5.22	35,866
Single – Married	54.52	16.58	22.37	6.54	24,739
Married – Separated/divorced/widowed	75.36	9.08	10.99	4.56	19,413
Limiting ling-term illness (2001–2011)					
No – No	71.72	9.50	13.02	5.77	232,767
Yes – Yes	81.19	5.34	9.24	4.23	40,682
No – Yes	78.18	6.11	11.32	4.39	40,231
Yes – No	75.41	7.08	12.08	5.44	15,342
Address one year ago (2001–2011)					
Same – Same	78.17	7.23	9.94	4.66	289,754
Different – Same	40.66	4.72	43.28	11.34	22,184
Same – Different	45.53	40.15	7.74	8.58	13,340
Different – Different	26.75	14.43	32.33	26.49	1,525
Housing tenure (2001–2011)					
Owned – Owned	78.08	6.35	10.64	4.93	233,520
Rented – Rented	65.35	13.06	15.17	6.41	78,585
Owned – Rented	65.44	17.42	12.94	4.20	7,978
Rented – Owned	47.49	8.01	34.53	9.97	7,966
Communal establishment – Communal establishment	43.34	39.41	8.45	8.80	2,898

Table 3. Relative risk ratios and associated 95% confidence intervals for individual-, household- and area-level variables from a multinomial logistic regression model of longitudinal XUPRN status (2001 and 2011) for the 16 to 74 age group; XUPRN is the anonymised Unique Property Reference Number.

Variable	XUPRN match 2001, mismatch 2011		XUPRN mismatch 2001, match 2011		XUPRN mismatch 2001 and 2011	
	Relative risk ratio	95% confidence interval	Relative risk ratio	95% confidence interval	Relative risk ratio	95% confidence interval
Constant	0.04*	0.04–0.05	0.02*	0.01–0.02	0.04*	0.04–0.05
Sex (ref: Female)						
Male	1.64*	1.58–1.69	1.55*	1.51–1.59	2.18*	2.10–2.26
Age group, 2001 (ref: 35 to 44)						
16 to 24	2.52*	2.36–2.69	0.70*	0.66–0.74	1.06	0.98–1.14
25 to 34	1.78*	1.69–1.88	1.03	0.99–1.07	1.36*	1.29–1.43
45 to 54	0.67*	0.63–0.71	0.92*	0.88–0.97	0.71*	0.67–0.75
55 to 64	0.60*	0.55–0.64	0.77*	0.72–0.81	0.53*	0.50–0.57
65 to 74	0.62*	0.57–0.67	0.79*	0.73–0.85	0.52*	0.48–0.57
Educational qualifications, 2001 (ref: None)						
Level 1 to 3/apprenticeship	1.14*	1.09–1.19	0.98	0.94–1.02	1.06*	1.01–1.11
Level 4 and above	1.24*	1.17–1.32	1.05*	1.00–1.10	1.06	1.00–1.13
Religion belong to, 2001 (ref: Protestant/other Christian)						
Catholic	1.11*	1.07–1.15	1.27*	1.23–1.31	1.28*	1.23–1.33
Other religions	0.89	0.66–1.22	1.29*	1.02–1.64	1.01	0.71–1.43
No religion/not stated	1.10	0.98–1.24	1.17*	1.06–1.30	1.28*	1.12–1.47
Country of birth, 2001 (ref: UK/Ireland)						
Non–UK/Ireland	1.16*	1.01–1.34	0.94	0.83–1.06	0.97	0.82–1.14
Frequency of address changes in the HCRS, 2001–2011 (ref: None)						
1 to 2	0.71*	0.68–0.74	20.13*	19.34–20.96	0.88*	0.85–0.92
3 and above	0.45*	0.41–0.48	22.45*	21.19–23.79	0.74*	0.68–0.81

Table 3. Continued.

Variable	XUPRN match 2001, mismatch 2011		XUPRN mismatch 2001, match 2011		XUPRN mismatch 2001 and 2011	
	Relative risk ratio	95% confidence interval	Relative risk ratio	95% confidence interval	Relative risk ratio	95% confidence interval
Marital status transition, (ref: Married – Married)						
Single – Single	1.75*	1.66–1.84	1.21*	1.15–1.26	1.24*	1.17–1.31
Single – Married	2.45*	2.31–2.61	0.97	0.93–1.02	1.15*	1.08–1.24
Married – Separated/divorced/widowed	2.23*	2.09–2.37	0.86*	0.81–0.91	1.10*	1.01–1.19
Other	1.82*	1.71–1.93	1.19*	1.14–1.25	1.39*	1.31–1.47
Limiting long-term illness (LLTI) transition, (ref: No LLTI – No LLTI)						
LLTI – LLTI	0.84*	0.79–0.89	0.79*	0.75–0.83	0.77*	0.72–0.82
LLTI – No LLTI	0.84*	0.77–0.91	0.89*	0.83–0.95	0.88*	0.81–0.96
No LLTI – LLTI	0.82*	0.77–0.86	0.96	0.92–1.01	0.84*	0.79–0.89
Other	0.56	0.26–1.23	0.45	0.19–1.05	0.48	0.19–1.24
Socio-economic classification transition, (ref: Professional – Professional)						
Routine — Routine	0.85*	0.81–0.90	0.89*	0.85–0.93	0.80*	0.76–0.85
LT unemployed/never worked – LT unemployed/never worked	0.87*	0.77–0.98	0.97	0.88–1.08	0.89	0.79–1.01
Student – In employment	1.12*	1.05–1.19	0.71*	0.66–0.76	0.85*	0.78–0.94
Other	0.96	0.92–1.00	0.95*	0.92–0.99	0.86*	0.82–0.91
Address one year ago transition, (ref: Same address – Same address)						
Same address – Mover	4.19*	3.95–4.45	0.63*	0.58–0.69	2.53*	2.31–2.77
Mover – Same address	0.96	0.88–1.04	4.96*	4.75–5.18	4.22*	3.98–4.47
Mover – Mover	1.59*	1.27–2.00	3.98*	3.37–4.70	13.58*	11.38–16.22
Other	2.33*	1.85–2.94	2.48*	2.15–2.85	2.47*	2.05–2.99

Table 3. Continued.

Variable	XUPRN match 2001, mismatch 2011		XUPRN mismatch 2001, match 2011		XUPRN mismatch 2001 and 2011	
	Relative risk ratio	95% confidence interval	Relative risk ratio	95% confidence interval	Relative risk ratio	95% confidence interval
Housing tenure transition, (ref: Owned – Owned)						
Owned – Rented	3.24*	3.08–3.41	0.62*	0.59–0.66	1.42*	1.32–1.53
Rented – Owned	1.12*	1.05–1.19	1.32*	1.26–1.39	1.42*	1.33–1.52
Rented – Rented	1.31*	1.24–1.39	1.03	0.98–1.08	1.50*	1.41–1.60
Communal establishment – Communal establishment	9.23*	5.92–14.40	8.31*	5.59–12.37	25.14*	17.57–35.97
Other	3.41*	3.13–3.72	0.93	0.85–1.02	1.75*	1.56–1.96
Urban/rural classification of SOA, 2001 (ref: Urban)						
Rural	1.04	0.97–1.11	1.23*	1.17–1.30	1.29*	1.20–1.39
2005 Multiple Deprivation Measure score by SOA, log transformed	0.95	0.90–1.01	1.15*	1.10–1.21	1.05	0.99–1.13
2001-based SOA population density, log transformed	0.98	0.93–1.02	0.96	0.93–1.00	0.94*	0.89–0.98
Proximity to Services score by SOA, log transformed (ref: Quartile 1)						
Quartile 2	1.00	0.96–1.05	0.87*	0.83–0.90	0.84*	0.79–0.89
Quartile 3	0.99	0.93–1.05	0.95*	0.90–1.00	0.84*	0.78–0.90
Quartile 4	1.00	0.91–1.10	1.19*	1.10–1.30	1.04	0.94–1.16
Address location transition, (ref: Elsewhere in NI – Elsewhere in NI)						
Fermanagh – Fermanagh	1.23*	1.07–1.42	1.36*	1.20–1.55	1.53*	1.33–1.76
Fermanagh – Elsewhere in NI	1.21	0.77–1.91	0.99	0.69–1.42	2.18*	1.40–3.40
Elsewhere in NI – Fermanagh	3.69*	2.71–5.03	0.66*	0.47–0.92	1.51	0.97–2.35

*Statistically significant.

Clear patterns were evident for the transition variables. Those with an unchanged marital status of single compared to individuals married at the two time points had a greater propensity for XUPRN mismatch in one or both years relative to matching XUPRN at both time points. Individuals with an LLTI in 2001 and/or 2011 and NILS members in routine employment in 2001 and 2011 relative to professionals were less likely to exhibit XUPRN mismatch in one or both years. There was a strong positive association between migration and address mismatch, with those who made an address change prior to each census at 14 times the risk of XUPRN mismatch in both 2001 and 2011. Housing tenure was influential, with renters and especially communal establishment residents at both time points having a greater likelihood of XUPRN mismatch in both years. Compared to the individual- and household-level factors, there was less variation in the relative risk ratios for the area-level variables, with many close to the threshold value of one. This indicates that from a longitudinal perspective, area characteristics were less influential in terms of the determinants of address mismatch in the HCRS. There were some noteworthy findings however, with rural relative to urban areas and having a census-recorded address in Fermanagh in both years as opposed to elsewhere in NI generally associated with a greater likelihood of XUPRN mismatch in 2001 and/or 2011.

The interaction of migration transition with age was investigated (see Supplemental data, Table 1 found online at: http://dx.doi.org/10.1515/JOS-2018-0004). Compared to older age groups, 16–24 and 25–34 year olds in 2001 who migrated in the year prior to each census had a greater likelihood of consistent address mismatch relative to non-movers. A further interaction of migration and housing tenure transitions indicated that NILS members in rented accommodation who made pre-census residential moves were at greater risk of XUPRN mismatch compared to those in owner-occupied housing.

The findings from the 2001 and 2011 cross-sectional logistic regression models of XUPRN mismatch (see Supplemental data, Table 2 found online at: http://dx.doi.org/10.1515/JOS-2018-0004) were similar to those from the longitudinal model of address mismatch (Table 3). The former models revealed common characteristics associated with address mismatch at each time point, namely males, young adults (aged 25–35) relative to older age groups, single people compared to those who were married, pre-census migrants, renters and communal establishment residents as opposed to those in owner-occupied housing, and in terms of geography, living in Belfast or Fermanagh relative to elsewhere in NI.

4.2.3. Residential Movement

The longitudinal analysis revealed 18,165 NILS members with an address mismatch in the HCRS in 2001 and 2011. Of particular interest was whether this group made more residential moves compared to those with a matching XUPRN in both census years. The updates of address data from the HCRS every six months from 2001 to 2014 facilitated analysis of the proportion in each of the four XUPRN status categories with a reported address change at each time point. Figure 3 indicates little difference in the proportion of individuals reporting an address change between 2001 to 2011 for the categories of matching and mismatching XUPRN in both years, which suggests similar levels of residential movement over the decade. The main explanation seems to be that individuals who mismatched in 2001 and 2011 had happened to change address in the year before the

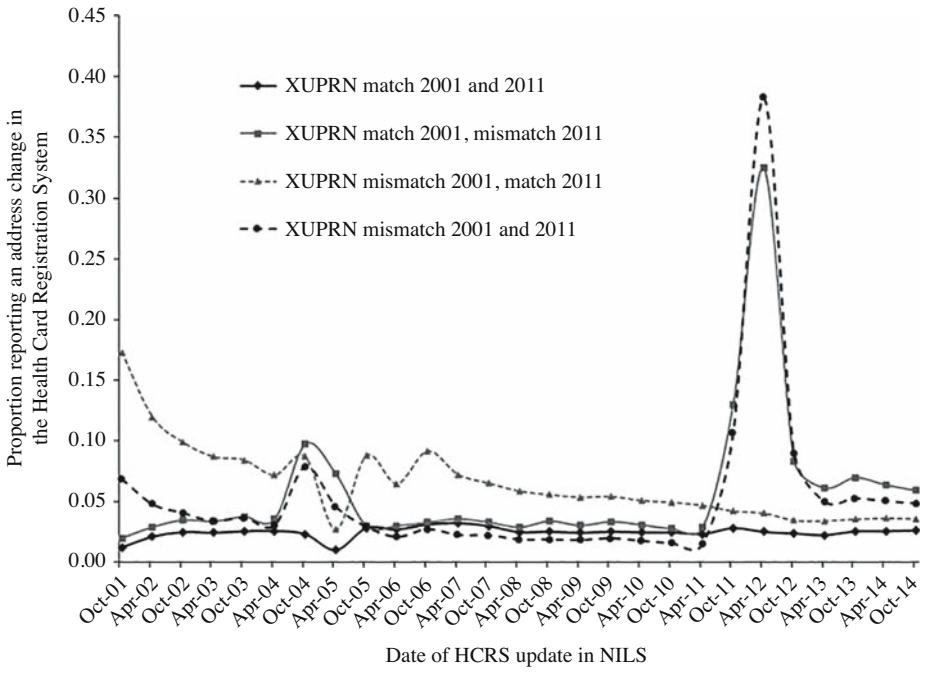


Fig. 3. Proportion of Northern Ireland Longitudinal Study (NILS) members reporting an address change by six-monthly extract of the Health Card Registration System (HCRS) from 2001 to 2014, according to longitudinal XUPRN status; XUPRN is the anonymised Unique Property Reference Number.

respective census but not that they were more spatially mobile in general. The sharp peak for those with a mismatched XUPRN in 2011 can be attributed in part to NILS members who made a residential move prior to the 2011 Census, which was not reported to the HCRS until after March 2011. However, this peak is somewhat inflated by a data cleaning exercise undertaken by BSO during 2011 and 2012, which resulted in some non-genuine address changes being recorded in the HCRS data provided to the NILS. Although the time series begins in October 2001, the remnants of a similar peak associated with lagged reporting of an address change to the HCRS is evident for those with a mismatched address in 2001.

5. Discussion

This study provides a useful insight on an element of data quality in a key administrative source and highlights the value of the NILS as a research resource. Of the individual-, household-, and area-level factors associated with address mismatch in the HCRS, many are intuitive and others less so. In general, the frequency of residential moves is highest among young adults for reasons such as leaving home, pursuing higher education and entering the labour market. In addition, this population group has a low level of engagement with primary care services, thus increasing the likelihood of address changes not being communicated to the associated administrative system. It is very likely that these combined factors contributed to the high rates of address mismatch in the HCRS for the 20 to 34 years age group observed in the cross-sectional and longitudinal analyses. Those

with an LLTI can be expected to engage with their general practitioner to a greater extent compared to individuals in good health, thereby increasing the likelihood of maintaining an accurate address in the HCRS (Shuttleworth and Barr 2011; Barr and Shuttleworth 2012). The greater prevalence of XUPRN mismatch among Catholics may be explained in part by their geographical distribution across NI; rural relative to urban SOAs had a lower likelihood of address match between the HCRS and census and both the former and SOAs with a majority Catholic resident population are prevalent in the west and south-west of the country. The often unforeseen movement of individuals into care-homes, hospitals or prisons along with the transience of many communal establishment populations is a probable explanation for the poor quality address information in the HCRS displayed by individuals residing in this accommodation type. Regarding area effects, the address referencing problem in Fermanagh appeared to exacerbate XUPRN mismatch, which was more prevalent in this part of NI. Although records with an address in this area amounted to just 1.5% and 1.9% of all NILS records with an assigned XUPRN in 2001 and 2011, respectively, the Fermanagh effect may represent a wider issue with address formats and referencing in rural parts of NI that warrants further investigation.

The 2011 cross-sectional analysis provided evidence of higher incidence of address mismatch in the HCRS for those in households that deviated from the traditional nuclear family structure. This is important as increasingly diverse households have become a feature of modern society. For example, in the UK in 2015, cohabiting couples were the fastest growing family type, accounting for 17% of all families, while 29% of households consisted of just one person (ONS 2015c). These household types are also becoming more common in Canada (SC 2012), the United States (USCB 2013), and several European countries (Sanchez Gassen and Perelli-Harris 2015). In addition, multi-family and multi-generational households are becoming more numerous (ONS 2015c; Fry and Passel 2014). For cohabiting couples without dependent children and single-person households in particular, there is likely to be greater scope for residential movement and, consequently, their address information to be out-dated in administrative systems. In the context of NSIs using administrative sources to inform population statistics, the frequency of engagement by particular population groups with the administrative system(s) in question is a key consideration regarding address information accuracy. Indeed, the increasing complexity of household structure presents a challenge to NSIs when conducting a census irrespective of whether they continue with a traditional approach or adopt a method more reliant on administrative sources. Households are a fundamental observation unit for population statistics; therefore, the potential of administrative sources to provide an insight on household structure should be explored so that NSIs can define households accurately and appropriately.

Student households were shown to be problematic in terms of their residents maintaining accurate address information in the HCRS; a contributory factor is likely to be the highly mobile nature of student populations (Duke-Williams 2009; Finney 2011). Areas around higher education institutions are subject to considerable population turnover, including inflows of new and continuing students in line with the academic year and outflows of graduates pursuing employment or further education opportunities elsewhere. This group present challenges for the production of high quality population statistics. For example, there were specific questions included in the 2011 UK Census to determine the appropriate address at which to classify a student as usually resident. Furthermore, the

methodology underlying the estimation of annual internal migration flows in UK constituent countries now incorporates higher education data to provide more accurate student address information (NISRA 2016b; ONS 2016). Regarding health register data, students are infrequent users of primary care services, which compounds the problem of address inaccuracy. In using secondary data such as health registers to produce population statistics, it is prudent for NSIs to incorporate supplementary sources that provide better quality address data on highly mobile groups such as students. Furthermore, these findings highlight the importance of addressing key areas of weakness in administrative sources when used by NSIs to augment or replace a traditional census; a robust statistical system drawing upon administrative sources should seek to improve the quality of the source data.

During the 2000s, international migration to NI experienced an unprecedented surge, with around 122,000 immigrants estimated to have arrived in the country over the decade. European and North-American studies have shown that the propensity to migrate internally among immigrants is greatest for recent arrivals to the host country (Reher and Silvestre 2009; King and Newbold 2011), which is relevant in the context of maintaining accurate address information in administrative systems that immigrants interact with. In general, this study did not find a greater prevalence of address mismatch in the HCRS for NILS members born outside the UK or ROI compared to the native population. A possible factor is that the vast majority of immigrants were from Poland and Lithuania (Krausova and Vargas-Silva 2014); there is evidence of the former developing strong ties with their place of residence after being most internally mobile in the initial stages of their immigration to the UK (Trevena et al. 2013). As immigration to NI increased steadily from 2000 before peaking in 2007, it is likely that a large proportion of this population group had settled in residential terms by 2011, thus providing greater scope for having accurate address information recorded in the HCRS.

Many of the determinants of address mismatch in the HCRS indicated by this study exhibit a strong relationship with residential mobility and, consequently, internal migration. This is unsurprising since all of these phenomena are intertwined. Males, young adults, individuals who are single, those in good health and professionals have been shown to be more residentially mobile and migrate internally to a greater extent (Owen and Green 1992; Bailey and Livingston 2007; Champion et al. 1998; Finney and Simpson 2008); these characteristics also exhibited a positive relationship with XUPRN mismatch in the HCRS. A number of the aforementioned factors are also relevant in the context of census under-enumeration. The issue of nonresponse when conducting a population census is most pronounced among the likes of young adult males and residents of single occupant and student households (Rahman and Goldring 2006; Martin 2010); likewise, these had a higher likelihood of exhibiting XUPRN mismatch. For NSIs planning to supplement their traditional census process with administrative data, it will be important to consider these individual and household characteristics not just in terms of the hard-to-enumerate population but also where address information in the administrative sources will be drawn upon for specific purposes such as informing an address register or data linkage.

Having migrated in the year prior to each census was strongly associated with address mismatch in the HCRS, based on the main effect and interactions with age and housing tenure. As migration is a continual demographic process, a cohort with questionable locational information persists, which needs to be accounted for when using address

information in health register data in the population statistics production process. Another interesting finding from the longitudinal analysis was the similar levels of residential mobility between individuals with a mismatched XUPRN in 2001 and 2011 and those with a matching XUPRN in both years. A reasonable initial hypothesis would have been more residential movement by the former group. Along with the post-census peak in address changes recorded in the HCRS for individuals with a mismatched XUPRN at census time, it reaffirms the existence of a cohort who persistently lagged in reporting or failed to report an address change to the health register, as observed by Barr and Shuttleworth (2012) and Shuttleworth and Martin (2016). For mobile population groups such as young adult males who are more prone to address inaccuracy in administrative systems, the longitudinal structure of the NILS would facilitate investigation of the extent to which address information at a point in time could be used to infer accurate address data at a previous juncture.

The comparable findings from the longitudinal and cross-sectional models suggests that it was the status at the time of census enumeration and not necessarily the 2001 to 2011 transitions that contributed to address inaccuracy in the HCRS. Therefore, the latter models were sufficient to gain an understanding of the factors associated with this aspect of data accuracy in the administrative source. Nonetheless, the longitudinal model was a novel extension of the approach adopted by Shuttleworth and Martin (2016) in their 2001-based study and was appropriate for the inclusion of the variable based on the frequency of address changing in the HCRS over the 2001–2011 period, which proved to be an influential factor. In this regard, the extremely high relative risk ratios for the category of XUPRN mismatch in 2001 only warranted further investigation. From the raw data, 74% and 15% of this cohort reported 1–2 and 3 or more address changes, respectively, to the HCRS over the decade; the corresponding distribution for the XUPRN mismatch in 2011 only group was 35% and 7%, and 29% and 7% for those mismatching in both years. This suggests a prevalence of address changing prior to March 2001 within the XUPRN mismatch in 2001 only group that was not reported to the HCRS until after the 2001 Census.

The extent of address mismatch in the HCRS has implications for the statistical application of this administrative source. National health registers elsewhere are likely to experience similar data quality issues, so while the potential impacts are described primarily in a UK context, they are relevant to other countries. The quality of internal migration estimates, based largely on address changes in health register data, is affected by address inaccuracy. Internal moves are missed or lagged where individuals fail to report a change of address or a period of time elapses before doing so. As this component greatly influences population size and composition at local level, inaccurate address information in the health register underlying the estimates affects the quality of annual subnational population estimates; this is important given the widespread use of these data, which includes informing the allocation of funding from central government to local authorities, councils and health bodies. For the next England and Wales census in 2021, there are a number of broad uses of administrative data proposed. These include using activity information to enhance the estimates of the size and location of the population (ONS 2015d); health registers are likely be a key source in this regard based on the successful use of HCRS data for the 2011 Census under-enumeration project in NI (NISRA 2015). Furthermore, in countries where the traditional census is discontinued, administrative

sources such as the health register are likely to have a key role in the production of population statistics. Record linkage between data sources will be a fundamental aspect of this alternative approach, which typically uses address information along with name and date of birth as linking variables. Address inaccuracies in health register data could therefore adversely affect the matching quality achieved in record linkage processing.

An important aspect underlying this analysis is the increasing difficulty in referencing individuals to a single address. Features of modern society such as second addresses for holiday, work or study purposes and fragmented family structures causing shared custody of children result in individuals residing at more than one address. This was acknowledged in the 2011 Census of England and Wales, with ONS releasing statistics based on alternative population bases, namely the out-of-term student population and usually resident dependent children with a parental second address (ONS 2014b, 2014c). Therefore, genuine address inconsistencies between census and administrative data sources can exist. The extent to which this contributed to the XUPRN mismatch between census and HCRS data observed in this study is unknown. However, it is one of many challenges for NSIs that move away from a traditional census in favour of a methodology based on linked administrative sources.

6. Conclusion

This study took a novel longitudinal approach, facilitated by the NILS, to investigate the extent of address mismatch in a key administrative source and the associated factors. Our findings provide evidence of the commonalities with internal migration, residential mobility and hard-to-enumerate groups from a census perspective. Currently, health registers are used by numerous NSIs to inform the ongoing production of population statistics. In addition, given the apparent discontinuation of the traditional census in many countries, these and other administrative sources are likely to be fundamental to the generation of population and census statistics in the future. While national health registers are a valuable statistical resource owing to their high population coverage, this and other studies have shown the quality of their record-level address information to be impaired for certain population groups. It is important to further improve understanding of quality issues in this and other administrative sources that are likely to underpin official statistical systems in the future. Many of the limitations of administrative data such as undercoverage and lags in updating are acknowledged in the context of their application for statistical purposes. Indeed, some NSIs incorporate additional administrative sources to address specific shortcomings. However, it is prudent to build upon existing research on administrative data quality; effective methodological improvements can then be developed and implemented to ensure that official statistics generated from these sources are fit for purpose and sufficiently accurate.

7. References

Agafiţei, M., F. Gras, W. Kloek, F. Reis, and S. Vâju. 2015. "Measuring Output Quality for Multisource Statistics in Official Statistics: Some Directions." *Statistical Journal of the IAOS* 31: 203–211. Doi: http://dx.doi.org/10.3233/sji-150902.

Andersson, C., A. Holmberg, I. Jansson, K. Lindgren, and P. Werner. 2013. "Methodological Experiences from a Register-Based Census." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 6, 2009. 3289–3296. Alexandria, US. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2013/Files/309549_82867.pdf (accessed August 2017).

Australian Bureau of Statistics (ABS). 2014. *Information Paper: Review of Interstate Migration Method*. Available at: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/3412.0.55.003Main+Features1Mar%202014?OpenDocument (accessed August 2017).

Bailey, N. and M. Livingston. 2007. *Population Turnover and Area Deprivation: Final Report.* Joseph Rowntree Foundation. Available at: https://www.jrf.org.uk/sites/default/files/jrf/migrated/files/2004-population-census-deprivation.pdf (accessed August 2017).

Barr, P. and I. Shuttleworth. 2012. "Reporting Address Changes by Migrants: The Accuracy and Timeliness of Reports via Health Card Registers." *Health & Place* 18: 595–604. Doi: http://dx.doi.org/10.1016/j.healthplace.2012.01.005.

Bycroft, C. 2015. "Census Transformation in New Zealand: Using Administrative Data without a Population Register." *Statistical Journal of the IAOS* 3(13): 401–411. Doi: http://dx.doi.org/10.3233/SJI-150916.

Champion, T., S. Fotheringham, P. Rees, P. Boyle, and J. Stillwell. 1998. *The Determinants of Migration Flows in England: A Review of Existing Data and Evidence*. Report for the Department of the Environment, Transport and the Regions, UK. Available at: http://www.geog.leeds.ac.uk/publications/DeterminantsOfMigration/report.pdf (accessed August 2017).

Duke-Williams, O. 2009. "The Geographies of Student Migration in the UK." *Environment and Planning A* 41(8): 1826–1848. Doi: http://dx.doi.org/10.1068/a4198.

Eurostat. 2003. *Quality Assessment of Administrative Data for Statistical Purposes*. Eurostat, Luxembourg. Available at: https://unstats.un.org/unsd/EconStatKB/KnowledgebaseArticle10264.aspx (accessed August 2017).

Finney, N. and L. Simpson. 2008. "Internal Migration and Ethnic Groups: Evidence for Britain from the 2001 Census." *Population, Space and Place* 14: 63–83. Doi: http://dx.doi.org/10.1002/psp.481.

Finney, N. 2011. "Understanding Ethnic Differences in the Migration of Young Adults within Britain from a Lifecourse Perspective." *Transactions of the Institute of British Geographers* 36(3): 455–470. Doi: http://dx.doi.org/10.1111/j.1475-5661.2011.00426.x.

Fry, R. and J.S. Passel. 2014. *In Post-Recession Era, Young Adults Drive Continuing Rise in Multi-generational Living*. Pew Research Center's Social and Demographic Trends project Washington DC, US. Available at: http://www.pewsocialtrends.org/files/2014/07/ST-2014-07-17-multigen-households-report.pdf (accessed August 2017).

Karr, A.F. 2012. "Discussion on Statistical Use of Administrative Data: Old and New Challenges." *Statistica Neerlandica* 66(1): 80–84. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00510.x.

King, K.M. and K.B. Newbold. 2011. "Internal Migration of Canadian Immigrants, 1993–2004: Evidence from the Survey of Labour and Income Dynamics." *Canadian Studies in Population* 38: 1–18.

Krausova, A. and C. Vargas-Silva. 2014. *Northern Ireland: Census Profile*. Migration Observatory briefing, COMPAS, University of Oxford, UK. Available at: http://www.migrationobservatory.ox.ac.uk/resources/briefings/northern-ireland-census-profile/ (accessed August 2017).

Kukutai, T., V. Thompson, and R. McMillan. 2015. "Whither the Census? Continuity and Change in Census Methodologies Worldwide, 1985–2014." *Journal of Population Research* 32(1): 3–22. Doi: http://dx.doi.org/10.1007/s12546-014-9139-z.

Lange, A. 2014. "The Population and Housing Census in a Register Based Statistical System." *Statistical Journal of the IAOS* 30(1): 41–45. Doi: http://dx.doi.org/10.3233/SJI-140798.

Lo Scalzo, A., A. Donatini, L. Orzella, A. Cicchetti, S. Profili, and A. Maresso. 2009. *Italy: Health system review*. World Health Organisation, Copenhagen, Denmark. Available at: http://www.euro.who.int/__data/assets/pdf_file/0006/87225/E93666.pdf (accessed August 2017).

Martin, D. 2006. "Last of the Censuses? The Future of Small Area Population Data." *Transactions of the Institute of British Geographers, New Series* 31: 6–18. Doi: http://dx.doi.org/10.1111/j.1475-5661.2006.00189.x.

Martin, D. 2010. "Understanding the Social Geography of Census Undercount." *Environment and Planning A* 42(11): 2753–2770. Doi: http://dx.doi.org/10.1068/a43123.

Millett, C., C. Zelenyanszki, K. Binysh, J. Lancaster, and A. Majeed. 2005. "Population Mobility: Characteristics of People Registering with Qeneral Practices." *Public Health* 119(7): 632–638. Doi: http://dx.doi.org/10.1016/j.puhe.2004.09.004.

Moore, T., L. Bailie, and G. Gilmour. 2008. *Building a Business Case for Census Internet Data Collection*. In Proceedings of Statistics Canada Symposium. Statistics Canada, Ontario, Canada. Available at: http://www.statcan.gc.ca/pub/11-522-x/2008000/article/10978-eng.pdf (accessed August 2017).

National Records of Scotland (NRS). 2012. *Overview of Administrative Comparator Data Used in 2011 Census Quality Assurance*. Edinburgh, UK: NRS. Available at: http://www.scotlandscensus.gov.uk/documents/methodology/census-overview-comparative-sources-uksa.pdf (accessed August 2017).

NRS. 2014. *Background information on the Beyond 2011 Programme*. Edinburgh, UK: NRS. Available at: http://www.nrscotland.gov.uk/files//census/2021-census/Background/2021-census-background.pdf (accessed August 2017).

Northern Ireland Longitudinal Study (NILS). 2015. *NILS Data Matching Methodology, NILS Working Paper 3.1*. Belfast, UK: NISRA. Available at: http://www.qub.ac.uk/research-centres/NILSResearchSupportUnit/FileStore/Filetoupload,425661,en.DOCX (accessed August 2017).

Northern Ireland Neighbourhood Information Service. 2013. *NISRA Geography Fact Sheet*. Belfast, UK: NISRA. Available at: http://www.ninis2.nisra.gov.uk/public/documents/NISRA%20Geography%20Fact%20Sheet.pdf (accessed August 2017).

Northern Ireland Statistics and Research Agency (NISRA). 2005a. *Northern Ireland Multiple Deprivation Measure 2005*. Belfast, UK: NISRA. Available at: https://www.nisra.gov.uk/sites/nisra.gov.uk/files/publications/NIMDM2005FullReport.pdf (accessed August 2017).

NISRA. 2005b. *Report of the Inter-Departmental Urban-Rural Definition Group: Statistical Classification and Delineation of Settlements*. Belfast, UK: NISRA. Available at: http://www.ninis2.nisra.gov.uk/public/documents/ur_report.pdf (accessed August 2017).

NISRA. 2012. *Quality Assurance of the 2011 Census in Northern Ireland*. Belfast, UK: NISRA. Available at: https://www.nisra.gov.uk/sites/nisra.gov.uk/files/publications/2011-census-quality-assurance-strategy.pdf (accessed August 2017).

NISRA. 2014. *The Future Provision of Census of Population Information for Northern Ireland*. Belfast, UK: NISRA. Available at: https://www.nisra.gov.uk/sites/nisra.gov.uk/files/publications/the-future-provision-of-census-of-population-information-for-northern-ireland.pdf (accessed August 2017).

NISRA. 2015. *Northern Ireland Census 2011 General Report*. Belfast, UK: NISRA. Available at: https://www.nisra.gov.uk/sites/nisra.gov.uk/files/publications/2011-census-general-report.pdf (accessed August 2017).

NISRA. 2016a. *Population Estimates and Projections Data Quality Document*. Belfast, UK: NISRA. Available at: https://www.nisra.gov.uk/sites/nisra.gov.uk/files/publications/Population-DataQuality_1.pdf (accessed August 2017).

NISRA. 2016b. *Methodology Paper – Mid-Year Population Estimates for Northern Ireland*. Belfast, UK: NISRA. Available at: https://www.nisra.gov.uk/sites/nisra.gov.uk/files/publications/Methodology-2015.PDF (accessed August 2017).

Office for National Statistics (ONS). 2011. *Improved Immigration Estimates to Local Authorities in England and Wales: Overview of Methodology*. Hampshire, UK: ONS. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/imps/improvements-to-local-authority-immigration-estimates/overview-of-improved-methodology.doc (accessed August 2017).

ONS. 2012a. *Overview of Administrative Comparator Data Used in 2011 Census Quality Assurance*. Hampshire, UK: ONS. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/data-quality-assurance/overview-of-administrative-comparator-data-used-in-2011-census-quality-assurance.pdf (accessed August 2017).

ONS. 2012b. *Beyond 2011: Administrative Data Sources Report: NHS Patient Register*. Hampshire, UK: ONS. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/sources-reports/beyond-2011–administrative-data-sources-report–nhs-patient-register–s1-.pdf (accessed December 2015).

ONS. 2012c. *Beyond 2011: Exploring the Challenges of Using Administrative Data*. Hampshire, UK: ONS. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011–exploring-the-challenges-of-using-administrative-data.pdf (accessed August 2017).

ONS. 2014a. *The census and future provision of population statistics in England and Wales: Recommendation from the National Statistician and Chief Executive of the UK Statistics Authority*. Hampshire, UK: ONS. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/beyond-2011-report-on-autumn-2013-consultation–and-recommendations/national-statisticians-recommendation.pdf (accessed August 2017).

ONS. 2014b. *Out of Term Population of England and Wales: An Alternative 2011 Census Population Base*. Hampshire, UK: ONS. Available at: http://www.ons.gov.uk/ons/dcp171776_377904.pdf (accessed August 2017).

ONS. 2014c. *Dependent Children Usually Resident in England and Wales with a Parental Second Address, 2011*. Hampshire, UK: ONS. Available at: http://www.ons.gov.uk/ons/dcp171776_372236.pdf (accessed August 2017).

ONS. 2015a. *Beyond 2011 research strategy and plan – 2015 to 2017*. Hampshire, UK: ONS. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/research-strategy-and-plan—2015-2017.pdf (accessed August 2017).

ONS. 2015b. *2021 Administrative Data Research Report 2015: Methodology and Analysis of Estimates Produced from a Statistical Population Dataset*. Hampshire, UK: ONS. Available at: http://www.ons.gov.uk/ons/guide-method/census/2021-census/progress-and-development/research-projects/beyond-2011-research-and-design/research-outputs/administrative-data-research-outputs–2015.pdf (accessed August 2017).

ONS. 2015c. *Families and Households: 2015*. Hampshire, UK: ONS. Available at: http://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2015-11-05/pdf (accessed March 2015).

ONS. 2015d. *2021 Census Design Document*. Hampshire, UK: ONS. Available at: http://www.ons.gov.uk/ons/guide-method/census/2021-census/progress-and-development/research-projects/2021-census-design-document.pdf (accessed December 2015).

ONS. 2016. *Internal migration by local authorities in England and Wales*. Hampshire, UK: ONS. Available at: https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/populationandmigration/migrationwithintheuk/qmis/internalmigrationestimatesqmi/qmiimelajune16finalforpub.pdf (accessed August 2017).

ONS. 2017. *Northern Ireland internal migration and Northern Ireland Medical Card information: quality assurance of administrative data used in population statistics, Feb 2017: ONS*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/northernirelandinternalmigrationandnorthernirelandmedicalcardinformationqualityassuranceofadministrativedatausedinpopulationstatisticsfeb2017/pdf (accessed August 2017).

O'Reilly, D., M. Rosato, G. Catney, F. Johnston, and M. Brolly. 2012. "Cohort Description: The Northern Ireland Longitudinal Study (NILS)." *International Journal of Epidemiology* 41(3): 634–641. Doi: http://dx.doi.org/10.1093/ije/dyq271.

Owen, D.W. and A.E. Green. 1992. "Migration Patterns and Trends." In *Migration Processes and Patterns Volume 1: Research Progress and Prospects*, edited by A.G. Champion and A. Fielding, 17–38. London: Belhaven.

Pedersen, K.M., J.S. Andersen, and J. Søndergaard. 2012. "General Practice and Primary Health Care in Denmark." *The Journal of the American Board of Family Medicine* 25(Suppl 1): S34–S38. Doi: http://dx.doi.org/10.3122/jabfm.2012.02.110216.

Rahman, N. and S. Goldring. 2006. "Factors Associated with Household Non-Response in the 2001 Census." Newport, UK: ONS. *Survey Methodology Bulletin* 59: 11–24.

Raymer, J., P. Rees, and A. Blake. 2015. "Frameworks for Guiding the Development and Improvement of Population Statistics in the United Kingdom." *Journal of Official Statistics* 31(4): 699–722. Doi: https://doi.org/10.1515/jos-2015-0041.

Reher, D.S. and J. Silvestre. 2009. "Internal Migration Patterns of Foreign-Born Immigrants in a Country of Recent Mass Immigration: Evidence from New Micro Data for Spain." *International Migration Review* 43: 815–849. Doi: http://dx.doi.org/10.1111/j.1747-7379.2009.00785.x.

Roland, M., B. Guthrie, and D.C. Thomé. 2012. "Primary Medical Care in the United Kingdom." *The Journal of the American Board of Family Medicine* 25(Suppl 1): S6–S11. Doi: http://dx.doi.org/10.3122/jabfm.2012.02.110200.

Sanchez Gassen, N. and B. Perelli-Harris. 2015. "The Increase in Cohabitation and the role of Union Status in Family Policies: A Comparison of 12 European Countries." *Journal of European Social Policy* 25(4): 431–449. Doi: http://dx.doi.org/10.1177/0958928715594561.

Shuttleworth, I. and P. Barr. 2011. "Who Reports Address Changes Through the Health Service System? The Characteristics of Laggers and Non-Reporters Using the Northern Ireland Longitudinal Study." Newport, UK: ONS. *Population Trends* 144: 48–54.

Shuttleworth, I. and D. Martin. 2016. "People and Places: Understanding Geographical Accuracy in Administrative Data from the Census and Health Service Systems." *Environment and Planning A* 48: 594–610. Doi: http://dx.doi.org/10.1177/0308518X15618205.

Smallwood, S. and K. Lynch. 2010. "An Analysis of Patient Register Data in the Longitudinal Study – What does it Tell Us about the Quality of the Data?" Newport, UK: ONS. *Population Trends* 141: 151–169.

StataCorp. 2016. Stata Statistical Software: Release14. StataCorp LP, College Station, Texas, US.

Statistics Canada (SC). 2012. *Canadian households in 2011: Type and growth*. Ontario, Canada: SC. Available at: https://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-312-x/98-312-x2011003_2-eng.cfm (accessed August 2017).

SC. 2015. *Population and Family Estimation Methods at Statistics Canada*. Ontario, Canada: SC. Available at: http://www.statcan.gc.ca/pub/91-528-x/2015001/ch/ch7-eng.htm (accessed August 2017).

Statistics New Zealand. 2013. *Evaluation of Administrative Data Sources for Subnational Population Estimates*. Wellington, NZ: SNZ. Available at: http://www.stats.govt.nz/~/media/Statistics/browse-categories/population/estimates-projections/eval-admin-subnat-pop-est/eval-admin-data-subnat-pop-est.pdf (accessed August 2017).

Trevena, P., D. McGhee, and S. Heath. 2013. "Location, Location? A Critical Examination of Patterns and Determinants of Internal Mobility Among Post-accession Polish Migrants in the UK." *Population, Space and Place* 19: 671–687. Doi: http://dx.doi.org/10.1002/psp.1788.

United Nations Economic Commission for Europe (UNECE). 2007. *Register-Based Statistics in the Nordic Countries: Review of Best Practices with Focus on Population and Social Statistics*. Geneva, Switzerland: UNECE. Available at: http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2764 (accessed August 2017).

UNECE. 2011. *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices*. Geneva, Switzerland: UNECE. Available at: https://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf (accessed August 2017).

UNECE. 2012a. *Conference of European Statisticians, The Swiss Census System: a Comprehensive System of Household and Person Statistics*. Geneva, Switzerland: UNECE. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2012/55-SP_Switzerland_-_the_swiss_census_system.pdf (accessed August 2017).

UNECE. 2012b. *Conference of European Statisticians. Lessons Learnt from a Mixed-Mode Census for the Future of Social Statistics*. Geneva, Switzerland: UNECE. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2012/37-SP_Germany.pdf (accessed August 2017).

UNECE. 2015. *Conference of European Statisticians. From the 2010 to the 2020 Census Round in the UNECE Region – Plans by Countries on Census Methodology and Technology*. Geneva, Switzerland: UNECE. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2015/mtg1/UNECE_paper_Paolo_draft_0925_rev2.pdf (accessed August 2017).

UNECE. 2016. *Conference of European Statisticians. Innovative Approaches Used in the 2016 Canadian Census*. Geneva, Switzerland: UNECE. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2016/mtg1/CES_GE.41_2016_16E.pdf (accessed August 2017).

United States Census Bureau (USCB). 2013. *America's Families and Living Arrangements: 2012*. Washington DC, US: USCB. Available at: https://www.census.gov/prod/2013pubs/p20-570.pdf (accessed August 2017).

USCB. 2015a. *2020 Census Operational Plan*. Washington DC, US: USCB. Available at: http://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-oper-plan.pdf (accessed August 2017).

USCB. 2016. *Methodology for the United States Population Estimates*. Washington DC, US: USCB. Available at: https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2010-2016/2016-natstcopr-meth.pdf (accessed August 2017).

Wallgren, A. and B. Wallgren. 2014. *Register-Based Statistics: Statistical Methods for Administrative Data*, *2nd Edition*. New York: John Wiley & Sons.

Received April 2016
Revised September 2017
Accepted October 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 83–106, http://dx.doi.org/10.1515/JOS-2018-0005

Typology and Representation of Alterations in Territorial Units: A Proposal

Francisco Goerlich[1] *and Francisco Ruiz*[2]

This article proposes a typology of boundary changes in territorial units at two points in time. The different types of changes are organized in a hierarchy and represented homogeneously, independently of the number of territorial units involved and of the changes to them. Each alteration is described precisely and unambiguously, and it is codified to allow the information to be treated automatically. In addition to providing efficient storage of the information about these changes, a canonical representation facilitates the automatic detection of inconsistencies in the database. At the same time, the typology allows us to define backward and forward equivalence rules, which helps in the task of generating homogeneous time series about territorial unit characteristics, such as population or surface area, or generating the full genealogy of a territorial unit over time. We also offer an application of the proposal to inconsistencies and error detection in the database *Alterations to the Municipalities in the Population Censuses since 1842* from the Spanish National Statistical Institute (INE).

Key words: Municipal boundary changes; population census; homogeneous series; standardized representation; equivalence rules; inconsistency criteria; typology.

1. Introduction

Alterations to territorial unit boundaries occures relatively frequent, particularly in the case of smaller units – municipalities or census tracts – or when a sufficiently long time perspective is considered. Historical records of such alterations, however, tend to be literal or descriptive, with no clear standardization, and therefore very difficult to deal with on paper and practically impossible to handle in digital format. Modern advances in Geographical Information Systems (GIS) have facilitated the generation of different administrative boundary layers of territorial units – countries, regions, counties, municipalities, and so on – with different reference dates; however they do not usually provide information about the changes between two reference periods.

[1] University of Valencia and Instituto Valenciano de Investigaciones Económicas (Ivie), Ivie, Calle Guardia Civil 22 Esc. 2 1°, 46020 Valencia, Spain. Email: Francisco.J.Goerlich@uv.es.
[2] University of Castilla-La Mancha, Information Systems and Technology Institute. Dpto. Tecnologías y Sistemas de Información, Escuela Superior de Informática, Universidad de Castilla-La Mancha, Edif. Fermín Caballero (3ª planta), Paseo de la Universidad 4, 13071 Ciudad Real, Spain. Email: Francisco.RuizG@uclm.es.
Acknowledgments: Useful comments by three reviewers and an associate editor have contributed to improving the article substantially; they are not, however, responsible for any remaining errors. Francisco Goerlich acknowledges funding from the BBVA Foundation-Ivie Research Program, and project ECO2015-70632-R of the Spanish Ministry of Education and Science. Results referred to in the article are available from the authors upon request.

Various studies have compiled historical boundaries in an attempt to provide a systematic record of the alterations to them (Tir et al. 1998). More recently several authors have worked on reconstructing these boundaries using GIS techniques, which require the genealogy of current territories to be reset (Gregory 2005; Gregory and Ell 2007; Flora et al. 2015), usually with the purpose of homogenizing specific characteristics over time (Marti-Henneberg 2005; Gregory and Ell 2006).

In the process of harmonizing European regional statistics over time, considerable effort has been devoted to modeling territorial changes in the regions (NUTS 2 and 3 in European Union terms) in order to create homogenous databases on a regional scale subject to temporal evolution of the territorial hierarchy (Ben Rebah et al. 2011; Milego and Ramos 2011). Official organizations and National Statistical Institutes (NSIs) also provide compilations of changes to administrative divisions at the various territorial scales into which the state is organized. However, as far as we know, no methodological proposal has attempted to harmonize, systematize, and computerize changes in territorial units over time. The European INSPIRE directive (Directive 2007/2/EC), establishing an infrastructure for spatial information in the European Union, would probably be the most suitable framework for this harmonization process. However, the *Technical Guidelines* on *Data Specification on Statistical Units* (INSPIRE 2013) only offer a very brief guide to the temporal representation of administrative units at the object level, and tend to emphasize the modeling, for all geographical objects, considering only a life cycle defined by the attributes *beginLifespanVersion* and *endLifespanVersion*, rather than explicitly linking the different temporal versions of the same spatial object. These simple rules are clearly unsuitable for a complete characterization of a territorial hierarchy subject to spatial changes over time that can be managed in a harmonized and systematized way, taking advantage of computerization. Different NSIs have attempted to go beyond a simple compilation of territorial alterations, trying to develop spatio-temporal information systems allowing for the temporal evolution of administrative boundaries in a consistent manner (Sindoni et al. 2002; Duque 2016).

This article aims to help bridge this gap by proposing a typology and representation of alterations in administrative territorial units whose boundaries are determined by criteria of political powers or state organization. The proposal has been developed from our experience with the database of alterations to Spanish municipalities since they first appeared as such in the population censuses in the mid-nineteenth century. The proposal was therefore based on literal descriptions of changes, and our efforts have focused on developing a system to codify the alterations that is consistent and could be automated. But clearly the proposed typology and coding principles are much more general, and can be used in tracking alterations in regions, cities or urban areas, as considered, for example, in the Urban Audit pan-European project. These territories are not always consistent with administrative divisions, and the efficient monitoring of changes in borders could be very useful.

However, the principles we detail below can be applied more generally, not only at other scales such as regions or census tracts, but also based on GIS layers at two moments in time, since the geometric accuracy of the two layers will be the same in the two periods. A simple "union" GIS operation between the two layers provides all the necessary information to implement the typology proposed in this article. All that is required are the

codes for the territorial units in the two periods, together with their surface areas, and the surface area of each polygon generated by the "union" operation. Although our application below is illustrated using historical municipalities, the typology proposed in this article has been successfully applied to generate types of alterations between census tracts at two moments in time using their GIS layers, without any other additional information on alterations to them. This information is contained in the mapping itself.

The article is structured as follows. The next section presents the proposed typology in detail and the criteria for it to be computerized. We then apply the alterations in Spanish municipalities to the database going back to the first population censuses of the mid-nineteenth century. This application allows us to examine how efficient our proposal is in detecting inconsistencies. The final section offers a brief conclusion.

2. Typology of Territorial Changes: A Proposal

One initial question that must be clarified from the start is *what do we understand by a territorial unit, from the perspective of a typology?* While this may appear to be a fairly trivial question, from the point of view of a typology it must be stressed that for our purposes, the only property by which a territorial unit can be unequivocally identified is a code. A name is not usually valid information to generate a typology, since it might not be the only name, two or more territorial units may have the same name, and a name may change at any given moment. The typology can also reflect name changes, as we shall see, even though they are not territorial changes. It is true, however, that other types of geographical entities, such as census tracts, have no name and the only information of consequence is their code. NSIs and international institutions are fully aware of this need to identify territorial units using unique codes.

This clarification is important because if a territorial unit code changes, and this is the only change made to it, from the perspective of our typology it will be treated as a 'territorial' alteration: one territorial unit disappears and another identical one is created. Fortunately, the typology allows these cases to be clearly identified.

Based on the premise that each territorial unit has its own code, and drawing on our experience of municipal alterations in a historical context, we propose a typology to classify the categories of changes to territorial units that is complete in that it incorporates all existing situations, but also open as new cases can be added to it. Because the typology was created on the basis of a specific experience – historical alterations in Spanish municipalities – certain unusual cases arise; however it is clear that the underlying philosophy can easily be adapted to other similar situations. The typology includes a codification that allows for efficient treatment of the information with computer systems, and databases in particular. We start from a two-dimensional classification and aim to establish a typology that meets the following characteristics:

i) It distinguishes the cause of the change, which may affect several territorial units at the same time, from the alteration or effect that this change produces in each of the territorial units.
ii) It considers all possible types of change. Notwithstanding, it can be extended to incorporate new types of changes or situations.

iii) It includes a textual definition for each type, as well as a precise specification that indicates the 'backward' and 'forward' rules to generate a homogenous structure of territorial units according to a given criterion, for example, homogeneous population series according to the structure of a given year, or to derive the genealogy of a territorial unit.
iv) It establishes a 'canonical form' of representing changes, namely, a common format that enables all possible situations to be dealt with by creating a database for consultation.
v) It establishes criteria to detect inconsistencies.

2.1. *Double Perspective*

When analyzing the possible types of change that can affect territorial units, and the best way of representing them, it is important to distinguish two perspectives or dimensions:

a) The cause or the type of change itself: various units merged into one new unit, one unit integrated into another, etc.
b) The alteration or specific effect that this type of change has on each of the affected territorial units: elimination, creation, modification, etc.

For example, if unit A is integrated into unit B, from the first perspective we would refer to it as an 'integration' change type, whereas from the second perspective we would say two alterations had occurred: an elimination, territorial unit A disappears as it is integrated into B, and a modification, since the territory of territorial unit B increases with the integration of A. This idea of pairs of alterations associated to types of changes is crucial to our proposal.

The following distinction must always be maintained: a change (of a certain type) is reflected (manifested) in one or more alterations. The term 'alteration' and the term 'change' must always refer to these two related ideas or concepts. From the outset the meaning of the terms used must be precisely understood so that there is no ambiguity in the way they are applied.

2.2. *Hierarchy of Basic Change Types*

The first perspective provides the base for the following hierarchy of change types that, at the first level, distinguishes between territorial and nonterritorial changes as by definition the territory is the key element of all Territorial Units (TU).

TERRITORIAL:

- *Territorial units are neither created nor eliminated:*
  - Transfer (T): one TU transfers part of its territory to another TU or other TUs.
  - Exchange (P): two TUs exchange part of their territories.
- *Territorial units are created and are not eliminated:*
  - Segregation (S): one part of a TU is separated to create a new TU.
  - Partial merger (Fp): parts of two or more TUs are combined to form a new TU.
  - Unspecified appearance (O): a new TU emerges without any specific information.

- *Territorial units are eliminated and are not created:*
  - Integration (I): a TU is fully incorporated into another TU or other TUs.
  - Distribution (R): a TU disappears when its territory is distributed among two or more pre-existing TUs.
  - Unspecified disappearance (O): a TU disappears without any specific information.
- *Territorial units are created and eliminated:*
  - Code change (C): a TU's code is changed (in practice, the old TU is eliminated and a new one is created).
  - Merger (F): two or more TUs are combined to form a new one.
  - Division (D): a TU is divided into two or more new TUs.

NONTERRITORIAL:

- Change of designation (G): a TU's name or designation is changed.
- Annotation (A*x*): other changes or information that do not affect the territory. A different letter '*x*' can be used for each situation we are interested in identifying, thus allowing the typology to be extended.

It should be noted that the cases of *Unspecified appearance* and *Unspecified disappearance* mentioned above are included because they appear in the historical lists of alterations, but they are infrequent in the present period. In any case, these situations do not arise in closed territorial systems and those with well-defined administrative divisions.

We believe the terminology is simple yet precise and each term is used unambiguously. Hence, a distribution indicates that the territory of the territorial unit that disappears results in an increase in the territory of other pre-existing territorial units, whereas a division indicates the appearance of new territorial units. In both cases the original territorial unit disappears, otherwise it would be considered as a territory transfer; but what happens to the destination territorial units, of which there must be more than one otherwise we would be dealing with an integration or a code change, depends on the specific term used.

2.3. Types of Alterations

Each type of change identified in the above hierarchy gives rise to a certain alteration or effect in each territorial unit involved in that change. There are four types of alterations and, as with the change types, we distinguish between those that have territorial effects and those that do not:

- Creation (C): the TU appears.
- Elimination (E): the TU disappears.
- Modification (M): the TU's territory changes.
- Others (O): a nonterritorial characteristic of the TU changes, such as its designation.

2.4. Codification: Canonical Representation

The alteration, and the type of change that causes it, is codified using a set of two or three letters that form a descriptive key representing the specific situation of a given territorial unit when the alteration occurs. The first letter represents the alteration in the municipality.

The remaining part of the descriptive key represents the type of change that causes the alteration. The full list of the keys used in the proposed typology is presented in Table 1, and a detailed description of all possible categories of changes, including a graphical representation, is provided in the Appendix. As can be seen from Table 1, usually two letters are sufficient for the codification, but sometimes an additional letter is convenient to denote partial splits or increasing/decreasing alterations by transfer of parts of other territoties.

We have looked for a way of representing the types of change, and their associated alterations, that can be used to represent all possible situations, and therefore generate a codification that can be represented and treated automatically. This way of structuring information corresponds to the concept of 'first normal form' used in the field of databases. In our case, this has two main consequences:

a) We represent all the changes by means of the list of the alterations they generate.
b) Each alteration is defined as a relation between two, and only two, territorial units.

Thus, the three central elements of the canonical representation are the two codes for the territorial units involved, and the key that identifies the relation between them (Table 1). The first territorial unit is the one that has the alteration, represented by the first letter of the key, while the second is related to the first by the type of change.

As an illustration, let us suppose that territorial unit A disappears because it is incorporated into territorial unit B. The 'standardized' representation of this 'integration' type change is formed by the following pair of alterations:

- (A, EI, B) $\rightarrow A$ is eliminated (E) by integration (I) into B; and
- (B, MI, A) $\rightarrow B$ is modified (M) because it integrates (I) A.

The keys that are related to the two territorial units involved are those given in Table 1, and as noted above, they use the first letter to denote the type of alteration (Elimination, E; or Modification, M), and the second and third letters to denote the type of change (Integration, I). Moreover, each key in Table 1 –first column– has a paired key –last column–, which denotes a reverse perspective, when we switch the codes of the territorial units involved in the alteration.

Nonterritorial change types only involve one territorial unit (code). For this reason, they are represented with a single alteration, in which the territorial unit code is repeated. For example, the change of name of F is represented as:

- (F, OG, F): the TU F was called . . .

Territorial changes can affect more than two territorial units at the same time, which is a fairly common case. For example, a division affects at least three territorial units: the one that is divided, which disappears, and the two or more that emerge from the division, which are created. The representation of this change must indicate all the pairs of territorial units related by an alteration caused by the change. Thus, a division in which territorial unit A disappears because it is divided into territorial units X, Y, and Z is represented with the following alterations:

- (A, ED, X) $\rightarrow A$ is eliminated (E) by division (D) into X, among others;
- (A, ED, Y) $\rightarrow A$ is eliminated (E) by division (D) into Y, among others;

Table 1. Keys representing alterations and changes.

Key	Alteration	Change	Relation (between the two municipalities)	Pair
CC	C Creation	Code Change	is created by code change of	EC
CD		Division	is created by division of	ED
CF		Merger	is created by merger, among others, of	EF
CFp		Partial merger	is created by merger of one part, among others, of	MFp
CO		Unspecified appearance	is created from territories not registered as municipalities	CO
CS		Segregation	is created by segregation of	MS
EC	E Elimination	Code change	is eliminated by code change of	CC
ED		Division	is eliminated by division, among others, into	CD
EF		Merger	is eliminated by merger into	CF
EI		Integration	is eliminated by integration into	MI
EO		Unspecified disappearance	is eliminated without further information	EO
ER		Distribution	is eliminated because it is distributed, among others, to	MR
MFp	M Modification	Partial merger	is modified because one part is merged with others to form	CFp
MI		Integration	is modified because it integrates	EI
MP		Exchange	is modified because it exchanges territories with	MP
MR		Distribution	is modified because, among others, it receives a part of the distribution of	ER
MS		Segregation	is modified because it is segregated	CS
MTc		Transfer	is modified, increasing, because it receives a transfer of a part of	MTd
MTd		Transfer	is modified, decreasing, because it transfers a part to	MTc
OA*x*	O Others (Nonterritorial)	Annotation *x*	descriptive annotation of *x*	OA*x*
OG		Name change	was called	OG

Note: Only one set of alterations in nonterritorial types of change is shown, namely *x,* simply to show that the list is open.

- (A, ED, Z) → A is eliminated (E) by division (D) into Z, among others;
- (X, CD, A) → X is created (C) by division (D) of A;
- (Y, CD, A) → Y is created (C) by division (D) of A;
- (Z, CD, A) → Z is created (C) by division (D) of A.

This standardized representation is what we refer to as 'canonical representation', which has the shortest possible label that can be used atemporally. However, when we have a sequence of alterations over time, for example municipalities in various censuses, the representation of each pair of territorial units related by an alteration must include other informative elements as well as the territorial unit codes and keys, for example, a temporal dimension of when the alteration took place, or the old and new names in the case of name changes. These extensions can easily be accommodated.

2.5. *Representation of Complex Change Types*

As well as the changes corresponding to the basic types presented above, other complex types of changes can occur where several basic types of change are combined. The proposed typology covers all the basic types needed to create, through combinations, any change, however complex it may be, by applying exactly the same ideas as for the basic types. For example, let us suppose that the change shown in Figure 1, in which territorial units X and Y are created as a result of the division of unit A, and unit B receives part of territory A. The labels that fully describe the change will be:

- (A, ED, X) → A is eliminated (E) by division (D) into X, among others;
- (X, CD, A) → X is created (C) by division (D) of A;
- (A, ED, Y) → A is eliminated (E) by division (D) into Y, among others;
- (Y, CD, A) → Y is created (C) by division (D) of A;
- (A, ER, B) → A is eliminated (E) because it is distributed (R) to B, among others;
- (B, MR, A) → B is modified (M) because, among others, it receives a part of the distribution (R) of A.

2.6. *Detecting Inconsistencies*

The 'canonical representation' has other advantages as well as allowing the homogeneous representation of all types of situations in modifications to municipalities. One particularly

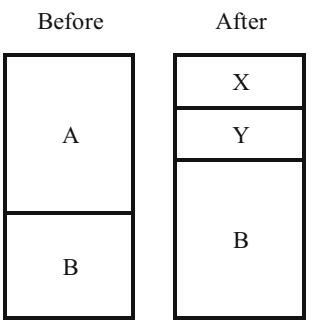


Fig. 1. Example of a complex change.

relevant advantage is that it automatically detects inconsistencies in the changes recorded. To do this, it takes advantage of the fact that alterations referring to territorial changes, those that link two or more units, must always be in pairs. Hence, in the examples given in the previous sections we can see that all type 'CD' alterations between units *A* and *B* are paired with another 'ED' type alteration between *B* and *A*. This property is due, simply, to the fact that the 'paired' alterations refer to the same information, but one is expressed actively and the other, passively. For example,

- (*A*, CD, *B*) $\rightarrow$ *A* is created (C) by division (D) of *B*;
- (*B*, ED, *A*) $\rightarrow$ *B* is eliminated (E) by the division (D) into *A*, among others.

In general, we can express a territorial alteration that affects a pair of territorial units *A* and *B* actively in terms of *A* (and passively in terms of *B*) or, conversely, actively in terms of *B* (and passively in terms of *A*). The pair of each type of alteration is shown in the last column of Table 1.

This property is very useful for detecting inconsistencies in changes to territorial units in the databases. One only has to check that the corresponding pairings match. Any unexpected pairing between two types of alterations will indicate an inconsistency in the base information that represents the changes. We apply this technique in Section 3.

2.7. Rules of Homogenization and Generating Genealogies

Another very useful application of the 'canonical representation' is the generation of homogenous series from the territorial unit structure existing at a given moment. Similarly, this allows us to generate the genealogy of changes in a territorial unit over time.

To do this we have identified backward and forward equivalence rules for each type of change presented in Table 1. These rules allow us to automate the knowledge we have about the different parts that a territorial unit, existing in a given moment, has in the territorial units at other moments in time, whether past or future, and regardless of whether or not these territorial units exist in the reference period. Thus, we establish a systematic way of creating homogeneous series for groups of territories, subject to territorial changes over time, based on applying a type of equivalence rule between the territorial units existing before the change and those existing afterwards. Hence, for each territorial unit *A* existing before a change, the forward rules allow us to establish the unit, units and/or parts of territorial units that are 'equivalent' to *A* after the change. Similarly, for each territorial unit B existing after the change, the backward rules allow us to establish the unit, units and/or parts of territorial units that are 'equivalent' to *B* before the change.

For example, let us suppose a change took place in year *T* that caused territorial unit *A* to disappear because it was divided into two new ones, *B* and *C*, as illustrated in Figure 2. Before year *T* only *A* existed, whereas after year *T*, *B,* and *C* exist, but *A* does not. The only forward rule will be $A \rightarrow B + C$. Its application to the case of the population will imply, for example, that we must compare the population of *A* before year *T* with the sum of the populations of *B* and *C* after year *T*. There are two backward rules for the same example, one for each territorial unit existing after year *T*: $B \rightarrow A(b)$ and $C \rightarrow A(c)$, where $A(x)$

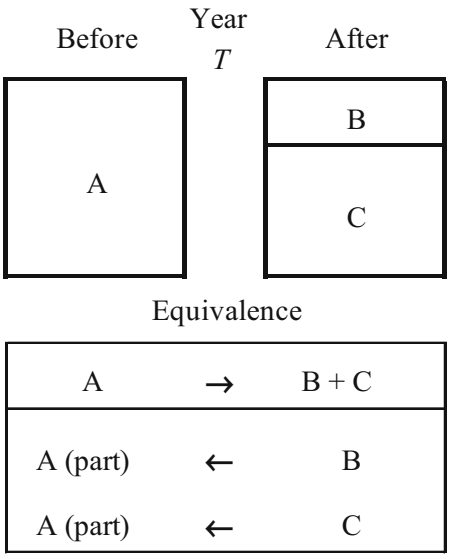


Fig. 2. Example of equivalence rules.

represents the part of territorial unit *A* that was transferred to territorial unit *X*. If we are interested in populations, applying these rules will mean that the population of *B* after year *T* should be compared with the population of a certain part of territorial unit A before year *T*; and the same for territorial unit *C*.

As a real example of the above rules, we may consider the case of *Santurce* (*A* in Figure 2), a municipality that disappeared at the beginning of the twentieth century by splitting its territory into two new municipalities, *Santurce Antiguo* and *Santurce Ortuella* (*B* and *C* in Figure 2). Application of the forward rule is very simple, because it entails comparing the particular characteristic of *Santurce*, that is population or surface, with the sum of the characteristic of the two new municipalities created by division, *Santurce Antiguo* and *Santurce Ortuella*. The application of the backward rules is not so straightforward, since it entails knowing the characteristic of the part of *Santurce* assigned to each of the new municipalities. For example, if we are interested in the homogeneous population of *Santurce Ortuella*, we must know or estimate the part of the population of *Santurce* that is in the actual territory of *Santurce Ortuella*. These backward and forward equivalence rules for each type of change included in Table 1 are fully described in the Appendix.

To automatically apply the equivalence rules it must be kept in mind that each change is represented in the form of one or various alterations. For this reason an algorithm has also been designed that allows the equivalence rules to be applied for a set of territorial units subject to a set of alterations. The algorithm is based on applying the substitutions shown in Table 2, where each row displays what substitutes a territorial unit *X* affected by an alteration (X, $<key>$, Y) by making a forward or backward homogenization. With this algorithm we can generate complete genealogies for a given territorial unit. Naturally, the rules are only applied to territorial type changes as they are not necessary for nonterritorial type changes. Some NSIs have been interested in keeping track of these forward and

Table 2. Substitutions for the automatic application of the equivalence rules.

		X is substituted by	
Key	Relation (X, <key>, Y)	Forward	Backward
CC	is created by code change of	none	Y
CD	is created by division of	none	Y
CF	is created by merger of, among others,	none	Y(part)
CFp	is created by merger of a part of, among others,	none	Y(part)
CO	is created from territories not registered as municipalities	none	none
CS	is created by segregation of	none	Y
EC	is eliminated by code change of	Y	none
ED	is eliminated by division into, among others,	Y(part)	none
EF	is eliminated by merger into	Y	none
EI	is eliminated by integration into	Y	none
EO	is eliminated without additional information about the circumstances	none	none
ER	is eliminated because it is distributed to, among others,	Y(part)	none
MFp	is modified because one part is merged with others to form	X(part) + Y(part)	X
MI	is modified because it integrates	X	X(part) + Y(part)
MP	is modified because it exchanges territories with	X(part) + Y(part)	X(part) + Y(part)
MR	is modified because, among others, it receives part of the distribution of	X	X(part) + Y(part)
MS	is modified because it is segregated	X(part) + Y(part)	X
MTc	is modified, increasing, because it receives a transfer of a part of	X	X(part) + Y(part)
MTd	is modified, decreasing, because it transfers a part to	X(part) + Y(part)	X

backward changes for developing spatio-temporal information systems (Sindoni et al. 2002; Duque 2016). Our typology, and the derived homogeneous rules, provides a complete and neat solution to this problem. An application of this technique to generate homogeneous population series at the municipal level for all the twentieth century censuses based on the municipality structure in the 2011 Census can be found in Goerlich et al. (2015).

3. Application to Detect Inconsistencies in the INE Database *Alterations to the Municipalities in the Population Censuses Since 1842*

This section provides an illustration of how the above typology was used to refine inconsistencies in the database of alterations to Spanish municipalities since their first appearance in censuses in the mid-nineteenth century (INE 2005).

The first attempts to compile the alterations to Spanish municipalities were made by the INE during the work carried out for the 1981 population census (INE 1981a, 1981b). The

aim was to discover which municipalities had disappeared since 1900, without offering any type of systematization, and to find out which municipalities had absorbed them and detect any name changes. Following these initial endeavors, various authors have studied alterations to municipalities, essentially with the aim of constructing homogeneous population series according to the structure of a given census (García 1985, 1994; Goerlich et al. 2006).

Eventually, INE released a database of the original population censuses at municipal level with literal descriptions of all inter-census changes: "*Alterations to the Municipalities in the Population Censuses since 1842*" (INE 2005). The current municipal codification system at a national level dates from the 1970 Census and comprises five digits in the form *PPNNN*, where *PP* is the provincial code and *NNN* a serial number for the municipality within each province. The INE, aware of the importance of a codification system that would allow municipalities to be accurately traced over time, extended the current codification to the municipalities that had disappeared before the 1970 Census. In addition, this process involved tracing the names of municipalities through the censuses, and creating a gazetteer of names, which is the database we use.

After the INE made its database public, the Ministry of Public Administrations prepared a database of municipal alterations since 1842 in which, first, the type of alteration was identified: creation (C), extinction (E), or modification (M) of the municipality in question; and second, the cause of the alteration was identified according to a series of keys (MAP 2008). Unfortunately this database does not have codes, which means it is practically unworkable for our purposes, and it focuses on name changes; however, it is this idea of double entry –alteration *versus* cause – that underlies our typology proposal.

3.1. Detecting Inconsistencies

The first step was to annotate the literal descriptors in the INE (2005) database with the classification keys in Table 1. This preliminary stage was peformed using standard Microsoft tools for extracting, loading, and transforming data – Access and Excel-Power Query – and required a fairly large amount of work, which shows again the benefits of a systematic treatment of the information. Once all the alterations are expressed in the 'canonical form', we apply the technique to detect inconsistencies based on the pairing of alterations presented in Subsection 2.6. The typology defined establishes that the alterations corresponding to territorial changes must always appear in pairs. This information is displayed in Table 3, generated from Table 1.

This information is used directly to detect inconsistencies in the alterations. It consists of identifying the situations in which for an alteration (row or entry in the table) between the territorial units A and B of type P in year T, (A, P, B, T), there is no paired alteration (B, Q, A, T), where Q is the pair key that corresponds to P according to Table 3. The result was that of the 13,424 alterations included in the table in the canonical form with the original INE data, 175 errors or inaccuracies were found, affecting a total of 334 alterations.

Once an inconsistency has been discovered the typology considerably reduces the effort involved in finding the problem, as the origin of the error can be pinpointed to within a few

Table 3. Pairing of types of alterations for each type of territorial change.

Type of change	*Key* 1	*Key* 2
Code change	CC	EC
Division	CD	ED
Merger	CF	EF
Partial merger	CFp	MFp
Integration	MI	EI
Exchange	MP	MP
Distribution	MR	ER
Segregation	CS	MS
Transfer	MTc	MTd

lines, from more than one hundred thousand in the original data. All the inconsistencies detected were investigated and corrected, resulting in a table of 13,415 alterations, of which 8,935 correspond to territorial changes. These figures are reported in Table 4, and the pairing rules are now satisfied in all cases.

It is worth mentioning that the statistics for territorial changes after the corrections displayed in Table 4 does not include all the errors found in the INE database (2005), but only those detected automatically by means of the pairing technique. Goerlich et al. (2015, Sec. 2.2 and 3.3) provide all the errors detected, together with the final statistics for the database of alterations to Spanish municipalities once it had been refined.

The full list of errors and inaccuracies can be consulted in Ruiz and Goerlich (2014). The corrections involved modifications to:

- The type of alteration (263 occasions).
- The census year (27 occasions).

Table 4. Statistics of territorial changes after the corrections.

Change	Alteration 1	Alteration 2	No. cases
Code change	CC	EC	60
Division	CD	ED	32
Merger	CF	EF	446
Partial merger	CFp	MFp	3
Segregation	CS	MS	452
Integration	EI	MI	3,380
Distribution	ER	MR	61
Exchange*	MP	MP	1
Transfer	MTc	MTd	16
Total pairs			**4,451**
Creation from others	CO	–	9
Unspecified disappearance	EO	–	24
Total			**8,935**

*Each exchange also involves two alterations, both MP type, but with the pair of municipalities the opposite way round.

Note: 4,451 pairs = 8,902 alterations.

- The code of the second municipality (22 occasions).
- Eliminating the alteration (14 occasions).
- Adding a new alteration (5 occasions).
- The type of alteration, together with the code of the second municipality (2 occasions).
- The type of alteration, together with the census year (1 occasion).

A comparative review of these errors, their corrections and the content of the entries on the INE website uncovered the reasons that most probably caused these errors in the original INE data. The following are highlighted:

a) Errata or isolated errors, such as a mistake in writing down the code for a municipality in an entry (for example, 1717007 instead of 17007) or in a date (for example, 1960 instead of 1860). Another frequent case involved a municipality being assigned the code that came either before or after it alphabetically in the province list. This seems to suggest that the entries recorded on the INE website were introduced manually one by one and, therefore, without any effective possibility of checking for consistency.
b) Inaccurate use of terms. The same verb was used for different situations, that is, different types of changes (for example, the verbs 'to group' for mergers or partial mergers, and 'to integrate' for distributions, segregations or divisions). It is very difficult to discover these situations through a manual review of the entries; however, they are quickly identified by applying the classification, codification and pairing techniques presented here.
c) Contradictory situations noted in two or more municipalities associated with the same change type. This usually occurs in municipalities where various types of alterations coincided (for example, a municipality grows in size because it incorporates others, and at the same time, because other municipalities transfer part of their territory to it). The system of representation used in our study was an essential tool to avoid these problems as it allowed us to 'visualize' the whole picture of all the alterations associated with each change.

3.2. Detecting Codification Errors

The pairing technique allowed us to uncover situations where the entries on the INE website were inconsistent. However, this does not rule out the possibility of other errors that are not reflected in inconsistencies. For example, if there is a mistake in the code for a municipality in the province of Madrid, 28979 instead of the correct one, 28079, and this mistake appears in every reference to this municipality, there is no way of knowing that this is the wrong code without comparing it with another external data source. To ensure that the codes the INE assigned in its alterations database are the correct ones, we checked that for each code, the municipality corresponds to that indicated in the official 1970 Census, which is when the municipal codes were first established, based on the alphabetical order generated in that census.

This verification revealed that the INE had wrongly assigned codes in the four municipalities reported in Table 5, all of which belong to the province of Teruel. On

Table 5. Errors in assigning codes in the INE alterations.

Code in the INE alterations website	Municipality	Correct code
44138	Luco de Bordón	44139
44139	Luco de Jiloca	44140
44140	Lledó	44141
44141	Loscos	44138

observing their names and the connection between the wrong codes and the correct codes, it seems that the four mistakes were due to the same human error on entering the data for the INE website, which arose because the alphabetical order created by computers differs from the traditional Spanish alphabetical order, which considered 'LL' as a separate letter following all other entries beginning with 'L'.

4. Conclusions

In this article we have presented a proposal for a typology that defines and classifies different types of change that may occur in different territorial units between two moments in time. The types of change were classified into two categories: territorial and nonterritorial. The first category includes various groups, depending on whether they create, eliminate or modify territorial units. The second group includes types of change that do not involve territorial modifications or appearances or disappearances in the list of territorial units, such as for example, name changes.

Each type of change is represented in what we term 'canonical form', which consists of expressing its effect (alterations) between pairs of territorial units. This allows every possible situation to be represented, however complex it may be, in a common format. Additionally, in this type of representation territorial alterations are presented in pairs, such that each type P alteration, between territories A and B, corresponds to another type Q alteration between B and A. In this way, errors due to inconsistencies in the original data sources can be detected automatically. This technique was applied to the data on municipal alterations available on the INE website (2005), which covers all alterations occurring between the censuses of 1842 and 2001. As a result of this procedure, 175 inconsistencies and inaccuracies in the database, as well as four wrongly assigned codes, were detected and corrected.

Appendix

Detailed Description of Basic Types of Change

This Appendix describes in detail the types of changes listed in the hierarchy presented in the study. A schema is used for each type of change that provides the specific name of each change type, a textual description, a graphical illustration of the state before and after the change, canonical representation – the list of alterations involved, using the keys from Table 1, forward and backward equivalence rules to generate homogeneous constructions over time, and an example taken from the INE database *Alterations to the Municipalities in the Population Censuses since 1842*. For this reason, territorial unit is equivalent to municipality in the annex. Additional comments are provided in some cases.

1. Territorial, without creation or elimination of municipalities

This includes types of changes that involve a territorial change, but neither the creation of new municipalities nor the elimination of pre-existing ones.

Type of change:	**TRANSFER**
Description:	One municipality transfers part of its territory to another or other pre-existing municipalities.
States:	Municipality *B* transfers part of its territory to municipality *A*. Before: **A** / **B** After: **A** / **B**
Representation:	(*A*, MTc, *B*): *A* is modified, increasing, because it receives a part of *B*. (*B*, MTd, *A*): *B* is modified, decreasing, because it transfers a part to *A*.
Forward rules:	$A \rightarrow A(a)$ $B \rightarrow B+A(b)$
Backward rules:	$A \rightarrow A+B(a)$ $B \rightarrow B(b)$
Example:	**09323 Rojas**: Between the 1897 Census and the previous one, the municipal area grew because it incorporated 095101 (Quintana Urria) which was previously an entity of 09071 (Carcedo de Bureba). => (09323, MTc, 09071) (09071, MTd, 09323)

Type of change:	**EXCHANGE**
Description:	Two municipalities exchange parts of their territories.
States:	Municipalities *A* and *B* exchange parts of their territories. Before: **A** / **B** After: **A** | **B**
Representation:	(*A*, MP, *B*): *A* is modified because it exchanges territories with *B*. (*B*, MP, *A*): *B* is modified because it exchanges territories with *A*.
Forward rules:	$A \rightarrow A(a)+B(a)$ $B \rightarrow A(b)+B(b)$
Backward rules:	$A \rightarrow A(a)+B(a)$ $B \rightarrow A(b)+B(b)$
Example:	**25138 Montgai**: Between the 1877 Census and the previous one, the municipal area altered because it transferred 25092 (Floresta) to 25153 (Omellons) and received 255079 (Butsenit) from 25153 (Omellons). => (25138, MP, 25153) (25153, MP, 25138)
Comments:	This is the same as two transfers in the opposite direction, one from *A* to *B* and another from *B* to *A*.

2. Territorial, where municipalities are created but not eliminated

This includes types of territorial changes in which a municipality is created but no municipality is eliminated.

Type of change:	**SEGREGATION**
Description:	A new municipality is created by the segregation of part of another municipality's territory.
States:	The new municipality *A* is created by the segregation of part of municipality *B*. Before: **B** After: **A** / **B**
Representation:	(*A*, CS, *B*): *A* is created by the segregation of *B*. (*B*, MS, *A*): *B* is modified because *A* is segregated from it.
Forward rules:	$B \rightarrow A+B$
Backward rules:	$A \rightarrow B(a)$ $B \rightarrow B(b)$
Example:	**02004 Albatana**: This municipality appeared between the 1920 Census and the previous one because it was segregated from municipality 02056 (Ontur). => (02004, CS, 02056) (02056, MS, 02004)

Type of change:	**PARTIAL MERGER**
Description:	Two or more municipalities cede parts of their territories to create a new municipality.
States:	Municipalities *B* and *C* transfer part of their territories to create the new municipality *A*. Before: **B** / **C** After: **B** / **C**, **A**
Representation:	(*A*, CFp, *B*): *A* is created by the merger of onepart, among others, of *B*. (*A*, CFp, *C*): *A* is created by the merger of onepart, among others, of *C*. (*B*, MFp, *A*): *B* is modified because one part is merged with others to create *A*. (*C*, MFp, *A*): *C* is modified because one part is merged with others to create *A*.
Forward rules:	$B \rightarrow B+A(b)$ $C \rightarrow C+A(c)$
Backward rules:	$A \rightarrow B(a)+C(a)$ $B \rightarrow B(b)$ $C \rightarrow C(c)$
Example:	**30902 Alcázares, Los**: This municipality appeared in the 1991 Census as a result of the merger of two parts, one from 30035 (San Javier) and the other from 30037 (Torre Pacheco). => (30902, CFp, 30035) (30902, CFp, 30037) (30035, MFp, 30902) (30037, MFp, 30902)

Type of change:	**UNSPECIFIED APPEARANCE OR CREATED FROM OTHER TERRITORIES**
Description:	A new municipality is created from unspecified territories or from territories that were not municipalities.
States:	Municipality *A* is created from territories that are not municipalities. Before: (empty) After: **A**
Representation:	(*A*, CO, *A*): *A* is created from territories that are not registered as municipalities.
Forward rules:	not applicable
Backward rules:	not applicable
Example:	**52001 Melilla**: This municipality appeared between the 1877 Census and the previous one; previously it was defined as a '*plaza de soberanía*' or sovereign stronghold. => (52001, CO, 52001)
Comments:	Used for special circumstances such as the change of category of certain sovereign strongholds in North Africa to municipalities. This alteration is not noted as such in the INE database, where Melilla appears with its present code, 52001, in the 1877 Census with no additional comment.

3. Territorial, where municipalities are eliminated but not created

This includes types of territorial changes involving the elimination of a pre-existing municipality without the creation of any new municipalities.

Type of change:	**INTEGRATION**
Description:	A pre-existing municipality is fully incorporated into one or more other municipalities.
States:	Municipality *B* disappears because it is integrated into municipality *A*. Before → After **A** / **B** → **A**
Representation:	(*A*, MI, *B*): *A* is modified because it is integrated into *B*. (*B*, EI, *A*): *B* is eliminated because it is integrated into *A*.
Forward rules:	$A \rightarrow A(a)$ $B \rightarrow A(b)$
Backward rules:	$A \rightarrow A+B$
Example:	**01016 Bernedo**: The surface area of this municipality increased between the 1981 Census and the previous one because it incorporated 01005 (Arlucea Marquínez). => (01016, MI, 01005) (01005, EI, 01016)

Type of change:	**DISTRIBUTION**
Description:	A municipality disappears because all its territory is distributed among two or more pre-existing municipalities.
States:	Municipality *C* disappears because it is distributed between municipalities *A* and *B*. Before → After **C** (**A** / **B**) → **A** / **B**
Representation:	(*A*, MR, *C*): *A* is modified because it receives, among others, one part from the distribution of *C*. (*B*, MR, *C*): *A* is modified because it receives, among others, one part from the distribution of *C*. (*C*, ER, *A*): *C* is eliminated because it is distributed to *A*, among others. (*C*, ER, *B*): *C* is eliminated because it is distributed to *B*, among others.
Forward rules:	$A \rightarrow A(a)$ $B \rightarrow B(b)$ $C \rightarrow A(c)+B(c)$
Backward rules:	$A \rightarrow A+C(a)$ $B \rightarrow B+C(b)$
Example:	**49001 Abelón**: This municipality disappeared between the 1981 Census and the previous one because it was integrated, in parts, into municipalities 49023 (Bermillo de Sayago) and 49152 (Pereruela). => (49001, ER, 49023) (49001, ER, 49152) (49023, MR, 49001) (49152, MR, 49001)

Type of change:	UNSPECIFIED DISAPPEARANCE
Description:	A municipality disappears and no information is available on the reason why or where its territory goes.
States:	Municipality *A* disappears with no further information about the circumstances. Before: **A** After:
Representation:	(*A*, EO, *A*): *A* is eliminated with no further information about the circumstances.
Forward rules:	not applicable
Backward rules:	not applicable
Example:	**255122 Cícera**: This municipality, which appears in the 1842 Madoz Census, is not found in the 1857 Census (it belonged to Partido de Cervera). => (255122, EO, 255122)
Comments:	In practice, this circumstance was only found in the 1857 Census with regard to municipalities recorded in the previous census of 1842 (the first to provide a list of municipalities). Presumably this is because the 1842 Census was not totally accurate in the way it distinguished between municipalities and submunicipal entities, since the concept of the municipality had not yet been fully clarified.

4. Territorial, where municipalities are created and eliminated

This includes the types of territorial changes involving both the creation of a new municipality and the elimination of a pre-existing municipality.

Type of change:	CODE CHANGE
Description:	The code for a municipality changes; that is, it disappears and another one appears in its place, although it is really the same one (with the same territory).
States:	Municipality with code *B* changes to *A*. Before: **B** After: **A**
Representation:	(*A*, CC, *B*): *A* is created due to the change of the code for *B*. (*B*, EC, *A*): B is eliminated due to the change of the code for *A*.
Forward rules:	$B \rightarrow A$
Backward rules:	$A \rightarrow B$
Example:	**27901 Baralla**: This municipality appeared between the 1981 Census and the previous one because its name changed, and municipality 27036 (Neira de Jusá) disappeared. => (27901, CC, 27036) (27036, EC, 27901)
Comments:	This is not, strictly speaking, a territorial change, but because what identifies the municipalities is the code, in practice it is the same as eliminating one municipality and creating a new one, which has the same territory and all its characteristics. Code changes are often due to an error when a simple name change occurs. A special case is when the code change is the result of a municipality transferring to a different province.

Type of change:	**MERGER**
Description:	Two or more municipalities disappear because they combine to form a new municipality.
States:	Municipalities *B* and *C* disappear because they are merged into a new municipality *A*. Before: **B**, **C** After: **A**
Representation:	(*A*, CF, *B*): *A* is created from the merger, among others, of *B*. (*A*, CF, *C*): *A* is created from the merger, among others, of *C*. (*B*, EF, *A*): *B* is eliminated by merger into *A*. (*C*, EF, *A*): *C* is eliminated by merger into *A*.
Forward rules:	$B \rightarrow A(b)$ $C \rightarrow A(c)$
Backward rules:	$A \rightarrow B+C$
Example:	**17902 Forallac**: This municipality appeared between the 1981 Census and the previous one following the merger of municipalities 17072 (Fonteta), 17131 (Peratallada) and 17235 (Vulpellach). => (17902, CF, 17072) (17902, CF, 17131) (17902, CF, 17235) (17072, EF, 17902) (17131, EF, 17902) (17235, EF, 17902)

Type of change:	**DIVISION**
Description:	A municipality disappears because it is divided into two or more new municipalities.
States:	Municipality *C* disappears because it is divided into municipalities *A* and *B*. Before: **C** After: **A**, **B**
Representation:	(*A*, CD, *C*): *A* iscreated by the division of *C*. (*B*, CD, *C*): *B* iscreated by the division of *C*. (*C*, ED, *A*): *C* is eliminated by the division into *A*, among others. (*C*, ED, *B*): *C* is eliminated by the division into *B*, among others.
Forward rules:	$C \rightarrow A+B$
Backward rules:	$A \rightarrow C(a)$ $B \rightarrow C(b)$
Example:	**48531 Santurce**: This municipality disappeared between the 1910 Census and the previous one, because it was integrated, in parts, into municipalities 48082 (Santurce Antiguo) and 48083 (Santurce Ortuella). => (48082, CD, 48531) (48083, CD, 48531) (48531, ED, 48082) (48531, ED, 48083)

5. Nonterritorial

This includes the types of change that involve no alterations to the municipalities' territory or to the list of existing municipalities.

Type of change:	**CHANGE OF DESIGNATION**
Description:	A municipality's name changes.
States:	The name of municipality A changes. Before: **A** → After: **A**
Representation:	(*A*, OG, *A*): *A* was called...
Forward rules:	not applicable
Backward rules:	not applicable
Example:	**01001 Alegría-Dulantzi**: In censuses 1842 to 1981 this municipality was called Alegría. => (01001, OG, 01001)
Comments:	The old and new names are recorded separately in additional columns.

Type of change:	**ANNOTATION**
Description:	Another type of nonterritorial change, or complementary information about changes in a municipality.
States:	Another change or other information about changes in municipality *A*. Before: **A** → After: **A**
Representation:	(*A*, OA*x*, *A*): where lower case letter *x* can mean (this list may be extended): - *a*: municipality is known to have changed province, but previous code is not available. - *b*: complementary information about the changes. - *c*: no population data in a given census. - *d*: also appears with a second alternative name. - *e*: change of capital entity ('cap' key in MAP 2008) - *f*: the capital entity name changes ('dca' key in MAP 2008)
Forward rules:	*not applicable*
Backward rules:	*not applicable*
Example:	**13012 Almadenejos**: In the 1842 Census it is mistakenly referred to as "Almagro". => (13012, OA*b*, 13012)
Comments:	The descriptive text must be recorded separately in an additional column.

5. References

Ben Rebah, M., C. Plumejeaud, R. Ysebaert, and D. Peeters. 2011. *Modeling Territorial Changes and Time Series Database Building Process: Empirical Approach and Applications*. ESPON Technical Report 2.1. ESPON 2013 Database. March 2011. Available at: https://www.espon.eu/sites/default/files/attachments/2.1_TR_time_series.pdf (accessed 5 September 2017).

Duque, I. 2016. "Improving Quality and Added Value of Geo-Statistical units." European Forum for Geography and Statistics (EFGS) Conference. Paris, November 15–17, 2016. Available at: http://www.efgs.info/wp-content/uploads/conferences/efgs/2016/S4-1_presentationV8_IgnacioDuque_EFGS2016.pdf (accessed 8 November 2017).

Flora, P., F. Kraus, R. Walton, D. Caramani, J. Marti-Henneberg, and J. Schweikart. 2015. *European Regions: The Territorial Structure of Europe since 1870*. Series: Societies of Europe. Gordonsville: Palgrave Macmillan.

García Fernández, P. 1985. *Población de los actuales términos municipales 1900–1981*. Madrid: Instituto Nacional de Estadística.

García Fernández, P. 1994. *Población de hecho de los municipios de España según la configuración geográfica del censo de 1991. Cifras homogéneas de los censos 1900-1991*. Madrid: Obra Social de la Confederación Española de Cajas de Ahorro.

Goerlich, F.J., M. Mas, J. Azagra, and P. Chorén. 2006. *La localización de la población española sobre el territorio. Un siglo de cambios: un estudio basado en series homogéneas (1900–2001)*. Bilbao: Fundación BBVA.

Goerlich, F.J., F. Ruiz, P. Chorén, and C. Albert. 2015. *Cambios en la estructura y localización de la población: Una visión de largo plazo (1842–2011)*. Bilbao: Fundación BBVA.

Gregory, I.N. 2005. "The Great Britain Historical GIS." *Historical Geography* 33: 132–134. Available at: https://ejournals.unm.edu/index.php/historicalgeography/article/view/2933 (accessed 8 November 2017).

Gregory, I.N. and P.S. Ell. 2006. "Error-Sensitive Historical GIS: Identifying Areal Interpolation Errors in Time-Series Data." *International Journal of Geographical Information Science* 20(2): 135–152. Doi: http://dx.doi.org/10.1080/13658810500399589.

Gregory, I.N. and P.S. Ell. 2007. *Historical GIS: Technologies, Methodologies and Scholarship*. Cambridge: Cambridge University Press.

INSPIRE. 2013. "D2.8.III.1 INSPIRE Data Specification on Statistical Units – Technical Guidelines" Available at: http://inspire.ec.europa.eu/id/document/tg/su (accessed 8 November 2013).

Instituto Nacional de Estadística. 1981a. *Relación de municipios desaparecidos desde principio de siglo*. Madrid: INE.

Instituto Nacional de Estadística. 1981b. *Relación de municipios y códigos al 31 de diciembre de 1980*. Madrid: INE.

Instituto Nacional de Estadística. 2005. Alteraciones de los municipios en los Censos de Población desde 1842. Database distributed by Instituto Nacional de Estadística. Available at: http://www.ine.es/intercensal/ (accessed 8 November 2013).

Ministerio de Administraciones Públicas (MAP). 2008. *Variaciones de los municipios de España desde 1842*. Madrid: Secretaría General Técnica, MAP. Available at: http://www.sefp.minhafp.gob.es/dam/es/web/publicaciones/centro_de_publicaciones_de_la_sgt/Monografias0/parrafo/011113/text_es_files/Variaciones-INTERNET.pdf (accessed 8 November 2017).

Marti-Henneberg, J. 2005. "Empirical Evidence of Regional Population Concentration in Europe, 1870–2000." *Population, Space and Place* 11: 269–281. Doi: http://dx.doi.org/10.1002/psp.37.

Milego, R. and M.J. Ramos. 2011. *Disaggregation of Socioeconomic Data into a Regular Grid: Results of the Methodology Testing Phase*. ESPON Technical Report 2.2. ESPON 2013 Database. March 2011. Available at: https://www.espon.eu/sites/default/files/attachments/M4D_FR_TechnicalReports.zip (accessed 8 November 2017).

Ruiz, F. and F.J. Goerlich. 2014. "Taxonomía y representación de los cambios en los municipios españoles." Working Paper No. WP-EC 2014-01. Valencia: Instituto Valenciano de Investigaciones Económicas. Doi: http://dx.medra.org/10.12842/WPEC_201401.

Sindoni, G., S. De Francisci, M. Paolucci, and L. Tininini. 2002. "Experiences in Developing a Spatio-Temporal Information System." *Research in Official Statistics* 1: 45–57. Available at: http://ec.europa.eu/eurostat/documents/3217494/5644045/KS-CS-02-001-EN.PDF/69d90c27-c788-45bc-87c9-ac3ffa09b06f?version=1.0 (accessed 5 September 2017).

Tir, J., P. Shafer, P. Diehl, and G. Goertz. 1998. "Territorial Changes, 1816–1996: Procedures and Data." *Conflict Management and Peace Science* 16(1)(Spring): 89–97. Doi: https://doi.org/10.1177/073889429801600105.

Received March 2015
Revised September 2017
Accepted November 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 107–120, http://dx.doi.org/10.1515/JOS-2018-0006

Calibration Weighting for Nonresponse with Proxy Frame Variables (So that Unit Nonresponse Can Be Not Missing at Random)

Phillip S. Kott[1] *and Dan Liao*[1]

When adjusting for unit nonresponse in a survey, it is common to assume that the response/nonresponse mechanism is a function of variables known either for the entire sample before unit response or at the aggregate level for the frame or population. Often, however, some of the variables governing the response/nonresponse mechanism can only be proxied by variables on the frame while they are measured (more) accurately on the survey itself. For example, an address-based sampling frame may contain area-level estimates for the median annual income and the fraction home ownership in a Census block group, while a household's annual income category and ownership status are reported on the survey itself for the housing units responding to the survey. A relatively new calibration-weighting technique allows a statistician to calibrate the sample using proxy variables while assuming the response/nonresponse mechanism is a function of the analogous survey variables. We will demonstrate how this can be done with data from the Residential Energy Consumption Survey National Pilot, a nationally representative web-and-mail survey of American households sponsored by the U.S. Energy Information Administration.

Key words: Model variable; calibration variable; weight-adjustment function; selection bias.

1. Introduction

Calibration weighting is a useful tool for treating unit nonresponse in a survey. It can implicitly estimate the probability of response given a known form of the response model. Moreover, the resulting weights tend to more efficient than the weights produced using maximum-likelihood methods to estimate the response model (Kim and Riddles 2012).

Deville (2000) has shown how calibration weighting can be used to treat unit (element-level) nonresponse that can be either missing at random (MAR) or not missing at random (NMAR). The former means that nonresponse is a function entirely of variables with either known population totals or known values for the entire sample, while the latter allows nonresponse to be at least partially a function of variables known only for responding sampled elements. The calibration-weighting framework in Särndal and Lundström (2005) also allows nonresponse to be not missing at random.

Unfortunately, there is no statistical way to determine whether or not nonrespondents are missing at random. Molenberghs et al. (2008) show that any data set fit by a model

[1] RTI International, 6110 Executive Blvd., Rockville, MD 20852, U.S.A. Emails: pkott@rti.org; dliao@rti.org.
Acknowledgments: Much of this work was supported by a grant from the National Science Foundation, award number SES-1424492.

assuming nonrespondents are not missing at random could also be fit by a model assuming nonrespondents are missing at random. As a result, many have argued that techniques like Deville's are best suited for sensitivity analyses. National Research Council (2010; 48, 59) discusses the limitations of what it calls the "inverse probability weighting" method for handling not-at-random missingness.

There are some situations, however, where unit nonresponse can logically be inferred to be not missing at random. In a survey of housing units (HUs), for example, unit nonresponse may be a function of whether or not the HU is owned by the household residing in it and by the annual income of that household. This information can be collected on the survey itself (assuming no item nonresponse), but can only be proxied for the sample as a whole. Such proxies are useful because Deville's method requires that there be variables on which to calibrate the respondent sample so that the weighted sum of those variables among respondents equal a known population total or a weighted total computed from the full sample (including nonrespondents). A potential source for proxy variables in the United States is the American Community Survey, which makes available estimates at the Census-block-group level of the average median annual income and the fraction of owned HUs.

Using data from the 2015 national pilot of the (United States) Residential Energy Consumption Survey (RECS) which was conducted by mail and web, we demonstrate how one can compare results of calibration weighting assuming nonresponse is missing at random using proxy variables available on the frame as response model variables with results of calibration weighting where survey variables, more logically related to response than their proxies, replace the proxy variables in the response model, showing in the process how to choose which survey variables to include in the response model.

Section 2 will review the underlying theory of calibration weighting assuming (for simplicity) a logistic response function. Section 3 will describe the RECS National Pilot and how it is being weighted to compensate for nonresponse assuming that unit respondents are missing at random. Section 4 compare some estimates and their estimated standard errors using the National-Pilot method and their alternatives that assume nonresponse is not missing at random. Section 5 offers some concluding remarks.

2. An Overview of Calibration Weighting Assuming a Logistic Response Function

To simplify matters, let us assume that there is only one type of unit nonresponse, and it takes place at the element level, denoted by the subscript k. Moreover, there is no coverage problems with the sampling frame nor is there any item nonresponse among element respondents.

In this article, we follow the quasi-randomization approach in Chang and Kott (2008) and treat unit response as an additional phase of probability sampling, where the response probabilities need to be estimated from the data. Although Kott and Chang (2010) showed that the methods they had proposed have good prediction-model properties, we will not discuss those here.

Suppose the unit (element) response mechanism can be represented by an independent logistic function that depends on a vector of values for each element. Letting ρ_k be the probability that element k responds, and $\mathbf{x}_k$ the vector of (response) *model* variables

governing that probability, which includes unity or the equivalent (i.e., a linear combination of the components of $\mathbf{x}_k$ is 1), we have

$$\rho_k = \rho\left(\mathbf{x}_k^T\boldsymbol{\gamma}\right) = 1/\left[1+\exp\left(\mathbf{x}_k^T\boldsymbol{\gamma}\right)\right], \tag{1}$$

for some unknown vector $\boldsymbol{\gamma}$.

Calibration weighting begins with the calibration equation:

$$\sum_R d_k\left[1+\exp\left(\mathbf{x}_k^T\mathbf{g}\right)\right]\mathbf{z}_k = \mathbf{T}_\mathbf{z} \tag{2}$$

where R denotes the respondent sample, d_k the design (initial sampling) weight of element k, $\mathbf{z}_k$ a vector of *calibration* variables, each having either a known population total or a total that can be estimated in the full sample (including the unit nonrespondents), $\mathbf{T}_\mathbf{z}$ the vector of (estimated) totals for the components of $\mathbf{z}_k$. Finally, $\mathbf{g}$ is a consistent estimator for $\boldsymbol{\gamma}$ under mild conditions, determined by solving for it in calibration equation (2) using Newton's method (repeated linearizations).

In practice, a $\mathbf{g}$ exists when one can be found satisfying (2). Moreover, that $\mathbf{g}$ is consistent for survey designs under which the expansion estimator for $\mathbf{T}_\mathbf{z}$ in the absense of nonresponse is consistent. The appendices in Chang and Kott (2008) lay out the theoretical conditions for $\mathbf{g}$ to exist and be consistent. A more lucid account of the theory underpinning this section can be found in Kott (2014).

The calibration weight for element k resulting from the solution of Equation (2) is

$$w_k = d_k\alpha\left(\mathbf{x}_k^T\mathbf{g}\right) = d_k\left[1+\exp\left(\mathbf{x}_k^T\mathbf{g}\right)\right].$$

The expression $\alpha\left(\mathbf{x}_k^T\mathbf{g}\right)$ is called the *weight-adjustment function* because it converts the design weight d_k into the nonresponse-adjusted or calibration weight w_k. The estimated total of a survey variable y using calibration weights is $t_y = \sum_R w_k y_k$.

In most applications, the components of calibration vector $\mathbf{z}_k$ are assumed to coincide with the components of the model vector $\mathbf{x}_k$. This means unit nonrespondents are assumed to be missing at random. When that is the case, the calibration equation (2) will almost always have a solution so long as unit nonresponse is truly a logistic function of the components of $\mathbf{x}_k$. When the components of $\mathbf{z}_k$ and $\mathbf{x}_k$ do not coincide, the calibration equation may not have a solution, especially if a component of $\mathbf{x}_k$ is linearly independent of all the components of $\mathbf{z}_k$.

Chang and Kott (2008) generalized the notion of calibration weighting to allow more calibration variables than model variables, but Kott and Liao (2017) maintained that a prudent approach would be to include in $\mathbf{z}_k$ all the components of $\mathbf{x}_k$ for which population totals or full-sample estimates are known. The rest they called *shadow variables*, which they suggested should be proxies for the *model-only variables* in $\mathbf{x}_k$ that could not themselves be calibration variables in $\mathbf{z}_k$.

Some variables in the RECS National Pilot sample, such as an indicator of whether (or not) an HU k is in an urban area, can be in both the model vector and the calibration vector, while other variables, such as home ownership (yes or no), are model-only variables in $\mathbf{x}_k$. At the same time, a reasonable proxy for each model-only variable, like the fraction of homes owned in its Census block group, can be a shadow variable in $\mathbf{z}_k$.

When the calibration equation has a solution, it is not hard to show that an asymptotically unbiased estimator for the variance of $\mathbf{g}$ under mild conditions is

$$\mathbf{V_g} = \mathbf{F}\ \mathbf{var}\left\{\sum_R d_k\left[1+\exp\left(\mathbf{x}_k^T\boldsymbol{\gamma}\right)\right]\mathbf{z}_k|\mathbf{T_z}\right\}\mathbf{F}^T, \tag{3}$$

where $\mathbf{F} = \left[\sum_R d_k \exp\left(\mathbf{x}_k^T\mathbf{g}\right)\right]\mathbf{z}_k\mathbf{x}_k^T]^{-1}$, and $\mathbf{var}\{\mathbf{q}|\mathbf{T_z}\}$ is an estimator of the variance-covariance matrix for $\mathbf{q}$ when $\mathbf{q}$ is viewed as an estimator for $\mathbf{T_z}$. To compute it, one treats $p_k = 1/\left[1+\exp\left(\mathbf{x}_k^T\mathbf{g}\right)\right]$ as if it equaled ρ_k in Equation (1).

An asymptotically unbiased estimator for the quasi-probability variance of $t_y = \sum_R w_k y_k$ (again under mild conditions) is

$$v_y = v\left\{\sum_S d_k\left[\mathbf{z}_k^T\mathbf{b} + \alpha\left(\mathbf{x}_k^T\mathbf{g}\right)e_k\right]\right\} \tag{4}$$

where $e_k = y_k - \mathbf{z}_k^T\mathbf{b}$, $\mathbf{b} = \left[\sum_R d_j\alpha'\left(\mathbf{x}_j^T\mathbf{g}\right)\mathbf{x}_j\mathbf{z}_j^T\right]^{-1}\sum_R d_j\alpha'\left(\mathbf{x}_j^T\mathbf{g}\right)\mathbf{x}_j y_j$, and $\alpha\left(\mathbf{x}_k^T\mathbf{g}\right) = \left[1+\exp\left(\mathbf{x}_k^T\mathbf{g}\right)\right]$ when $k \in R$ and 0 otherwise is treated as a constant within the probability-sampling variance estimator $v\{.\}$. For the variance of $m_y = \sum_R w_k y_k / \sum_R w_k$, replace y_k by $(y_k - m_y)/\sum_R w_j$.

It is easy to see that due to calibration $\sum_R w_k y_k - \sum_S d_k y_k = \sum_R w_k e_k$ (which also provides a heuristic justification for Equation (4)). We thus have the following estimate for the increase in quasi-probability variance due to nonresponse and nonresponse adjustment:

$$\begin{aligned}\mathrm{var}\left\{\sum_R d_k\left[1+\exp\left(\mathbf{x}_k^T\boldsymbol{\gamma}\right)\right]y_k|\sum_S d_k y_k\right\} &= \sum_R d_k^2\left(1/p_k^2\right)(1-p_k)e_k^2 \\ &= \sum_R d_k^2\left[1+\exp\left(\mathbf{x}_k^T\mathbf{g}\right)\right]\exp\left(\mathbf{x}_k^T\mathbf{g}\right)e_k^2, \end{aligned} \tag{5}$$

The estimate assumes the probabilities of element response are independent of each other. Again, the reader can consult Kott (2014) for proofs and details.

3. The RECS National Pilot

The RECS National Pilot was an attempt to convert what historically has been an in-person interview survey into one conducted by web and mail. More information on it can be found elsewhere (Berry and O'Brien 2016). For our purposes, the RECS National Pilot (hereafter the "National Pilot") used four randomly-assigned protocols and two randomly-assigned incentive levels in data collection from a stratified, two-stage sample of 9,650 dwelling units drawn using an address-based sampling frame with mail invitation and up to six mailings. The protocols were, 1, web only, 2, choice of web or mail, 3, choice of web or mail but with an added USD 10 incentive to respond via web, and, 4, web in the first mailing followed by a choice in subsequent mailings. The two incentive levels both provided the sampled HU USD 5 initially. One provided an extra USD 10 upon completion while the other provided an extra USD 20. There was a shortened mail follow-up survey (NRFU) for nonrespondents, but that does not concern us here — except in a

design-weight adjustment to be described shortly – nor does the poststratification designed to capture HUs not on the address-based sampling frame.

Two issues with the enumerations of the National Pilot do have an impact on our analysis. Not all HUs in the sampling frame were occupied, and some were occupied but not primary residents. Only data from primary residents were to be used in making National-Pilot estimates.

A latent-variable model (Biemer et al. 2016) has been used to estimate the probability that a sampled HU was occupied based on frame characteristics, the disposition of the first three mailings, and whether they responded to the survey. Those estimates have been incorporated into the design weights (the d_k in Equation (2)). Also incorporated into the design weights are the inverse of an estimated probability of a non-vacant HU being a primary residence. All responding primary residences had an estimated probability of 1, and all HU determined not to be primary residences a probability of 0. The rest have been assigned a probability of being a primary residence based on a logistic regression conducted among partially or fully responding HUs to either the National Pilot or its NRFU survey for which primary residence status could be determined.

Roughly 40% of eligible HUs responded to the RECS National Pilot. After investigating a longer list of candidate variables, the logistic model used to fit a response model in the National Pilot contains indicators for 17 geographic area (groups of states), indicators for the four protocols, indicators for the two incentive levels, an urbanicity indicator, an indicator of whether the HU is a single-family dwelling units from the frame, the fraction of HUs owned in the Census block group (CBG) containing the HU, and the fraction of HUs in its CBG with annual incomes less than USD 60,000. The latter two are estimated from the 2010 American Community Survey.

The WTADJUST procedure in SUDAAN® (Research Triangle Institute 2012) has been used to compute the calibration weights for the National Pilot. The procedure removes the extraneous calibration variables that would cause a singularity in matrix inversion (e.g., because the four protocol levels and two incentive levels cannot all define non-singular calibration variables).

WTADJUST has also been used to choose the variables for the National Pilot's missing-at-random logistic response model, which assumed the components of $\mathbf{x}_k$ in Equations (1) and (2) were the same as those in $\mathbf{z}_k$. WTADJUST fits a logistic model very much like SUDAAN's pseudo-maximum-likelihood logistic regression procedure (RLOGIST) but with a different estimating equation (WTADJUST solves for $\mathbf{g}$ in Equation (2) rather than in $\sum_R d_k \mathbf{z}_k = \sum_S \{d_k / [1 + \exp(-\mathbf{z}_k^T \mathbf{g})]\} \mathbf{z}_k$). The logistic functional form is, in fact, only a special case of the weight-adjustment functions fit by WTADJUST, but we restrict our attention to that form here until the concluding section.

4. Converting Proxy Variables into Model-Only Variables

The response model fit for the National Pilot contains three model variables that logic suggests would be more reasonably replaced by survey variables: the frame indicator for a single-family dwelling unit, the CBG fraction of owned HUs, and the CBG fraction of HUs with annual income less than USD 60,000.

Using the model variables described in the previous section as the calibration variables in fitting a missing-at-random (MAR) logistic response model, Table 1 shows the adjusted F values and their associated p-values produced by the WTADJUST (which uses Equation (3) to estimate variances by setting DESIGN = WR ADJUST = NONRESPONSE and NEST _ONE_). All the model variables are significant at the .15 level and have an F value greater than 2.5.

Table 2 show what happens when the three survey variables discussed above replace their proxy frame values in the model vector but not in the calibration vector. This is denoted as NMAR1 and fitted using WTADJX. Only annual income less than USD 60,000 remains significant at the .15 level, while the F values of the other two fall below 1. This is partly due to collinearity among them. In Table 3, NMAR2 removes whether the HU is a single-family dwelling unit from the model vector. All the remaining variables are significant at the .1 level. It should be noted that estimation treats mobile homes and attached single-family units as single-family dwelling units. Removing one of both does not meaningfully change the results however.

A fourth fit, NMAR3, containing the same model variables as NMAR2 with similar results is not shown. It replaces the two shadow calibration variables in NMAR2, the CBG fraction of owned HUs and the CBG fraction of HUs with annual incomes less than USD 60,000, with ordinary-least-squares (OLS) predictions of the probability of HU ownership and the probability of having an annual income less than USD 60,000, as suggested in Kott and Liao (2017). The regressors in those OLS predictions are the two CBG fractions and the frame indicator of the HU being a single-family dwelling unit.

Table 4 displays a number of estimated means and (quasi-probability) standard errors computed (with SUDAAN and NEST _ONE_ replaced by NEST STRATUM PSU to capture stratification and clustering effects on the estimated means) first assuming missingness is completely at random (MCAR; i.e., unit response does not depend on any frame or survey variables and both model and calibration vectors only have an intercept), then missing at random as in Table 1, and after that missing not at random under the NMAR assumption and using the three NMAR methods described above. All five methods treat the original sample as a stratified two-stage sample, with the original design's 19 strata collapsed into 17 variance strata to avoid variance strata containing only a single primary sampling unit (PSU). The PSUs in the RECS National Pilot design are 2010 US Census Public Use Microdata Areas (PUMAs, http://www.census.gov/geo/reference/puma.html).

Table 1. MAR: Model variable and calibration variables are the same.

Variable	Adjusted Wald F	p-value
GEOGRAPHICAL AREA	4.63	0.0000
INCENTIVE	17.63	0.0000
PROTOCOL	8.76	0.0000
URBANICITY INDICATOR	3.19	0.0741
CBG ANNUAL INCOME ≤ $60K?	8.44	0.0037
FRACTION OWNED IN CBG	2.52	0.1128
SINGLE-FAMILY UNIT(FRAME	6.95	0.0000

CBG – Census Block Group.

Table 2. NMAR1: three model-only variables and three shadow proxies.

Variable	Adjusted Wald F	*p*-value
GEOGRAPHICAL AREA	4.51	0.0000
INCENTIVE	14.43	0.0001
PROTOCOL	7.37	0.0001
URBANICITY INDICATOR	2.71	0.0996
ANNUAL INCOME ≤ $60K?	3.30	0.0695
HU OWNED	0.28	0.5938
SINGLE-FAMILY UNIT(SURVEY	0.00	0.9548

HU – Housing Unit.

The adjustments for the vacancies and non-primary residences are treated in variance estimation here as part of the design weights. Although this is a simplification, it is the same simplification for all five nonresponse-adjustment methods.

The results in Table 4 are summarized in Tables 5 and 6 and extended to three domain estimates: one for owned HUs (a model-only variable in the NMAR models), one for detached standing HUs, and one for HUs built before 1970. The measure log(X) − log(Y) = Log(X/Y) used in those tables is close to the percent difference between X and Y when that difference is less than 40% (Log(X/Y) ≈ (X − Y)/Y). Unlike percent differences, however, it is a symmetric measure (i.e., Log(X/Y) = − Log(Y/X)).

In Table 5, we see that the estimates from using the three NMAR methods always fall within 0.5% of each other. Assuming that these models more reasonably reflect reality than the MAR model, which in turn is more reasonable than the MCAR model, it appears that adjusting for nonresponse using an MAR model removes more than half of the bias relative to not adjusting at all (i.e., assuming unit nonresponse is completely at random). The sizes of the relative biases vary, with those associated with the two model-only variables (the fractions of HU owned and with annual income less than USD 60K) being the largest. Observe that the relative biases tend to be smaller for a domain related, or correlated to, the model-only variables (e.g., having a detached HU is correlated with both ownership and HU annual income).

In Table 6, we see that the estimated standard errors are, on average, lowest when the MAR is used, except for the domain of owned HUs. Using NMAR1 has, on average, the highest estimated standard errors while using NMAR3 has, on average, the lowest among the three NMAR methods but still higher estimated standard errors than when the MAR is used. The results appear to vary by variable, however.

Table 3. NMAR2: NMAR1 with an insignificant model-only variable removed.

Variable	Adjusted Wald F	*p*-value
GEOGRAPHICAL AREA	4.53	0.0000
INCENTIVE	14.89	0.0001
PROTOCOL	7.98	0.0000
URBANICITY INDICATOR	2.89	0.0894
ANNUAL INCOME ≤ 60K?	5.60	0.0179
HU OWNED	4.73	0.0297

HU – Housing Unit.

Table 4. Comparing alternative models of nonresponse adjustment.

Housing unit variable	Models				
	MCAR	MAR	NMAR1	NMAR2	NMAR3
Estimated means					
Fraction Built Before 1970	0.3802	0.3864	0.3946	0.3947	0.3947
Fraction Built After 1999	0.2227	0.2194	0.2124	0.2124	0.2124
Days Someone is at Home	3.5096	3.4940	3.5209	3.5203	3.5198
Fraction with an Attic	0.4953	0.4726	0.4607	0.4609	0.4607
Number of Bedrooms	2.9096	2.8412	2.8078	2.8084	2.8079
Fraction with a Cellar	0.3504	0.3187	0.3153	0.3155	0.3154
Fraction with Central Air	0.6871	0.6750	0.6654	0.6652	0.6651
Fraction with Clothes Washer	0.8706	0.8499	0.8429	0.8431	0.8429
Number of Desktops	0.5518	0.5404	0.5293	0.5292	0.5291
Fraction with Dishwasher	0.7417	0.7305	0.7119	0.7117	0.7116
Fraction with Dryer	0.8569	0.8344	0.8262	0.8263	0.8261
Fraction with Home Heating	0.9585	0.9538	0.9530	0.9530	0.9530
Fraction with Internet	0.8745	0.8725	0.8616	0.8617	0.8617
Fraction Owned	0.7134	0.6803	0.6468	0.6451	0.6445
Fraction with Annual Income $<$ \$60k	0.5356	0.5516	0.6181	0.6176	0.6173
Humber of Household Members	2.5583	2.5372	2.5236	2.5255	2.5257
Number of Refrigerators	1.4204	1.3941	1.3774	1.3774	1.3773
Number of Laptops	1.0701	1.0676	1.0327	1.0333	1.0335
Number of Tablets	0.9964	0.9783	0.9418	0.9425	0.9426
Fraction with a Garage	0.4653	0.4391	0.4270	0.4272	0.4270
Number of Rooms	6.5633	6.4211	6.3422	6.3429	6.3420
Number of Colors TVs	2.3651	2.3290	2.3009	2.3015	2.3013
Fraction of Detached Standing Units	0.6677	0.6330	0.6236	0.6239	0.6236
Fraction with a Central Furnace	0.6227	0.6081	0.6066	0.6067	0.6067
Fraction with Natural Gas Heating	0.4884	0.4764	0.4721	0.4721	0.4720
Fraction with Electric Heating	0.3541	0.3655	0.3691	0.3692	0.3693

Table 4. Continued.

Housing unit variable	Models				
	MCAR	MAR	NMAR1	NMAR2	NMAR3
Estimated standard errors					
Fraction Built Before 1970	0.0148	0.0142	0.0135	0.0142	0.0140
Fraction Built After 1999	0.0113	0.0113	0.0115	0.0114	0.0112
Days Someone is at Home	0.0384	0.0371	0.0418	0.0387	0.0388
Fraction with an Attic	0.0119	0.0093	0.0086	0.0093	0.0086
Number of Bedrooms	0.0331	0.0263	0.0257	0.0259	0.0251
Fraction with a Cellar	0.0147	0.0130	0.0129	0.0136	0.0129
Fraction with Central Air	0.0136	0.0122	0.0122	0.0119	0.0122
Fraction with Clothes Washer	0.0088	0.0071	0.0068	0.0072	0.0069
Number of Desktops	0.0133	0.0124	0.0125	0.0118	0.0116
Fraction with Dishwasher	0.0129	0.0112	0.0114	0.0103	0.0103
Fraction with Dryer	0.0094	0.0075	0.0071	0.0074	0.0073
Fraction with Home Heating	0.0061	0.0065	0.0067	0.0068	0.0067
Fraction with Internet	0.0070	0.0065	0.0071	0.0068	0.0067
Fraction Owned	0.0107	0.0073	0.0535	0.0201	0.0127
Fraction with Annual Income $<$ \$60k	0.0148	0.0120	0.0351	0.0322	0.0321
Humber of Household Members	0.0283	0.0269	0.0588	0.0349	0.0342
Number of Refrigerators	0.0152	0.0125	0.0114	0.0114	0.0115
Number of Laptops	0.0260	0.0246	0.0288	0.0249	0.0249
Number of Tablets	0.0237	0.0210	0.0290	0.0239	0.0238
Fraction with a Garage	0.0164	0.0125	0.0111	0.0108	0.0113
Number of Rooms	0.0683	0.0537	0.0493	0.0511	0.0495
Number of Colors TVs	0.0315	0.0283	0.0326	0.0294	0.0297
Fraction of Detached Standing Units	0.0132	0.0092	0.0091	0.0110	0.0090
Fraction with a Central Furnace	0.0111	0.0102	0.0104	0.0108	0.0105
Fraction with Natural Gas Heating	0.0151	0.0135	0.0136	0.0137	0.0134
Fraction with Electric Heating	0.0135	0.0129	0.0134	0.0131	0.0128

NMAR3 – ols fits of the two model-only variables using the three shadow variables.

Table 5. Summarizing the relative percent differences of the estimated means using alternative models of nonresponse adjustment across 26 variables.

	Mean	Median	3rd Q	Max
All				
$\lvert\log(\text{MAR}) - \log(\text{MCAR})\rvert \times 100$	2.56	2.15	2.96	9.48
$\lvert\log(\text{NMAR1}) - \log(\text{MAR})\rvert \times 100$	2.09	1.24	2.58	11.38
$\lvert\log(\text{NMAR2}) - \log(\text{NMAR1})\rvert \times 100$	0.04	0.02	0.04	0.26
$\lvert\log(\text{NMAR3}) - \log(\text{NMAR2})\rvert \times 100$	0.02	0.02	0.03	0.10
Owned housing unit				
$\lvert\log(\text{MAR}) - \log(\text{MCAR})\rvert \times 100$	0.86	0.66	0.95	5.58
$\lvert\log(\text{NMAR1}) - \log(\text{MAR})\rvert \times 100$	1.42	0.73	1.17	12.80
$\lvert\log(\text{NMAR2}) - \log(\text{NMAR1})\rvert \times 100$	0.05	0.03	0.06	0.22
$\lvert\log(\text{NMAR3}) - \log(\text{NMAR2})\rvert \times 100$	0.02	0.01	0.02	0.11
Detached housing unit (*excludes mobile homes*)				
$\lvert\log(\text{MAR}) - \log(\text{MCAR})\rvert \times 100$	0.72	0.47	0.80	5.13
$\lvert\log(\text{NMAR1}) - \log(\text{MAR})\rvert \times 100$	1.76	1.09	2.11	14.09
$\lvert\log(\text{NMAR2}) - \log(\text{NMAR1})\rvert \times 100$	0.04	0.02	0.04	0.24
$\lvert\log(\text{NMAR3}) - \log(\text{NMAR2})\rvert \times 100$	0.01	0.00	0.01	0.08
Built before 1970				
$\lvert\log(\text{MAR}) - \log(\text{MCAR})\rvert \times 100$	2.61	2.25	3.27	8.21
$\lvert\log(\text{NMAR1}) - \log(\text{MAR})\rvert \times 100$	1.77	0.97	1.50	9.68
$\lvert\log(\text{NMAR2}) - \log(\text{NMAR1})\rvert \times 100$	0.05	0.03	0.06	0.32
$\lvert\log(\text{NMAR3}) - \log(\text{NMAR2})\rvert \times 100$	0.03	0.02	0.04	0.11

Table 7 tries to get a cleaner picture of the impact of unit nonresponse and the alternative methods of adjusting for it when estimating means for all occupied residences. It computes the square root of the estimated added variance due to nonresponse adjustment computed using Equation (5) (which could not be done in SUDAAN). This measure ignores the impact of any correlation, whether real or random, between the sampling and nonresponse errors. Similarly, it ignores the impact of any within PSU correlations across HUs. The conclusions from Table 6 are amplified. The additional estimated variance from using MAR is always less than that from using any NMAR method. The additional estimated variance from using MAR is also less than that from doing nothing (i.e., MCAR) at least 75% of the time (since the third quartile is negative), contrary to popular belief, a possibility pointed out by Little and Vartivarian (2005). Similarly, the added estimated variances drop in over 75% of the cases (in fact, all but one case) when NMAR2 replaces NMAR1 and NMAR3 replaces NMAR2.

5. Concluding Remarks

The primary purpose of this article was to show how the theoretical and simulation results from Kott and Liao (2017) could be applied to a real survey suffering from a relatively large fraction of unit nonresponse (roughly 60%). In creating calibration weights to compensate for units nonresponse to RECS National Pilot survey, element response was at first modeled as a function of variables with known values for the entire sample, where

Table 6. Summarizing the relative differences of the estimated standard errors using alternative models of nonresponse adjustment across 26 variables.

	Mean	Min	1st Q	Median	3rd Q	Max
All						
(log(MAR) − log(MCAR)) × 100	−14.19	−38.80	−22.27	−11.78	−5.69	6.67
(log(NMAR1) − log(MAR)) × 100	16.40	−11.16	−3.36	1.17	11.87	199.48
(log(NMAR2) − log(NMAR1)) × 100	−6.96	−97.94	−8.69	−1.26	3.61	18.75
(log(NMAR3) − log(NMAR2)) × 100	−3.74	−45.83	−2.99	−1.85	0.11	4.51
Owned housing unit						
(log(MAR) − log(MCAR)) × 100	1.94	−5.51	−0.40	0.58	3.36	13.36
(log(NMAR1) − log(MAR)) × 100	17.51	−0.44	4.99	17.58	21.83	99.71
(log(NMAR2) − log(NMAR1))) × 100	−10.48	−25.02	−16.48	−9.74	−1.99	−0.33
(log(NMAR3) − log(NMAR2)) × 100	−0.56	−3.62	−1.10	−0.47	0.01	1.98
Detached housing unit (*excludes mobile homes*)						
(log(MAR) − log(MCAR)) × 100	−3.45	−13.84	−7.26	−3.16	−0.09	6.57
(log(NMAR1) − log(MAR)) × 100	21.32	−7.56	−0.13	3.60	22.63	194.02
(log(NMAR2) − log(NMAR1)) × 100	−10.32	−117.46	−12.08	−2.92	0.39	6.40
(log(NMAR3) − log(NMAR2)) × 100	−1.81	−17.75	−2.40	−0.64	−0.06	1.40
Built before 1970						
(log(MAR) − log(MCAR)) × 100	−5.08	−15.61	−8.86	−4.81	−1.73	10.44
(log(NMAR1) − log(MAR)) × 100	12.72	−5.11	−2.83	2.27	8.68	154.28
(log(NMAR2) − log(NMAR1)) × 100)	−6.93	−92.88	−7.75	−1.32	2.23	6.93
(log(NMAR3) − log(NMAR2)) × 100	−2.69	−34.99	−3.17	−0.37	0.17	1.75

Table 7. Summarizing the relative differences of the increases of estimated standard errors using alternative models of nonresponse adjustment across 26 variables.*

	Mean	Min	1st Q	Median	3rd Q	Max
$(\log(\text{MAR}) - \log(\text{MCAR}^{**})) \times 100$	−19.02	−78.53	−36.08	−12.52	−4.19	6.50
$(\log(\text{NMAR1}) - \log(\text{MAR}))) \times 100$	66.57	8.71	19.03	30.45	60.70	432.61
$(\log(\text{NMAR2}) - \log(\text{NMAR1})) \times 100$	−10.96	−205.68	−28.86	−1.97	23.38	76.69
$(\log(\text{NMAR3}) - \log(\text{NMAR2})) \times 100$	−16.95	−88.08	−27.98	5.23	−2.27	0.04
$(\log(\text{NMAR3}) - \log(\text{NMAR1})) \times 100$	−27.90	−285.54	−34.27	−6.42	−2.59	0.85
$(\log(\text{NMAR3}) - \log(\text{MAR})) \times 100$	38.67	2.58	10.03	19.72	45.93	261.25

*The increases are computed as the square root of the right-hand side of Equation (5).
**MCAR treats the intercept as the lone model and calibration variable.

some were of those obvious proxies for variables with known values only for respondents. When those proxies were replaced by their model-only analogues in a calibration-weighting equation, one was found no longer to be a contributor of response. Still, following Kott and Liao (2017), this type of variables was shown to have value in creating shadow variables for model-only values using OLS. As Kott and Liao demonstrated, the resulting calibration-weighted estimator retains its near quasi-probability-sampling unbiasedness despite the somewhat ad-hoc use of OLS.

With this data, there appeared to be gains in bias reduction from assuming reasonably that nonresponse was a logistic function of survey variables rather than their frame proxies (which were several years old when based on the ACS). With the largest bias reductions in those survey variables added to the response model. There was, however, a marked tendency for the standard errors to increase when NMAR modeling replaced MAR modeling. The reader should keep in mind that the results from the RECS National Pilot may not generalize to other surveys.

It is a simple matter to extend the methodology used here to other element response functions. In SUDAAN, the weight adjustment function in Equation (2) can be replaced by:

$$\alpha\left(\mathbf{x}_k^T\mathbf{g}\right) = \left[L + \exp\left(\mathbf{x}_k^T\mathbf{g}\right)\right] / \left[1 + U^{-1}\exp\left(\mathbf{x}_k^T\mathbf{g}\right)\right],$$

the inverse of which is a truncated logistic response model where the probabilities of element response are bound between $1/U \geq 0$ and $1/L \leq 1$. Other smooth monotonic functions can also be used $\alpha(.)$, but the user may have to do his/her own programming for that. Choosing an appropriate form for the response function and the penalty for failing to do so is an area for future research.

Finally, the reader should be aware that there are packages in R that can implement calibration weighting similar to the routine in SUDAAN. One such is 'Sampling' (Tille and Matei 2013).

6. References

Berry, C. and E. O'Brien. 2016. "Managing the Fast-Track Transformation of a 35-Year Old Federal Survey." Presented at the 2016 FedCASIC Workshop, Washington DC, May 4, 2016. Available at: https://www.census.gov/fedcasic/fc2016/ppt/2_2_Speed.pdf (accessed September 2016).

Biemer, P., P. Kott, and J. Murphy. 2016. "Estimating Mail or Web Survey Eligibility for Undeliverable Addresses: A Latent Class Analysis Approach." Proceedings of the Survey Research Methods Section: American Statistical Association, Chicago, IL, August 2016, 1166–1172. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2016/files/389587.pdf.

Chang, T. and P.S. Kott. 2008. "Using Calibration Weighting to Adjust for Nonresponse under a Plausible Model." *Biometrika* 95: 557–571. Doi: http://dx.doi.org/10.1093/biomet/asn022 (A version for the full appendices can be found here: http://ageconsearch.umn.edu/handle/234362).

Deville, J.C. 2000. "Generalized Calibration and Application to Weighting for Non-response." COMPSTAT: Proceedings in Computational Statistics, 14th Symposium,

Utrecht, The Netherlands, edited by J.G. Bethlehem and P.G.M. van der Heijden. New York: Springer-Verlag. Doi: http://dx.doi.org/10.1007/978-3-642-57678-2.

Kim, J.K. and M. Riddles. 2012. "Some Theory for Propensity Scoring Adjustment Estimator." *Survey Methodology* 38: 157–165.

Kott, P. 2014. "Calibration Weighting When Model and Calibration Variables Can Differ." In *Contributions to Survey Statistics* - ITACOSM 2013 *Selected Papers* (pp. 1–18). Cham: Springer, Contributions to Statistics. Doi: https://doi.org/10.1007/978-3-319-05320-2_1.

Kott, P. and T. Chang. 2010. "Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse." *Journal of the American Statistical Association* 105: 1265–1275. Doi: http://dx.doi.org/10.1198/jasa.2010.tm09016.

Kott, P. and D. Liao. 2017. "Calibration Weighting for Nonresponse that is Not Missing at Random: Allowing for More Calibration than Response-model Variables." *Journal of Survey Statistics and Methodology* 5(2): 159–174. Doi: https://doi.org/10.1093/jssam/smx003.

Little, R. and S. Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31: 161–168.

Molenberghs, G., C. Beunckens, and C. Sotto. 2008. "Every Missingness Not at Random Model has a Missingness at Random Counterpart with Equal Fit." *Journal of Royal Statistical Society B* 70: 371–388. Doi: http://dx.doi.org/10.1111/j.1467-9868.2007.00640.x.

National Research Council. 2010. *The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials*. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. Doi: http://dx.doi.org/10.17226/12955.

Research Triangle Institute. 2012. *SUDAAN Language Manual*, Volumes 1 and 2, Release 11. Research Triangle Park, NC: Research Triangle Institute.

Särndal, C-E. and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. John Wiley & Sons: Chichester. Doi: http://dx.doi.org/10.1002/0470011351.

Tille, Y. and A. Matei. 2013. Package 'Sampling.' A software routine available at" http://cran.r-project.org/web/packages/sampling/sampling.pdf (accessed October 2017).

Received September 2016
Revised November 2017
Accepted November 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 121–148, http://dx.doi.org/10.1515/JOS-2018-0007

Optimal Stratification and Allocation for the June Agricultural Survey

Jonathan Lisic[1], *Hejian Sang*[2], *Zhengyuan Zhu*[2], *and Stephanie Zimmer*[2]

A computational approach to optimal multivariate designs with respect to stratification and allocation is investigated under the assumptions of fixed total allocation, known number of strata, and the availability of administrative data correlated with thevariables of interest under coefficient-of-variation constraints. This approach uses a penalized objective function that is optimized by simulated annealing through exchanging sampling units and sample allocations among strata. Computational speed is improved through the use of a computationally efficient machine learning method such as K-means to create an initial stratification close to the optimal stratification. The numeric stability of the algorithm has been investigated and parallel processing has been employed where appropriate. Results are presented for both simulated data and USDA's June Agricultural Survey. An R package has also been made available for evaluation.

Key words: Area survey; optimal allocation; optimal stratification; multivariate design; simulated annealing.

1. Introduction

An attribute of many federal surveys is the use of a stratified design. Stratified designs allow the inclusion of knowledge about the Primary Sampling Units (PSU)s in a population through administrative variables that are well correlated with the desired estimators. Many federal surveys have the additional requirements of agency-mandated quality constraints. These constraints are typically based on the Coefficient-of-Variation (CV) or other functions of variance placed on either administrative variables or the survey estimates. Besides quality constraints, financial constraints are also imposed on federal surveys. This brings forth the question, "How can a federal survey practitioner optimally stratify and allocate a survey to meet imposed quality constraints without spending any more money?" In this article, a solution to this question is presented for the case of CV quality constraints and a fixed sample size.

An important aspect of this question is the concept of an optimal stratified design, where an optimal stratified design implies the joint optimization of stratification and allocation with respect to a predeterminated objective function. If an optimal design can be found, then it is optimal over all possible pairings of stratifications and allocations under the design constraints. Joint optimization differs from optimal stratification with an *a priori*

[1] Cigna, 900 Cottage Grove Rd, Bloomfield, CT 06002, U.S.A. Email: jonathan.lisic@cigna.com
[2] Iowa State University, Osborn Dr, Ames, IA 50011 U.S.A. Emails: hjsang@iastate.edu, zhuz@iastate.edu, and sazimme2@iastate.edu.

allocation or optimal allocation conditioned on a prior fixed stratification. Here a priori allocation is defined as designs where the allocations are predeterminated functions of the strata population sizes, such as proportional, uniform or other allocation methods that do not admit administrative data. Both the a priori and conditional allocation place additional constraints on the objective function. In the case of conditional allocation, the objective function used for stratification lacks information about the optimal allocation, making it necessary to optimize an alternative objective function. If an optimal allocation is performed with the desired objective function, then the allocation is restricted by the prior stratification. This restriction can lead to a nonoptimal design. In the case of a priori allocation, only a subset of pairings of all possible stratifications and allocations are considered; this subset is unlikely to contain the optimal allocation for a given objective function due to ignoring administrative data. The importance and improvements provided by assuming neither a priori allocation nor using conditional allocation are discussed and displayed through empirical results in Benedetti et al. (2008), Day (2009), Baillargeon and Rivest (2009), and Ballin and Barcaroli (2013). A comparison of a priori allocated designs for multivariate surveys can be found in Kozak (2006b); further discussion can be found in Gonzalez and Eltinge (2010).

One major advantage that a priori and conditional allocation designs have over optimal stratified designs is that they are easy to obtain. Optimal stratified designs require an exploration of a combinatorial space to find an optimal design. This is a nontrivial problem for even small population and sample sizes. A solution to the problem of finding a univariate optimal stratified design subject to a CV constraint using Neyman allocation for a fixed sample size was proposed by Dalenius and Hodges (1959). This method is commonly known as the cum $\sqrt{f}$ method (Särndal et al. 1991, Section 3.7) (Horgan 2006). Lavallée and Hidiroglou (1988) and the multivariate extensions in Benedetti et al. (2010) and Benedetti and Piersimoni (2012) provide optimal designs under CV constraints, but restrict the strata to either two or three stratum. These stratum in Benedetti and Piersimoni (2012) include a census (take-all) and sampled (take-some) strata and do not restrict the sample size. Benedetti et al. (2010) included a third (take-none) stratum for the purposes of cut-off sampling. These approaches are designed for highly skewed populations, exploiting the similarity of the underlying population to a geometric progression (Gunning et al. 2004). Benedetti and Piersimoni (2012) introduced a method for stratification which uses multiple administrative variables. This method, which is motivated by the Lavallée and Hidiroglou method, partitions the population into two strata, one which is sampled and one, which is a take-all stratum. The partitioning is determined such that the sample size is minimized for a target coefficient of variation of a response variable. In addition to allocations with goals of increasing precision, allocations also consider data collection costs and other practical constraints such as the method proposed by Valliant et al. (2014) to allocate sample in household surveys using Address-Based Sampling Frames and available commercial data.

Other multivariate approaches to optimal stratified designs can be found in Ballin and Barcaroli (2013) and Benedetti et al. (2008). Both of these methods are designed to work on a set of categorical administrative variables. These administrative variables work as a means of data reduction by assigning each PSU to an initial stratum, called an atomic stratum, defined by a unique combination of administrative values. To admit optimization

under variance or CV constraints, each atom is assigned a variance estimate. Simultaneous allocation in both cases is performed through optimal allocation as defined in Bethel (1986), where strata allocations are round up to the nearest integer. The major difference between these algorithms is how they explore the combinatorial space of the atomic strata. A divisive tree-based approach is used in Benedetti et al. (2008). In this approach, at each layer of the tree, a stratum is split by the administrative variable that results in the greatest reduction of total sample size according to optimal allocation in Bethel (1986). This is continued until a set of CV quality constraints are met. In Ballin and Barcaroli (2013), a Genetic Algorithm (GA) is used to explore the space of strata formed by merging atomic strata. At each iteration (generation) of the GA, a large collection of possible stratifications are generated and evaluated through minimum sample size under CV constraints. A set of sufficiently well-performing stratifications and allocations and a small number of less optimal stratifications and allocations are retained to contribute to future generations. The less optimal stratifications and allocations are retained to provide genetic diversity. These stratifications and allocations, along with combinations and mutations of these stratifications are then used as the next generation. Combinations of strata are formed by exchanging atomic strata assignments between two stratifications, and mutations are formed by randomly assigning an existing atomic stratum to another stratum. Iteration is continued until changes in the objective function plateau. Both of these approaches consider a variable number of strata with means to specify a maximum number of strata to avoid an unstable stratified design.

A separate but important issue not directly addressed by the previously mentioned works involves the relationship between the administrative variables being optimized and the desired estimators. Even when the administrative variables and the desired estimators are highly correlated, the optimal stratification and allocation under the administrative variables may not be optimal for the desired estimators. In particular, an assumption that meeting quality constraints for the administrative variables may not imply meeting assumed quality constraints for the desired estimators. A discussion of this issue and proposed solution for univariate stratified designs using anticipated moments can be found in Baillargeon and Rivest (2009). Anticipated moments are moments of a random variable calculated under the sample design and the super population model (Isaki and Fuller 1982). When the super population model, referred to as the model in this article, is correctly specified or a sufficiently robust model is used, it is possible to construct strata that on average meet the quality constraints for the desired estimators.

The prior literature on the subject of optimal stratified designs does not address three important use cases; (1) multivariate optimal allocation with continuous administrative variables with more than two strata, (2) multivariate optimal allocation using anticipated moments to attain CV constraints for desired estimators, (3) application to fixed sample sizes with CV constraints. In this article, a method to construct optimal multivariate stratified designs for an arbitrary, but fixed, number of self-representing strata from continuous valued administrative data is presented. This method admits a combination of *hard* and *soft* constraints, where soft constraints are handled by a penalized objective function and hard constraints are handled through traditional nonlinear programming constraints. Anticipated moments can be used within the objective function to account for the relationship between administrative variables and desired response. Accounting for

this relationship allows for optimization with CV constraints on the desired estimators. Optimization of this objective function is performed by simulated annealing, by moving individual PSUs between strata. The use of soft constraints allows a survey practitioner to find potential solutions by relaxing less important constraints. Unlike prior multivariate methods, this choice of simulated annealing for optimization provides the theoretical result of guaranteed convergence to the global optima, and good performance characteristics. Simultaneous stratification and allocation is provided by also considering changes in the allocation as part of the simulated annealing algorithm.

The problem of targeting multiple responses that are not-necessarily correlated with each other is a characteristic of area surveys in agriculture. Agricultural production of crops such as corn excludes using the land for another purpose such as growing soybeans. The agricultural area survey examined in this article is the United States Department of Agriculture (USDA) National Agricultural Statistics Service's (NASS) June Agricultural Survey (JAS). In particular, a proposed redesign of JAS using a permanent and fixed area frame is examined. Prior work on optimizing the existing JAS design using simulated annealing procedures has been proposed by Gentle and Perry (2000). This work focused on creating strata that are homogeneous with respect to remote sensing imagery. Given the quality of remotely sensed imagery at the time of publication, this approach provided remarkable improvements in efficiency. However, the approach did not consider optimal allocation, CV constraints, sampling unit dependent costs, or agricultural practices such as crop rotations. The application of the proposed method does consider all four topics and can be consider a modern revisit of the topic with the benefit of higher quality remote sensing data and faster computing resources.

This article is broken into the following sections. In Section 2, details on the proposed method, including the algorithm and objective function, are presented. In Section 3, simulated data are used to illustrate the proposed method and the result is compared to those from other stratification and allocation methods. In Section 4, the JAS is introduced and the results of applying the proposed method to the JAS are compared to those from the current univariate allocation method. The article concludes with a discussion and future extensions to the proposed method that would account for measurement error and improve computational efficiency.

2. Optimal Stratified Design Algorithm

The method proposed here uses a sequence of exchanges of PSUs between a set of initial strata to improve an objective function. The objective function is a weighted vector norm applied to the vector of administrative variable or modeled CVs attained by the current stratification and allocation. A penalty function is added to this objective function, where the penalty function is the sum of the element-wise products of penalty weights and penalty values. The penalty weights serve as importance weights, as in Kozak (2006b) with the exception that the weights can be set to zero once the constraint is met. The penalty values are the difference between the attained CVs and the target CVs for each administrative or modeled variable. This approach can be considered a weighted or approximate constraint satisfaction problem in operations research (Freuder and Wallace 1992). Traditional hard constraints can be considered by setting the penalty weight to

infinity. By allowing a combination of hard and soft constraints, survey managers and stakeholders have flexibility in identifying essential CV constraints and nonessential, but desirable, CV constraints. By defining these constraints separately, infeasibility of the constrained design by fixed sample sizes can potentially be avoided. If the nonessential constraints are violated, a design that minimizes the departure from the nonessential CV constraints can be found, and the constraints can be prioritized through the choice of penalty values.

The objective function optimizes the stratification and allocation through functions of moments; in particular, the population total and variance. The population total and variance are calculated from values assigned to individual PSUs. Therefore, either administrative data highly correlated with the desired response or a modeled response is required to find optimal stratification and allocation for a set of desired estimators. In the case of administrative data, it is assumed that the data is complete and available for each PSU. In the case of modeled response, it is similarly assumed that a model can be constructed for each PSU; variances based on this model can be incorporated within the objective function through anticipated moments. The later case will be discussed in Section 3.

The strata formed by PSU exchanges are self-representing. Self-representing strata are defined by PSU assignments, as opposed to any bounds on the administrative variables. Since the self-representing strata are not defined by a set of hyper-planes from administrative data bounds, they allow for strata that have nonlinear partitions and possibly disjoint subsets of the space of the administrative data.

Self-representing strata are also found in Benedetti et al. (2010) and Benedetti and Piersimoni (2012), however the approach presented can exceed three strata and is not restricted to highly asymmetric populations. Instead, it relies on the observation that, given an initial allocation and stratification, then a sequence of exchanges can be taken along primarily stratum boundaries to attain a more optimal design. Optimal allocation is performed simultaneously by potentially changing the sample size at each iteration. Optimization of this problem is performed by a stochastic optimization method known as simulated annealing (Metropolis et al. 1953). This optimization method is also used in Benedetti et al. (2010) and Benedetti and Piersimoni (2012). This stochastic optimization method is a metaheuristic that uses a Monte Carlo method to obtain an optimal solution by generating a sequence of possible solutions that slowly converge to an optimal solution. Simulated annealing has the useful property of being able to explore nonoptimal PSU exchanges and sample size changes, allowing for a more exhaustive search of the feasible region than deterministic optimization methods. This property allows for simulated annealing to guarantee convergence of the sequence to an optimal solution given sufficient run-time and precision.

Computational speed of the algorithm can be accelerated by starting with an initial stratification and allocation close to the optimal stratification and allocation. Such a stratification can be obtained through machine-learning methods such as K-means, and optimal allocation can be provided though the popular multivariate optimal allocation approach of Bethel (1986) with minor adjustments to ensure integer allocation and sample size constraints. It should be noted that, given hard constraints, the initial starting point must be in the feasible region; otherwise, single PSU exchanges are unlikely to result in a

finite objective function. Furthermore, the proposed method can also be used for optimal allocation, by running with only allocation changes. Additional hard and soft constraints may also be added to the objective function. Such constraints include costs for PSU collection, bounds on maximum stratum size, or spatial penalty functions.

An implementation of the proposed method has been written in the R Language (R Core Team 2015), as an R package (Lisic 2016). This package supports second order moments through per-PSU additive components and scaling. These second-order moment adjustments allow for the adoption of linear or linearized relationships between administrative variables and survey response. Furthermore, the additive per-PSU component can be used to impose fixed per-PSU costs.

In this section the objective function is presented using administrative data with weighting. Anticipated moments are introduced, followed by a discussion of the simulated annealing algorithm.

2.1. *Objective Function*

Multivariate objective functions are vector valued functions that map from $\mathbb{R}^J \rightarrow \mathbb{R}$, where J is the length of the vector input. If the objective function is bounded over the domain of the problem, then both a maximum and a minimum exist. Constrained objective functions carve out a subset of the unbounded region known as the feasible region. In methods to find optimal designs, the goal is to either minimize or to maximize an objective function over this feasible region. Since all maximization problems can be written as minimization problems by multiplying the objective function by negative one, only minimization problems will be considered in this article.

For constrained objective functions, the feasible region may be empty, implying a solution does not exist. This can be trivially seen when unrealistically small CV constraints are imposed with fixed sample sizes. A way to avoid this issue is through replacing hard constraints with soft constraints. Soft constraints can be violated without reducing the feasible region, but at the expense of increasing the objective function. A standard way of implementing soft objective functions is through the use of a penalty function. Penalty functions impose a positive-valued penalty for violating a constraint.

The unpenalized objective function is a p-norm of the vector of CVs for a stratified survey design with simple random sampling (SRS), SRS with replacement is shown for brevity,

$$\|f(\mathbf{X}|\mathbf{I},\eta)\|_p = \|\big(f(\mathbf{X}_{\cdot,1}|\mathbf{I},\eta),\ldots,f(\mathbf{X}_{\cdot,j}|\mathbf{I},\eta),\ldots,f(\mathbf{X}_{\cdot,J}|\mathbf{I},\eta)\big)\|_p, \qquad (1)$$

where

$\mathbf{x}_i =$ the vector valued administrative variable of length J available for all PSUs, identified with row i from matrix $\mathbf{X}$;
$\|\mathbf{x}_i\|_p =$ a p-norm of vector x_i equal to $\left(\sum_{j=1}^{J} x_{i,j}^p\right)^{1/p}$;
$\mathbf{I} =$ a vector of strata assignment parameters;
$\eta =$ a vector of sample sizes for each stratum;

$f(\mathbf{X}_{\cdot,j}|\mathcal{I},\eta) = \frac{\sqrt{\sum_h^H \eta_h^{-1} N_h^2 S_{h,j}^2}}{T_j}$ the CV for the j^{th} characteristic, $T_j = \sum_{i=1}^{N} x_{i,j}$ and $S_{h,j}^2 = \text{var}\left(\sum_{i=1}^{N} x_{i,j}\mathbb{I}_{\mathbf{I}_i=h} N_h^{-1}\right)$ with $h \in \{1,\ldots,H\}$ strata.

It is assumed that the number of strata is known *a priori*, the goal is to estimate the Horvitz-Thompson estimators for population totals, each element of $\mathbf{X}$ is nonnegative, and there is at least one positive valued $x_{i,j}$ for each j. To avoid issues with dividing by zero, it is also assumed that all strata have a minimum sample and population size of two. An extension to probability proportional to size sampling has been developed, but only SRS will be covered in this article.

If a set of quality constraints c are set through target CVs for $J^* \leq J$ administrative variables with a fixed total sample size, then for a given set of strata the problem can be written as

$$\text{argmin}_{\zeta,\mathbf{I}} \|\mathbf{\Phi} f(\mathbf{X}|\mathbf{I}, \eta)\|_p \tag{2}$$

subject to

1. $c_j \geq T_j^{-1} \sqrt{\sum_h^H \zeta_h N_h^2 S_{h,j}^2}, \ j \in \{1, \ldots, J^*\}$;
2. $\sum_h^H \zeta_h^{-1} = n$;
3. ζ_h^{-1} is an integer;
4. each $0 < \zeta_h \leq 1/2$.

where $\mathbf{\Phi}$ is a square matrix of dimension $J \times J$ with a diagonal equal to a vector of positive valued penalty weights. The diagonal of penalty weights from $\mathbf{\Phi}$ is identified as the vector ϕ; elements of ϕ help prioritize reduction of the CVs, regardless of the target CVs being met or not.

Soft constraints can be added to the objective function through the dot product of the penalty weights $\boldsymbol{\lambda}$ and penalty value vector $g(\mathbf{X}|\mathbf{c}, \mathbf{I}, \boldsymbol{\eta})$:

$$\text{argmin}_{\zeta,\mathbf{I}} \|\mathbf{\Phi} f(\mathbf{X}|\mathbf{I}, \boldsymbol{\eta})\|_p + \sum_{j \in J^{**}} \lambda_j g\left(\mathbf{X}_{\cdot,j}|\mathbf{I}, \boldsymbol{\eta}\right) \tag{3}$$

where J^{**} is the set of administrative variables with soft constraints subject to the same population and sample constraints in (2).

Each element of the penalty value vector $g(\mathbf{X}|\mathbf{I}, \boldsymbol{\eta})$ is equal to the maximum of 0 or $f(\mathbf{X}_{\cdot,j}|\mathbf{I}, \boldsymbol{\eta}) - c_j$. By this objective function, if all constraints are met, the problem is simply minimizing the norm of the vector formed by the product of $\mathbf{\Phi}$ and the vector of CVs. It is possible to simplify (3) by removing the hard constraints and replacing them with soft constraints using infinite valued penalty weights.

The choice of penalty weight vector $\boldsymbol{\lambda}$ can be motivated by targeting specific variables over others, or as a method to relate the administrative variables to the targeted response variables $\mathbf{Y}$. In the latter case, there are two potential solutions. The first would be to consider weights proportional to the absolute value of the correlation between the administrative variable and the response. This would favor a reduction in target CVs for variables with stronger relationships between $\mathbf{x_j}$ and $\mathbf{y_j}$ over weaker relationships for $j \in \{1, \ldots, J\}$. The second approach is the use of anticipated moments to explicitly model the response in the objective function. For brevity, only the second approach is covered.

Categorical administrative variables can be used through binning or grouping PSUs into disjoint sets identified by unique categorical values. This allows for the accommodation of industry by occupational groupings in establishment surveys or census blocks in area surveys.

2.2. Anticipated Moments

In practice, we are interested in estimating the Horvitz-Thompson estimators of unknown population characteristics that are correlated with an administrative variable. In Subsection 2.1, CV constraints placed on administrative variables serve as proxies for quality constraints on the estimators for the unknown population characteristics. A more direct approach is to place the CV constraints on the unknown population characteristic $\mathbf{Y}$ by an assumed model. This is accomplished by substituting the moments of $\mathbf{X}$ in (3), namely T_j and $S^2_{h,j}$, of the objective function with the complimentary anticipated moments of $\mathbf{Y}$. This is the multivariate generalization of the univariate approach by Baillargeon and Rivest (2009).

The exact form of the objective function is dependent on the choice of model for $\mathbf{Y}$ given $\mathbf{X}$. For many establishment surveys where y_i is a scalar, the model

$$y_i = x_i\beta + x_i^{\gamma}\epsilon_i, \tag{4}$$

$\mathbb{E}[\epsilon_i] = \mathbb{E}[\epsilon_i, \epsilon_{i'}] = 0$ where $(i \neq i')$, and $\mathbb{E}[\epsilon_i^2] = \sigma^2$ can provide a reasonable model (Kott et al. 2000). In the multivariate case considered, a generalization of this model for vectored value $\mathbf{y}_i$ is

$$\mathbf{y}_i = \mathbf{x}_i\mathbf{B} + \mathbf{V}_i\boldsymbol{\epsilon}_i, \tag{5}$$

$\mathbb{E}[\boldsymbol{\epsilon}_i] = \mathbb{E}[\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_{i'}] = 0$ where $(i \neq i')$, $\mathbb{E}[\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i] = \Sigma$, and $\mathbf{V}_i$ is a symmetric matrix of heteroscedastic weights for Σ.

To integrate the modeled response into the objective function, Anticipated Variance (AV) is used. Anticipated variance is simply the expectation both using the model ($\mathbb{E}_m$) and design, ($\mathbb{E}_d$),

$$\mathrm{AV}(\hat{T}_j) = \mathbb{E}_m\mathbb{E}_d\left[\hat{T}_j - \mathbb{E}_m\mathbb{E}_d[\hat{T}_j]\right]^2. \tag{6}$$

In the multivariate simple linear regression case with heteroscedastic variance, AV takes the form

$$\mathrm{AV}(\hat{T}_j) = \sum_{h=1}^{H}\left(1 - \frac{n_h}{N_h}\right)\frac{N_h^2}{n_h}\left(\sum_{i\in U_h} v_{i,j}^2\sigma^2 + \sum_{i\in U_h}(x_iB_j - \bar{x}_hB_j)^2\right). \tag{7}$$

Likewise, the Horvitz-Thompson estimator of the population total is,

$$\mathbb{E}_m\mathbb{E}_d[\hat{T}_j] = \sum_{i=1}^{N} = x_iB_j \tag{8}$$

and the anticipated coefficient of variation (ACV) can be estimated as

$$\widehat{ACV}(\hat{T}_j) = \frac{\sqrt{\mathrm{AV}(\hat{T}_j)}}{\sum_{i=1}^{N} x_iB_j} \tag{9}$$

Although not explored in this article, other models could be used in this framework.

2.3. Simulated Annealing

To minimize the objective function (3), we only make changes to $\boldsymbol{\eta}$ and $\mathbf{I}$. Since $\mathbf{I}$ is binary, this optimization is a combinatorial optimization problem, where simulated annealing is applicable. Simulated annealing is a stochastic optimization process that minimizes an objective function (possibly with constraints), and avoids the pitfalls of ending up in a local minima by admitting nonoptimal states. The general form of an algorithm to perform this stochastic process on the objective function m optimized over parameters θ in the finite dimensional parameter space Θ is detailed in Figure 1. Line five is the key component of the simulated annealing algorithm, where nonoptimal states can be accepted with nonzero probability ρ. This probability decreases as the number of iterations increases, allowing for both early exploration and eventual convergence of the simulated annealing sequence. The sequence $t(l)$ is called the cooling schedule and is a nonincreasing function that governs how quickly the probability of accepting a nonoptimal state decreases. Examples of $t(l)$ include $(l+1)^{-1}$ and $(\log(l+1))^{-1}$. The algorithm continues until either a fixed number of iterations L or threshold δ are met.

An advantage that simulated annealing has over other searches of binary spaces as seen in Benedetti et al. (2008) and Ballin and Barcaroli (2013) is the guaranteed theoretical convergence to a global minima by the simple condition that there is a non-zero transition probability between all possible states (Hajek 1988). This is not the case in Benedetti et al. (2008) where strata splits are chosen not by global optimality but local optimality. Similarly, genetic algorithms do not guarantee convergence to a global optimal solution in general.

One disadvantage of simulated annealing is its computational speed. Simulated annealing can be quite slow relative to other methods such as tree based methods that can partition a large number of sampling units at once. Similarly, genetic algorithms easily admit parallel implementations as opposed to simulated annealing which has serial dependence between each iteration (Henderson et al. 2003).

Algorithm 1 Simulated annealing algorithm.

1: **while** $l \leq L$ or $|m(\theta^{(l)}) - m(\theta^{(l-1)})| < \delta$, for objective function m **do**
2: Randomly generate a candidate state $\theta_*^{(l)}$, $l \geq 1$
3: **if** $\theta_*^{(l)}$ has a lower objective function than $\theta^{(l-1)}$ **then**
4: set $\theta^{(l)} = \theta_*^{(l)}$
5: **else if** $\rho = exp\{\Delta h_l / t(l)\} \geq U_l$, where $\Delta h_l = m(\theta^{(l-1)}) - m(\theta_*^{(l)})$ and $U_l \sim \text{Uniform}(0,1)$ **then**
6: set $\theta^{(l)} = \theta_*^{(l)}$
7: **else**
8: set $\theta^{(l)} = \theta^{(l-1)}$
9: **end if**
10: **end while**

Fig. 1. Pseudocode for the simulated annealing algorithm.

Simulated annealing applied to strata formation and allocation is straightforward and detailed in Figure 2. Each new candidate state consists of a PSU exchange and a change in allocation. The PSU exchange is generated by selecting a single PSU and a stratum to move it to. The change in allocation is generated by increasing the sample allocation for a randomly selected stratum by one, and decreasing the stratum allocation for another stratum by one. The new stratum can be the same stratum in which the PSU resides,

likewise there may be no change in allocation. To help improve the chance of retaining optimal allocation, multiple allocation exchanges are allowed per PSU change. In this case, subsequent assignment changes are only accepted if they improve the proposed objective function. In practice, the number of assignment changes required to maintain near-optimal allocation in each iteration is small for nonhighly skewed populations. This is due to a change in allocation dominating objective function changes when the sampling fraction is small.

In this application only linear cooling functions will be used, $t(l) = \alpha(l+1)^{-1}$ where α is a tunable parameter. Hard constraints on the objective function are handled by generating states that satisfy the imposed constraints.

Although Benedetti et al. (2010) and Benedetti and Piersimoni (2012) also uses simulated annealing, these methods differ in PSU selection. In the prior two papers, each PSU is iteratively selected ensuring each member of the population is offered a chance to move strata in a finite time-frame. The random search approach does not make this guarantee. Instead, it is assumed that for a sufficient number of iterations all PSUs are likely to be visited at least once. Furthermore, the introduction of nonuniform weighting in PSU selection for random searches could greatly improve performance of the proposed method by considering more likely PSU exchanges near stratum boundaries more frequently than more-extreme valued PSUs.

Algorithm 2 Simulated annealing algorithm for optimal stratification and allocation.

1: Start with initial stratification $\mathbf{I}^{(0)}$ and allocation $\boldsymbol{\eta}^0$
2: **while** $l \leq L$ or $m(\mathbf{y}|\mathbf{I},\boldsymbol{\eta}) - m(\mathbf{y}|\mathbf{I},\boldsymbol{\eta}) < \delta$, for objective function m **do**
3: Randomly generate a candidate state $\mathbf{I}_*^{(l)}, l \geq 1$
4: Randomly generate a candidate allocation $\boldsymbol{\eta}_*^{(l)}$ with respect to $\mathbf{I}_*^{(l)}$ to obtain $\boldsymbol{\eta}_*^{(l)}$
5: **if** $\left(\mathbf{I}_*^{(l)}, \boldsymbol{\eta}_*^{(l)}\right)$ has a lower objective function than $\left(\mathbf{I}^{(l-1)}, \boldsymbol{\eta}^{(l-1)}\right)$ **then**
6: set $\left(\mathbf{I}^{(l)}, \boldsymbol{\eta}^{(l)}\right) = \left(\mathbf{I}_*^{(l)} \boldsymbol{\eta}_*^{(l)}\right)$
7: **else if** $\rho = exp\{\Delta h_l / t(l)\} \geq U_l$, where $\Delta = m\left(\mathbf{I}^{(l)}, \boldsymbol{\eta}^{(l)}\right) - m\left(\mathbf{I}_*^{(l)} \boldsymbol{\eta}_*^{(l)}\right)$ and $U_l \sim \text{Uniform}(0,1)$ **then**
8: set $\left(\mathbf{I}^{(l)}, \boldsymbol{\eta}^{(l)}\right) = \left(\mathbf{I}_*^{(l)} \boldsymbol{\eta}_*^{(l)}\right)$
9: **else**
10: set $\mathbf{I}^{(l)} = \mathbf{I}^{(l-1)}$
11: **end if**
12: **end while**

Fig. 2. Pseudocode for the simulated annealing algorithm applied to joint stratification and allocation.

The allocation approach is similar to the random search of Kozak (2006a), divergence between the two methods occurs in both the use of penalized objective functions and the use of simulated annealing to achieve a final allocation. Instead, Kozak (2006a) considers only a sequence of allocations that are monotone increasing in objective function value. However, both Kozak (2006b) and the proposed method do produce integer based allocations, unlike Ballin and Barcaroli (2013) and Benedetti et al. (2008) that use Bethel (1986) allocation. Where the final allocation from Bethel (1986) is rounded up to the nearest integer. This rounding is generally nonoptimal, particularly when stratum sample sizes are small.

The performance of simulated annealing is governed by three primary factors (Henderson et al. 2003): choice of cooling schedule, the shape of the objective function surface, and the application or domain. The shape of the cooling schedule governs the

speed of convergence and the rate of accepting nonoptimal states. The literature on the choice of cooling schedule is largely based on heuristics balancing run-time and acceptable conditions (Romeo and Sangiovanni-Vincentelli 1991). Strenski and Kirkpatrick (1991) provide some theoretical results for extremely small populations with respect to optimal cooling schedules. The results of this analysis suggest that the linear, used here, or geometric cooling schedules tend to out perform more complex methods. Beyond the choice of cooling function, the only other controllable aspect of the optimization is the objective function. Objective functions that have shallow local minima tend to yield shorter run-times and better results due to the ease of escaping from nonoptimal states. In the application to optimal stratification and allocation, the depth of the local minima is a function of the shape of the function. Successful exchanges of PSUs with large values relative to the other PSUs, such as large operations in highly skewed populations, may cause changes in allocation. This can create local minima that are difficult to escape, causing nonoptimal solutions (Hajek 1988).

Implementation of a high-dimensional simulated annealing algorithm for nontrivial cases, however, is not so straightforward. The primary issues are:

- The computational cost of calculating the objective function;
- The likelihood of selecting a move that would reduce the objective function (improving convergence speed).

Calculating the objective function directly is computationally challenging. An alternative is to retain the $S_{h,j}^2$ component and to update a temporary candidate for $S_{h,j}^{2,*}$. Updates are performed through the numerically stable online sample variance calculation algorithm given by (Knuth 1997, 232), where online methods provide an iterative method to update the variance as opposed to recalculation of the variance. Periodic recalculation of the variances is provided to preserve numeric precision over a large number of updates, this recalculation occurs every 1,000,000 accepted exchanges. The numeric stability of the online algorithm was tested on simulated data using 1,000,000 iterations of the algorithm. This result was compared against $S_{h,j}^2$ calculated directly from the current stratification. The difference between these methods was less than e^{-12}. This error rate should be acceptable for most applications. However, care should be taken for exceptionally large populations. The application of a stable online method is also used by Benedetti et al. (2008) and Ballin and Barcaroli (2013) without periodic recalculation of variances.

3. Simulated Examples

The effectiveness of the multivariate joint stratification and allocation method using Simulated Annealing (SA) proposed in this article is compared to other methods in the literature through two examples: A univariate example comparing SA to the univariate joint stratification and allocation method of Lavallee-Hidiroglou (LH) using the R package *stratification* (Baillargeon and Rivest 2011), and a multivariate example comparing SA to the multivariate joint allocation and stratification using a genetic algorithm (GA) provided in theR package *SamplingStrata* (Barcaroli et al. 2014). The tree based method in Benedetti et al. (2008) was not considered due to lack of an available software package. Each of the two examples model PSU response through one of two linear models.

A homoscedastic linear model, and a heteroscedastic linear model where the variance is proportional to administrative data.

The univariate comparison between SA and LH has four goals: (1) Provide a diagnostic to ensure that SA has similar performance to known univariate optimal methods, as in Barcaroli et al. (2014). (2) Provide empirical results with respect to penalty weight selection. (3) Compare results using design variance and anticipated variance. (4) Show improvements that can be obtained over univariate methods with the presence of correlation. Similarly, the multivariate comparison between SA an GA has a two goals: (1) Compare SA and GA with respect to statistical efficiency. (2) As in the univariate example, to illustrate the advantage of using anticipated variance as a criterion for optimization.

In both examples, a population of 5,000 PSUs is simulated from two sets of linear models, a homoscedastic model and a heteroscedastic model. The homoscedastic model provides a simple case where the variance of the response is independent of the administrative data; the heteroscedastic model provides a more complex case where the variance of the response is proportional to the administrative data. This later case is common in many establishment surveys. In each set of linear models, each PSU, indexed by i has a vector valued response $\mathbf{y}_i = \{y_{i,1}, y_{i,2}, y_{i,3}, y_{i,4}, y_{i,5}\}$, and each element of the PSU is correlated with a vector $\mathbf{z}_i = \{z_{i,1}, z_{i,2}\}$ by a varying amount. Both elements of $\mathbf{z}_1$ and $\mathbf{z}_2$ are generated from a Chi-squared distribution with three degrees of freedom and scaled by 50 to produce values largely in the range of 0 to 1,000. This distribution is chosen to mimic the response of skewed populations common in establishment surveys. The relationship between a response y_i and z_i is determined by a linear model. The linear component of these models $x_i = z_i\beta_j \; j \in \{1, \ldots, 5\}$ will be used as an administrative variable for both examples and models, where β_j is assumed to be known. Examples and model will instead vary on the objective functions used for SA and comparisons to alternative methods.

In the homoscedastic linear model, the response vector for the i^{th} observation, y_i, is generated from linear models of the form,

- $y_{i,1} = z_i\beta_1 + \epsilon_{i,1}\bar{z}_1^{\gamma}$ with $\mathbb{E}[\epsilon_{i,1}^2] = \sigma_1^2$,
- $y_{i,p} = z_i\beta_p + \epsilon_{i,p}\|(\bar{z}_1, \bar{z}_2)\|_2^{\gamma}$ with $\mathbb{E}[\epsilon_{i,p}^2] = \sigma_p^2$ for $p \in \{2, 3, 4\}$, and
- $y_{i,5} = z_i\beta_5 + \epsilon_{i,5}\bar{z}_5\gamma$ with $\mathbb{E}[\epsilon_{i,5}^2] = \sigma_5^2$.

Each model error $\boldsymbol{\epsilon}_i = \{\epsilon_{i,1}, \epsilon_{i,2}, \epsilon_{i,3}, \epsilon_{i,4}, \epsilon_{i,5}\}$ is distributed $\mathcal{N}(0, \Sigma)$ where Σ is a diagonal matrix, and $\bar{z}_j$ is the mean of the variable z_j over all PSUs. To avoid cases where the response is less than zero, all negative values are truncated to zero. To provide similar magnitude of the variances in the heteroscedastic model the variance is scaled by the mean of norm of means from the vectors in $\mathbf{Z}$ raised to $\gamma = 0.75$.

The heteroscedastic model is similar in form and follows from the generalization of the linear model found in Brewer (1963). Specifically,

- $y_{i,1} = z_i\beta_1 + \epsilon_{i,1}z_{i,1}^{\gamma}$ with $\mathbb{E}[\epsilon_{i,1}^2] = \sigma_1^2$,
- $y_{i,p} = z_i\beta_p + \epsilon_{i,p}\|z_i\|_2^{\gamma}$ with $\mathbb{E}[\epsilon_{i,p}^2] = \sigma_p^2$ for $p \in \{2, 3, 4\}$, and
- $y_{i,5} = z_i\beta_5 + \epsilon_5 z_{i,2}^{\gamma}$ with $\mathbb{E}[\epsilon_{i,5}^2] = \sigma_5^2$.

Each model error ϵ_i is distributed $\mathcal{N}(0, \Sigma)$ where Σ is the identity matrix and γ is also set to 0.75. This value of γ was described by Kott et al. (2000) to be appropriate for many establishment surveys.

In both the homoscedastic and heteroscedastic models β is set to allow for different levels of correlation: β_1 and β_5 are fixed to the vectors $(1, 0)$ and $(0, 1)$ respectfully, and β_p is set at three different levels $(0.75, 0.25)$, $(0.5, 0.5)$, and $(0.25, 0.75)$ respectfully for $p \in \{2, 3, 4\}$. ACV constraints for y in both models are set at 0.04 for all variables. All calculations for SA use ϕ set to one for all variables, 50,000 PSU exchanges with ten iterations of optimal allocation per exchange, and cooling schedule $(l + 1)^{-1}$ where l is the current iteration. For simplicity, only soft constraints are used and are specified in each example. Details on the objective functions used in GA and SA, as well as their associated administrative functions are provided in Table 1.

3.1. Univariate Example

In this example, we consider a univariate approach to multivariate response using the LH joint allocation and stratification method, as described in Baillargeon and Rivest (2009), and compare it to the multivariate approach of SA. In this example, LH will be used as a diagnostic measure to ensure that SA can obtain an optimal result in a simple setting, identify performance characteristics using targeted penalty weights and no penalty weights, identify the importance of using anticipated variance, and finally to see if SA can further improve the results of the univariate allocation by finding an allocation that meets the univariate CV target while simultaneously improving the CVs of nontargeted administrative variables. Results of this example can be found in Tables 2 and 3.

Before examining the results of this comparison it is useful to consider some properties of LH relative to SA. LH can use either design variances or anticipated variance to determine the partitioning of the population into a fixed number of strata either with CV constraints without a fixed sample size, or with a fixed sample size and without CV constraints. In this example, we will be using the former case to set an initial sample size for SA.

To form the strata with fixed CV targets, LH uses an iterative algorithm. Since LH only works on univariate administrative data, strata can be identified as a set of disjoint intervals of the real line. Two approaches to find the boundaries of these intervals are found in Baillargeon and Rivest (2011): a model based approach used in the original Lavallée and Hidiroglou (1988) paper, and a random search method proposed in Kozak (2004). Due to the excellent performance characteristics without model assumptions, the random search method was chosen for this example.

Table 1. Objective functions used in the univariate and multivariate examples.

Method	Second moment	Objective function	Administrative data
LH	$\mathbb{E}_m\left[S^2_{h,1}\right]$	Neyman	$\mathbf{x} = \mathbf{Z}\beta_1$
GA	$S^2_{h,j}$	Bethel	$\mathbf{X} = \mathbf{Z}\beta$
SA	$S^2_{h,j}$	Equation (3)	$\mathbf{X} = \mathbf{Z}\beta$
SA (ACV)	$\mathbb{E}_m\left[S^2_{h,j}\right]$	Equation (3)	$\mathbf{X} = \mathbf{Z}\beta$

The random search approach starts with a set of initial intervals. Given these intervals, interval boundaries are perturbed at each iteration. This perturbation is performed by sorting the PSUs by administrative variable values and either moving the boundary forward or backward a random number of places in this sorted order. If the perturbed set of strata is more-optimal than the prior set per an objective function, then the perturbed set of strata is taken; otherwise, the perturbed set is discarded. The algorithm terminates when there is no change in stratification.

The objective function used in LH is simply Neyman allocation with a CV constraint:

$$n_h = n\frac{\sqrt{N_h^2 S_h^2}}{\sum_{k=1}^{H}\sqrt{N_k^2 S_k^2}} \text{ s.t. } f\left(\mathbf{X}|\mathbf{I},\eta\right) \leq c \tag{10}$$

In this example, $c = c_1$ is the univariate target, $S_h^2 = S_{h,1}^2$, and $f\left(\mathbf{x}|\mathbf{I},\eta\right) =$ is the coefficient of variance of the administrative data $\mathbf{x} = \mathbf{z}_1\beta_1$. S_h^2 can be substituted with the expected value of S_h^2 with an assumed model to provide an approximation to optimization with the anticipated variance

$$\mathbb{E}_m\left[S_h^2\right] = \frac{1}{N_h - 1}\left(\sum_{i\in U_h} v_i^2\sigma^2 + \sum_{i\in U_h}(z_{i,1}\beta_1 - \bar{z}_{1,h}\beta_1)^2\right) \tag{11}$$

where $v_i = \bar{z}_1$ in the homoscedastic case and z_i in the heteroscedastic case. All calculations are performed using the R package *stratification* (see Baillargeon and Rivest 2011).

Results for SA are calculated with four possible combinations of objective functions and penalty weighting. The two configurations of the objective function are identified in Tables 2 and 3 as "SA" and "SA (ACV)" with objective function specification identified in Table 1. The first configuration "SA" uses design variance in the objective function (3) with the administrative variables $\mathbf{X} = \mathbf{Z}\beta$. The second configuration "SA (ACV)" uses anticipated variance in the objective function (3) using the homoscedastic or heteroscedastic model. Targeted penalty weighting towards y_1 is used to provide a comparable result to LH, while the nontargeted weighting is used to compare the changes in attained ACV due to targeting a specific variable.

Results in ACV for the homoscedastic and heteroscedastic models are respectively provided in Tables 2 and 3. In both cases, LH chose six total strata and total sample size $n = 23$ in the homoscedastic case and $n = 65$ in the heteroscedastic case. All results are provided in anticipated coefficients of variation.

In addressing the goals of this example, SA (ACV) provides similar results to LH in that both methods attain the desired target CV for the same sample size in both homoscedastic and heteroscedastic cases. However, LH was able to reduce the ACV of y_1 further below the target of SA (ACV). The benefit of this further reduction is debatable, particularly if there is a benefit from reducing the ACV of other characteristics of interest in a survey. When there are other characteristics of interest, as in y_2 through y_5, SA (ACV) clearly outperforms LH.

With respect to penalty weight selection, the targeted weighting clearly had an effect on the result. This effect can be seen through the reduction of SA (ACV) and SA in the

Table 2. Anticipated CVs for simulated population generated from a homoscedastic linear model (univariate example).

Method Target	λ_1	λ_2	λ_3	λ_4	λ_5	ACV(y_1) 0.04	ACV(y_2) 0.04	ACV(y_3) 0.04	ACV(y_4) 0.04	ACV(y_5) 0.04	$\|\text{ACV}(y)\|_2$
Targeted weighting on y_1											
LH						0.0391	0.0540	0.0958	0.1406	0.1856	0.2604
SA	100	0	0	0	0	0.0411	0.0487	0.0873	0.1310	0.1754	0.2442
SA (ACV)	100	0	0	0	0	0.0400	0.0492	0.0865	0.1285	0.1713	0.2395
No targeted weighting											
SA	0	0	0	0	0	0.0820	0.0579	0.0451	0.0507	0.0698	0.1399
SA (ACV)	0	0	0	0	0	0.0693	0.0491	0.0446	0.0580	0.0802	0.1378

Table 3. Anticipated CVs for simulated population generated from a heteroscedastic linear model (univariate example).

Method	λ_1	λ_2	λ_3	λ_4	λ_5	ACV(y_1)	ACV(y_2)	ACV(y_3)	ACV(y_4)	ACV(y_5)	$\|\|ACV(y)\|\|_2$
Target						0.04	0.04	0.04	0.04	0.04	
Targeted weighting on y_1											
LH						0.0397	0.0647	0.0820	0.1050	0.1273	0.1993
SA	100	0	0	0	0	0.0542	0.0576	0.0564	0.0605	0.0609	0.1297
SA (ACV)	100	0	0	0	0	0.0400	0.0619	0.0778	0.0996	0.1204	0.1895
No targeted weighting											
SA	0	0	0	0	0	0.0553	0.0577	0.0559	0.0594	0.0593	0.1288
SA (ACV)	0	0	0	0	0	0.0553	0.0574	0.0555	0.0591	0.0594	0.1284

direction of y_1. The no target weighting case shows the degree of change that occurs by targeting a particular variable. As would be assumed, the amount of change in attained ACVs is associated with the degree of correlation with the targeted variable. Variables with high positive correlation with y_1 have ACVs that increase the least when y_1 is targeted (e.g., y_2); variables such as y_5 tend to have their ACVs increase the most.

The heteroscedastic case is important, in that strata containing larger values of the administrative variable will have higher model variance. Therefore, the impact of ignoring the model variance is more extreme than only using the design variance. This can clearly be seen in the results of SA compared to those of LH and SA (ACV). The later two results that use anticipated variance tend to consistently meet targets in both cases, while SA using just the design variance almost hits the target using the homoscedastic model, but considerably misses the target in the heteroscedastic case.

3.2. Multivariate Example

In the multivariate example, we reuse the prior population in the univariate example but apply the joint optimal allocation and stratification methods GA and SA. The goal of this example is to compare the statistical efficiency of the resulting survey designs using GA and SA, as well as revisit the topic of using ACVs in optimization.

Because GA as presented in Ballin and Barcaroli (2013) does not support targeting ACVs, the algorithm uses $X = \boldsymbol{Z}\beta$ as known administrative data. As in the univariate example, the results identified as SA are also fit in the same manner; SA (ACV) uses anticipated variance in the objective function. Results are found in Tables 4 and 5. To illustrate the importance of specifying a design using ACVs, the stratifications attained for GA and SA are presented both using attained CVs from the administrative data and ACVs. It is important to note that optimization using design variances are identical in the homogenous and heterogenous case, as they ignore the model variance. Therefore attained CVs are only listed in Table 4.

To provide comparable results between GA and SA the optimal sample size and number of strata from GA are used for the SA based optimizations. In this example, the optimal sample size using GA is 193 and the total number of strata is five.

Individual PSUs are used for atomic strata for GA, and minor performance tuning was performed. Tuning proved problematic due to the long run-time of GA, averaging two hours and 35 minutes per run. Run-times of SA, on the other hand, averaged 25 seconds for both the CV and ACV optimizations.

In both the homoscedastic and heteroscedastic cases, GA was less efficient than SA when only considering CV targets, but produced more robust stratifications. This robustness appears to be a result of attaining a local minima, instead of a feature of the GA algorithm. As in the univariate example, SA (ACV) provided uniformly better results with respect to attained ACV than SA and GA. GA did do reasonably well in the homoscedastic case, meeting all CV targets. With ACV evaluation criteria, GA met one ACV target for both the homoscedastic and heteroscedastic cases, but did not suffer from larger departures from the target as in the case of SA.

The result for SA demonstrate the importance of using ACVs in an objective function. In this result, there was considerable reduction in CV. However, this reduction in CV was

Table 4. CVs and ACVs for simulated population using the homoscedastic model. The first set of results ignore anticipated variance in the objective function and are CV targets are reported, the resulting ACV of the first set of results and SA using anticipated variance in the objective function are provided in the second set of results.

Method	λ_1	λ_2	λ_3	λ_4	λ_5	$\text{CV}(y_1)$	$\text{CV}(y_2)$	$\text{CV}(y_3)$	$\text{CV}(y_4)$	$\text{CV}(y_5)$	$\|\text{CV}(y)\|_2$
Target						0.04	0.04	0.04	0.04	0.04	
GA						0.0397	0.0308	0.0274	0.0311	0.0397	0.0763
SA	0	0	0	0	0	0.0274	0.0200	0.0171	0.0202	0.0272	0.0509
Method	λ_1	λ_2	λ_3	λ_4	λ_5	$\text{ACV}(y_1)$	$\text{ACV}(y_2)$	$\text{ACV}(y_3)$	$\text{ACV}(y_4)$	$\text{ACV}(y_5)$	$\|\text{ACV}(y)\|_2$
GA						0.0419	0.0418	0.0348	0.0443	0.0420	0.0919
SA	0	0	0	0	0	0.0630	0.0493	0.0439	0.0490	0.0616	0.1205
SA (ACV)	0	0	0	0	0	0.0238	0.0242	0.0235	0.0252	0.0255	0.0547

Table 5. ACVs for simulated population using the heteroscedastic model.

Method	λ_1	λ_2	λ_3	λ_4	λ_5	ACV(y_1)	ACV(y_2)	ACV(y_3)	ACV(y_4)	ACV(y_5)	$\|\|ACV(y)\|\|_2$
GA						0.0451	0.0423	0.0397	0.0422	0.0451	0.0960
SA	0	0	0	0	0	0.0673	0.0590	0.0544	0.0584	0.0660	0.1369
SA (ACV)	0	0	0	0	0	0.0344	0.0346	0.0328	0.0344	0.0342	0.0762

done at the expense of ACV in both the homoscedastic and heteroscedastic cases, where the attained ACV was over double the attained CV. This is an over fitting and model misspecification problem, where assuming the control data as the response yielded a nondesirable outcome. In practice, care should be taken to test multiple potential response models on a potential stratification.

4. June Agricultural Survey

Application of the simulated annealing based multivariate optimal stratification and allocation method is examined in the proposed redesign of the United States Department of Agriculture (USDA) National Agricultural Statistics Service's (NASS) June Agricultural Survey (JAS). In this section, JAS and the proposed redesign are introduced along with a discussion of administrative variables and implementation details. A proxy of JAS provides a comparable design using the same administrative data and PSUs. This proxy is then compared to the simulated annealing based stratification and allocation method in this article, followed by a discussion of the results.

JAS is a two-stage annual area survey of the continental United States, producing estimates of acreage devoted to various agricultural land uses and other spatially associated estimates (Davies 2009). JAS is administered by NASS, with data collected by The National Association of State Departments of Agriculture (NASDA) employees. The first stage of JAS is a stratified simple random sample design with replacement, where strata are formed by grouping PSUs based on the percentage of cultivated acres within each PSU. When needed, specialty strata are used to target rare commodities or demographic groups. Each PSU in the first stage is a contiguous one-to-eight square mile region of land manually delineated along permanent geographic features such as roads. Cultivated acreage for each PSU is calculated using NASS's Cropland Data Layer (CDL), a remotely-sensed administrative data set of land-cover and land-use (Boryan et al. 2011). PSUs are sampled using systematic sampling of a spatial index, allowing for a spatially well distributed sample. In the second stage, selected PSUs are partitioned into smaller areas of land known as segments, serving as Secondary Sampling Units (SSU)s. Segments are manually formed by the delineation of PSUs into approximately one-square-mile contiguous regions of high agricultural production; larger segments can be drawn in areas with no-to-low agricultural activity. A single segment is selected randomly from each PSU, and all land within the selected segment is fully enumerated. Nonresponse is handled through observation, remote sensing, or subject matter experts. Allocation in JAS is performed by Bethel (1986) using historic data with equal cost per PSU.

To lower design costs and to allow for estimation of year-to-year change JAS is replicated. A set of replicates are created every five years and all of these replicates are

rotated into the sample one year at a time. Each replicate is collected for approximately five years, and then rotated out of the sample.

National level target CVs are chosen by NASS to ensure quality estimates. The CV targets are predictive in nature, as they are set for the estimates, not the administrative data. Target CVs are considered satisfied if on-average the attained CVs are less than the target CVs. This comparison occurs at the level of precision of the target CV. Target CVs are typically met each year, but occasionally some targets cannot be attained for a given sample size.

Iowa State in cooperation with NASS, considered an update of the current JAS design (Zimmer et al. 2013). In this proposed redesign, the two-stage design is replaced by a single-stage design with optimal stratification based on areal units of one-square-mile in size. These PSUs, known as sections, are part of a permanent area frame based on the Public Land Survey System (PLSS). This permanent frame greatly reduces survey cost, as the current JAS requires the labor intensive manual delineation of PSUs and SSUs. Stratification of the proposed redesign's area frame is based on the optimal joint allocation and stratification algorithm described in this article. Like JAS, this design is calculated using equal PSU costs. Spatial balance is attained by using the local pivotal method in Grafström et al. (2012) and implemented using the *BalancedSampling* R package (Grafström and Lisic 2016). Unfortunately, the current implementation of the simulated annealing procedure only supports the Sen-Yates-Grundy variance estimates (Sen 1953), over estimating the variance when using locally balanced sampling; instead, simple random sampling with replacement is used as an upper bound for the variance with an assumptions of spatial clustering (Grafström et al. 2012).

As in the current JAS design, remotely sensed CDL data is used as administrative data. The CDL has accuracy above 90% for corn and soybeans as well as above 80% for winter and spring wheat for all years used in this research (2008–2015). This makes the CDL a fairly useful tool in evaluating surveys, unfortunately, linear models of section acreage are not particularly good at predicting future land use. This is due to the agricultural practice of crop rotations, where individual fields within a section tend to follow crop specific sequences to maximize yield, mitigate pests, and reduce erosion. Instead of directly modeling these crop sequences, an assumption is made that sequences observed within a period of time are likely to re-occur within a future window of time. Using this assumption, we use the prior four years to predict the next four years. This is similar to the current JAS practice where a single stratification is used for multiple years.

Due to lack of correlation between nonacreage based estimators, such as number of farms and livestock, with available administrative data, the joint optimal allocation and stratification method is only used for acreage based estimates. To ensure that quality constraints are met for nonacreage estimators, prior JAS response is used to calculate historical variances. These historical variances are used to ensure that the total sample size is of sufficient size to meet the imposed quality constraints. To check for any potential deleterious effects caused by unforeseen relationships with the nonacreage responses and the optimal stratification and allocation, quality constraints are checked by post stratifying geo-referenced, but not complete, historical data by the new design.

In this article, multivariate stratification and allocation are only performed on the PSUs of the redesign (sections). The original JAS design is proxied by a set of univariate bounds based on cultivated acres (Table 6). A proxy is used, instead of the original JAS design,

Table 6. Approximation of JAS stratification.

Stratum	Definition
11	75% or more cultivated land
20	50–74% cultivated land
30	15–49% cultivated land
40	1–14% cultivated land
50	0% cultivated land

due to the intractability of modeling the manual segmentation of secondary segmentation units. For the purpose of evaluation, only results for South Dakota were considered with the acreages of interest including corn, soybeans, winter wheat, spring wheat, and cultivated acres. South Dakota provides a reasonable use case for multivariate allocation with a large number of crop types and large frame of close to 80,000 sections.

National level target CVs from JAS cannot be applied to a single state; therefore, historical JAS CVs (2008–2011) for South Dakota are used as CV targets. Multiple years of land use are used to form strata to account for year-to-year land cover variability. In this example 2008–2011 are used to form the stratification, and 2012–2015 are used to evaluate future land use. The CDL variables used for stratification are cultivated, corn, soybeans, winter wheat, and spring wheat acreages. These variables were used for all segments in the population, and optimization was applied to each variable and year combination from 2008–2011 simultaneously treating each combination as a separate variable. The algorithm is run for 5,000,000 iterations with five sample size optimization steps per allocation. Initial stratification is performed by K-means to accelerate convergence. The penalty weights φ and λ were set to 1 and 1,000 respectively for each combination of year and land cover. The total run-time using this parameterization using the proposed method is 30 minutes. In both the univariate and the multivariate cases all 80,000 segments were assigned to five strata. Allocation for the JAS strata is performed by the multivariate allocation method described in (Bethel 1986). The total sample size from this allocation is used for the simulated annealing based approach. The highly correlated administrative data is used as a proxy for the true response. The resulting CVs from both methods provided in Table 8 for the multivariate method and Table 7 for the method approximating the current JAS.

Table 7. CVs for specified variables based on an approximate JAS stratification and allocation of South Dakota.

	Cultivated	Corn	Soybeans	Winter wheat	Spring wheat
Target	0.01	0.05	0.05	0.19	0.16
2008	0.0162	0.0470	0.0524	0.1078	0.1129
2009	0.0153	0.0453	0.0510	0.1148	0.1153
2010	0.0168	0.0455	0.0484	0.1239	0.1151
2011	0.0137	0.0410	0.0483	0.1139	0.1275
2012	0.0147	0.0395	0.0477	0.1378	0.1389
2013	0.0158	0.0409	0.0500	0.1546	0.1404
2014	0.0167	0.0428	0.0481	0.1514	0.1327
2015	0.0168	0.0457	0.0488	0.1606	0.1310

Table 8. CVs for specified variables based on multivariate joint stratification and allocation of South Dakota. Changes relative to the univariate method (Table 7) in parenthesis.

	Cultivated	Corn	Soybeans	Winter wheat	Spring wheat
Target	0.01	0.05	0.05	0.19	0.16
2008	0.0137 (−15.43%)	0.0496 (5.53%)	0.0534 (1.91%)	0.1138 (5.57%)	0.1200 (6.29%)
2009	0.0118 (−22.88%)	0.0475 (4.86%)	0.0508 (−0.39%)	0.1197 (4.27%)	0.1219 (5.72%)
2010	0.0138 (−17.86%)	0.0465 (2.20%)	0.0475 (−1.86%)	0.1293 (4.36%)	0.1207 (4.87%)
2011	0.0119 (−13.14%)	0.0429 (4.63%)	0.0482 (−0.21%)	0.1178 (3.42%)	0.1338 (4.94%)
2012	0.0125 (−14.97%)	0.0403 (2.03%)	0.0492 (3.14%)	0.1449 (5.15%)	0.1452 (4.54%)
2013	0.0137 (−13.29%)	0.0419 (2.44%)	0.0511 (2.20%)	0.1599 (3.43%)	0.1466 (4.42%)
2014	0.0144 (−13.77%)	0.0436 (1.87%)	0.0487 (1.25%)	0.1571 (3.76%)	0.1352 (1.88%)
2015	0.0146 (−13.09%)	0.0469 (2.63%)	0.0494 (1.23%)	0.1623 (1.06%)	0.1339 (2.21%)

The results of this empirical example showed a general improvement in the CVs for the multivariate method relative to the univariate method in areas where the CV targets were difficult to attain. Considering JAS rounding rules, this improvement allows the multivariate method to meet the target CVs of all crops both on average (2012–2015) and per-year. The rounded CV targets in the univariate case are generally met, with the exception of total cultivated acreage that was only met in 2012. However, the rounding rule tends to favor the multivariate method over stating its performance relative to the univariate method.

For both methods, the most difficult to attain target CV is total cultivated acreage, where the multivariate method averaged a -13.78% decrease in CV relative to the univariate method in the evaluation years (2012–2015). The univariate method has lower CVs for other crops within the evaluation years, but most of these CVs for other crops are well below the target CVs for both methods. Other results included indications of model misspecification in the multivariate method through the general increase in attained CVs for the predicted years.

5. Discussion

In this article, a method to construct optimal multivariate stratified designs for an arbitrary, but fixed, number of self-representing strata was presented. This method admits a combination of hard and soft constraints, where soft constraints are handled using a penalized objective function and hard constraints are handled through traditional nonlinear programming constraints. Optimization of the objective function is performed using simulated annealing, by moving individual PSUs between strata. Simultaneous allocation is provided by also considering changes in the allocation as part of the simulated annealing algorithm. Penalized weighting in the objective function allowed for flexibility in design specification, allowing for penalty weights to target specific commodities based on preference or correlation with administrative variables. The use of anticipated variance in the objective function was shown to account for uncertainty in the relationship between administrative data and targeted estimates, and opens the door to modeling nonsampling error. Applications to both a simulated population and the proposed redesign of JAS were provided. Important issues with the proposed method, beyond the scope of this research, include investigations of nonsampling error, handling poor quality administrative data, model misspecification when using a model-assisted objective function, and improvements to computational speed. Future application specific research with respect to the JAS redesign, and potentially other spatial surveys, includes improvements to the objective function to reflect better the variance using spatially balanced sampling methods and the development of better prediction models for agricultural land use for individual PSUs.

In application, the proposed method was shown to be computationally tractable for reasonably large populations and more flexible than existing methods through the use of soft constraints and the use of anticipated variance in the objective function. The two examples are chosen to show utility of the method in existing establishment and area surveys. Application to more complex designs has not been considered in this research, but the model-assisted objective function could be extended to account for subsampling and other traits of complex designs. Application to more complicated designs, requiring

optimization of multiple samples may also be possible for paneled or split-questionnaire designs as in Ioannidis et al. (2016).

In a univariate example, both homoscedastic and heteroscedastic populations are investigated to describe establishment surveys with different dispersion patterns. In both the anticipated moments-based approach, SA performed better than the existing anticipated moments-based LH approach. Provided that the relationship between the administrative variables and the target variables are reasonably well known, the proposed method should provide improvements over LH in multivariate scenarios.

In the multivariate example, the proposed method greatly out-performs the genetic algorithm approach in optimization. Improvements to the genetic algorithm-based approach, such as the adoption of anticipated moments in the objective function and finer tuning of parameters, may provide parity between the results. Furthermore, model misspecification, and prospective use of related goodness-of-fit diagnostics should be explored for both methods. However, the proposed method may be more applicable for larger populations due to the long run-time of the genetic algorithm relative to the proposed method.

In the JAS redesign, the proposed method met or exceeded the attained CVs of the JAS approximation under JAS rounding rules, and in general had lower target CVs for hard-to-attain targets, providing a strong argument for the use of multivariate designs onthis survey. This method was also shown to be computationally feasible for population sizes of 80,000 with a run-time of thirty minutes: computation for larger populations should be possible at the expense of longer run-times. The computational speed and stability of the proposed method improves on existing methods through the use of online-variance calculation with periodic recalculation of variances. The use of prior information, as in the case of JAS may not be possible for other area surveys, nor advisable if the underlying stochastic process changes over time.

The applicability to other area surveys is largely dependent on the variance estimation method employed, number of PSUs, availability of administrative data, and the ability to model individual PSUs. The current objective function only considers SRS with or without replacement, not accounting for increases in efficiency that could be attained using more advanced sampling methods. The application of the proposed method to a population of 280,000 PSUs has been explored by Lisic et al. (2015), but general applicability to surveys with considerably larger populations has not been explored. Modeling individual PSUs is not needed to apply the proposed method to a survey using administrative data based quality constraints. However, if quality constraints are placed on the estimates either correlation-based weights or a model should be introduced. The correlation-based weights may be useful under linear relationships. However, their use for more complicated relationships is uncertain. The applicability in the case of a model would depend on how well the model describes the uncertainty in the relationship between the administrative data and the response.

Although not explored in this research, this anticipated moment approach allows incorporation of estimates of nonsampling errors such as assumed nonresponse in the response variable. This feature can already be found in (Baillargeon and Rivest 2014) for univariate optimal allocation and stratification. Correlation-based penalty weighting can also incorporate nonsampling error within the correlation function. However, this

approach may be limited to cases when the relationships between the administrative variables and the survey response is fairly linear.

A similar problem addressed within the context of the JAS redesign, but not in general is the handling of poor quality administrative variables. This can occur when only a subset of the frame has complete records, such as using prior survey data or incomplete databases. Provided that an accurate measure of uncertainty can be obtained for each observation, the anticipated variance framework can provide an optimal allocation and stratification. However, this question is beyond the scope of this research in this article.

Another interesting, but unexplored, area of research within this article is the importance of model specification for the anticipated moment approach. In the simulated populations, it is assumed that the model is known. In application, this is an unlikely case. Therefore, future analysis of the effect of model misspecification, and prospective use of related goodness-of-fit diagnostics should be explored more thoroughly.

Further acceleration of the proposed method may extend the applicability to larger populations. Two ways to improve the computational speed of the presented method for larger populations is through selecting PSUs near stratum boundaries with greater probability and exchanging multiple PSUs. These PSUs are more likely to be accepted during exchanges, allowing for faster convergence of the algorithm. The current implementation already supports the use of static weights to increase the probability that a particular PSU is selected. However, finding the ideal properties of these weights has not been considered yet. For these large population sizes moving individual PSUs between strata may result in infeasible run-time. One solution to this problem is by exchanging clusters of PSUs or partitioning strata by identifying useful hyperplanes in the space of administrative variables.

6. References

Baillargeon, S. and L.-P. Rivest. 2009. "A General Algorithm for Univariate Stratification." *International Statistical Review* 77(3): 331–344. Doi: http://dx.doi.org/10.1111/j.1751-5823.2009.00093.x.

Baillargeon, S. and L.-P. Rivest. 2011. "The Construction of Stratified Designs in R with the Package Stratification." *Survey Methodology* 37(1): 53–65.

Baillargeon, S. and L.-P. Rivest. 2014. "Stratification: Univariate Stratification of Survey Populations." Available at: https://CRAN.R-project.org/package=stratification, r package version 2.2-5 (accessed January 18, 2017).

Ballin, M. and G. Barcaroli. 2013. "Joint Determination of Optimal Stratification and Sample Allocation Using Genetic Algorithm." *Survey Methodology* 39(2): 369–393.

Barcaroli, G. 2014. "Sampling Strata: An R Package for the Optimization of Stratified Sampling." *Journal of Statistical Software* 61(4): 1–24. Available at: https://www.jstatsoft.org/article/view/v061i04/v61i04.pdf (accessed January 20, 2018).

Benedetti, R., M. Bee, and G. Espa. 2010. "A Framework for Cut-Off Sampling in Business Survey Design." *Journal of Official Statistics* 26(4): 651–671.

Benedetti, R., G. Espa, and G. Lafratta. 2008. "A Tree-Based Approach to Forming Strata in Multipurpose Business Surveys." *Survey Methodology* (34): 195–203.

Benedetti, R. and F. Piersimoni. 2012. "Multivariate Boundaries of a Self Representing Stratum of Large Units in Agricultural Survey Design." *Survey Research Methods* 6: 125–135. Doi: http://dx.doi.org/10.18148/srm/2012.v6i3.5127.

Bethel, J.W. 1986. *An Optimum Allocation Algorithm for Multivariate Surveys*. Technical report, United States Department of Agriculture, Statistical Reporting Service. Available at: https://naldc.nal.usda.gov/naldc/download.xhtml?id=43260&content=PDF (accessed January 20, 2017).

Boryan, C., Z. Yang, R. Mueller, and M. Craig. 2011. "Monitoring US Agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program." *Geocarto International* 26: 341–358. Doi: http://dx.doi.org/10.1080/10106049.2011.562309.

Brewer, K.R.W. 1963. "Ratio Estimation and Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process." *Australian Journal of Statistics* 5(3): 93–105. Doi: http://dx.doi.org/10.1111/j.1467-842X.1963.tb00288.x.

Dalenius, T. and J.L.J. Hodges. 1959. "Minimum Variance Stratification." *Journal of the American Statistical Association* 54: 88–101.

Davies, C. 2009. *Area Frame Design for Agricultural Surveys*. Technical report, United States Department of Agriculture, National Agricultural Statistics Service. Available at: https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Advanced_Topics/AREA%20FRAME%20DESIGN.pdf (accessed January 20, 2018).

Day, C.D. 2009. "Evolutionary Algorithms for Optimal Sample Design." A paper presented at the 2009 Federal Committee on Statistical Methodology Conference, November 4, 2009. Washington DC: Federal Committee on Statistical Methodology. Available at: https://fcsm.sites.usa.gov/files/2014/05/2009FCSM_Day_XI-A.pdf (accessed January 20, 2018).

Freuder, E.C. and R.J. Wallace. 1992. "Partial Constraint Satisfaction." *Artificial Intelligence* 58: 21–70. Doi: https://doi.org/10.1016/0004-3702(92)90004-H.

Gentle, J. and C. Perry. 2000. "Optimal Stratification of Area Frames." In Proceedings of the second international conference on establishment surveys - II, June 17–21, 2000. 1354–1382. Buffalo, NY: American Statistical Association. Available at: https://ww2.amstat.org/meetings/ices/2000/proceedings/S04.pdf (accessed January 20, 2018).

Gonzalez, J.M. and J.L. Eltinge. 2010. "Optimal Survey Design: A Review." In Proceedings of the Survey Research Methods Section of the American Statistical Association. Joint Statistical Meeting, October 4, 2010. 4970–83. Vancouver, BC: American Statistical Association. Available at: https://www.bls.gov/osmr/pdf/st100270.pdf (accessed January 20, 2018).

Grafström, A. and J. Lisic. 2016. *Balanced Sampling: Balanced and Spatially Balanced Sampling*. Available at: https://CRAN.R-project.org/package=BalancedSampling, r package version 1.5.1. (accessed March 8, 2016).

Grafström, A., N.L. Lundström, and L. Schelin. 2012. "Spatially Balanced Sampling through the Pivotal Method." *Biometrics* 68: 514–520. Doi: http://dx.doi.org/10.1111/j.1541-0420.2011.01699.x.

Gunning, P., J. Horgan, and W. Yancey. 2004. "Geometric Stratification of Accounting Data." *Contaduría y Administración*. Available at: http://www.ejournal.unam.mx/rca/214/RCA21401.pdf (accessed January 20, 2018).

Hajek, B. 1988. "Cooling Schedules for Optimal Annealing." *Mathematics of Operations Research* 13(2): 311–329. Doi: https://doi.org/10.1287/moor.13.2.311.

Henderson, D., S.H. Jacobson, and A.W. Johnson. 2003. "The Theory and Practice of Simulated Annealing." In *Handbook of Metaheuristics*, edited by F. Glover and G.A. Kochenberger, 287–319. Boston, MA: Springer.

Horgan, J.M. 2006. "Stratification of Skewed Populations: A Review." *International Statistical Review* 74: 67–76. Doi: http://dx.doi.org/10.1111/j.1751-5823.2006.tb00161.x.

Ioannidis, E., T. Merkouris, L.-C. Zhang, M. Karlberg, M. Petrakos, F. Reis, and P. Stavropoulos. 2016. "On a Modular Approach to the Design of Integrated Social Surveys." *Journal of Official Statistics* 32(2): 259–286. Doi: https://doi.org/10.1515/jos-2016-0013.

Isaki, C.T. and W.A. Fuller. 1982. "Survey Design under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77(377): 89–96. Doi: http://dx.doi.org/10.2307/2287773.

Knuth, D.E. 1997. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, 3 edition.

Kott, P.S. and J.T. Bailey. 2000. "The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling." In Proceedings of the second international conference on establishment surveys - II, June 17–21, 2000. 269–304. Buffalo, NY: American Statistical Association. Available at: https://ww2.amstat.org/meetings/ices/2000/proceedings/S04.pdf (accessed January 20, 2018).

Kozak, M. 2004. "Optimal Stratification Using Random Search Method in Agricultural Surveys." *Statistics in Transition* 6(5): 797–806. Available at: http://omega.sggw.waw.pl/~m.kozak/SIT2004_6_5.pdf (accessed November 16, 2017).

Kozak, M. 2006a. "Multivariate Sample Allocation: Application of Random Search Method." *Statistics in Transition* 7(4): 889–900. Available at: https://www.researchgate.net/profile/Marcin_Kozak/publication/242329930_MULTIVARIATE_SAMPLE_ALLOCATION_APPLICATION_OF_RANDOM_SEARCH_METHOD/links/02e7e51cc8ede254cc000000.pdf (accessed November 16, 2017).

Kozak, M. 2006b. "On Sample Allocation in Multivariate Surveys." *Communications in Statistics-Simulation and Computation* 35: 901–910. Doi: http://dx.doi.org/10.1080/03610910600880286.

Lavallée, P. and M. Hidiroglou. 1988. "On the Stratification of Skewed Populations." *Survey Methodology* 14: 33–43.

Lisic, J.J. 2016. "saAlloc: Stratification and Allocation of Sampling Units Using Simulated Annealing." Available at: https://github.com/jlisic/saAlloc, r package version 2.0 (accessed January 20, 2018).

Lisic, J.J., H. Sang, Z. Zhu, and S. Zimmer. 2015. "Optimal Stratification and Allocation for the June Area Survey." A paper presented at the 2015 Federal Committee on Statistical Methodology Conference, December 2, 2015. Washington, DC: Federal Committee on Statistical Methodology. Available at: https://fcsm.sites.usa.gov/files/2016/03/E2_Lisic_2015FCSM.pdf (accessed January 20, 2018).

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21: 1087–1092. Doi: https://doi.org/10.1063/1.1699114.

R Core Team. 2015. "R: A Language and Environment for Statistical Computing." *R Foundation for Statistical Computing*, Vienna, Austria, Available at: https://www.R-project.org/ (accessed July 1, 2017).

Romeo, F. and A. Sangiovanni-Vincentelli. 1991. "A Theoretical Framework for Simulated Annealing." *Algorithmica* 6(1): 302–345. Doi: https://doi.org/10.1007/BF01759049.

Särndal, C.-E., B. Swensson, and J. Wretman. 1991. *Model Assisted Survey Sampling*. New York: Springer.

Sen, A.R. 1953. "On the Estimate of the Variance in Sampling with Varying Probabilities." *Journal of the Indian Society of Agricultural Statistics* 5(1194): 127.

Strenski, P.N. and S. Kirkpatrick. 1991. "Analysis of Finite Length Annealing Schedules." *Algorithmica* 6(1): 346–366. Doi: https://doi.org/10.1007/BF01759050.

Valliant, R., F. Hubbard, S. Lee, and C. Chang. 2014. "Efficient Use of Commercial Lists in US Household Sampling." *Journal of Survey Statistics and Methodology* 2(2): 182–209. Doi: http://dx.doi.org/10.1093/jssam/smu006.

Zimmer, S., J.-K. Kim, and S. Nusser. 2013. "Automatic Stratification for an Agricultural Area Frame Using Remote Sensing Data." In Proceedings of the 59th ISI World Statistics Congress, 25–30 August 2013, Hong Kong, 1952–1957 Available at: www.statistics.gov.hk/wsc/STS043-P5-S.pdf (accessed June 12, 2014).

Received May 2016
Revised July 2017
Accepted November 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 149–180, http://dx.doi.org/10.1515/JOS-2018-0008

Components of Gini, Bonferroni, and Zenga Inequality Indexes for EU Income Data

Leo Pasquazzi[1] *and Michele Zenga*[1]

In this work we apply a new approach to assess contributions from factor components to income inequality. The new approach is based on the insight that most (synthetic) inequality indexes may be viewed as (weighted) averages of point inequality measures, which measure inequality between population subgroups identified by income. Assessing contributions of factor components to point inequality measures is usually an easy task, and based on these contributions it is straightforward to define contributions to the corresponding (synthetic) overall inequality indexes as well. As we shall show through an analysis of income data from Eurostat's European Community Household Panel Survey (ECHP), the approach based on point inequality measures gives rise to readily interpretable results, which, we believe, is an advantage over other methods that have been proposed in literature.

Key words: Inequality decomposition; factor components; point inequality measures; synthetic inequality index.

1. Introduction

A great deal of literature about income inequality is concerned with evaluation of contributions to inequality from factor components. A common approach to this problem is to express some given (synthetic) inequality index as sum of terms, with one term corresponding to each factor component, which are then interpreted as contributions to inequality. The interpretations are justified by showing that the terms representing the contributions are functions of some descriptive statistics for the joint distribution of the factor components and total income. In connection with the well-known Gini index, this approach has, for example, been applied by Rao (1969); Lerman and Yitzhaki (1984, 1985), and Radaelli and Zenga (2005).

Shorrocks (1982), on the other hand, explores an axiomatic approach. He considers a broad class of inequality indexes, but is faced with the problem that under a fairly general set of restrictions there exists an infinite number of potential decomposition rules for every given inequality index. To solve this nonuniqueness problem, he adds two further restrictions which imply that the relative contributions (or "proportional contributions" in his language) from the components are the same for all inequality indexes and are equal to those corresponding to what he calls the "natural decomposition rule" for the variance. In a

[1] University of Milano-Bicocca, Dept. of Statistics and Quantitative Methods, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy. Emails: leo.pasquazzi@unimib.it and michele.zenga@unimib.it
Acknowledgment: This research was supported by the FAR 2014-ATE-0200 and the FAR 2015ATE-0353 grants from the University of Milano-Bicocca.

later paper (Shorrocks 1983) Shorrocks acknowledges that not everyone might agree on the restrictions imposed to derive the "unique" decomposition rule, but he still defends that rule by showing that in applications to some empirical datasets it gives rise to reasonable results.

In the present article we illustrate, through an application to income data from the European Community Household Panel (ECHP), a new approach to factor component decomposition. This approach has been recently suggested by Zenga et al. (2012), and was originally developed for the inequality index I (Zenga 2007a). In a later paper it has been extended to the Gini and Bonferroni indexes as well (Zenga 2013). The new approach is based on the fact that these three inequality indexes are, *by their original definitions*, (weighted) averages of point inequality measures which measure inequality between population subgroups identified by income. Defining factor component contributions to the point inequality measures is, as we shall show below, an easy and straightforward task, and taking appropriate averages of these contributions yields decomposition rules for the (synthetic) inequality indexes as well.

The rest of this article is organized as follows. In Section 2 we recall the definitions of the Gini, Bonferroni, and Zenga indexes in terms of point inequality measures. In Section 3 we show how the decomposition rules based on the point inequality indexes are derived and in Section 4 we highlight some interesting relations between factor component contributions to inequality and shares on total population income. Since income distributions are usually available in the form of survey data with weights associated to each sample unit, we devoted Section 5 to estimation from survey data. Finally, in Section 6 we provide an application to data from the 2001 wave of the ECHP in order to give some insight into the range of possible outcomes. To help the reader to recall the meaning of certain symbols which we shall introduce in the course of this article, we added a list of notations at the end of the article.

2. The Gini, Bonferroni, and Zenga Indexes as Averages of Point Inequality Indexes

Let

$$y_1 \leq y_2 \leq \cdots \leq y_N \tag{1}$$

denote total income Y of individuals or families belonging to a given population, and let

$$p_i := \frac{i}{N}, \quad i = 1, 2, \ldots, N, \tag{2}$$

and

$$q_i := \frac{\sum_{\nu=1}^{i} y_\nu}{\sum_{\nu=1}^{N} y_\nu}, \quad i = 1, 2, \ldots, N, \tag{3}$$

denote the cumulative population and income shares, respectively. When Gini (1914) first proposed what later became the virtually most widely used inequality index, he set out from the fact that the cumulative income shares q_i can never exceed their corresponding

cumulative population shares p_i. Thus, he proposed

$$R_i := \frac{p_i - q_i}{p_i}, \quad i = 1, 2, \ldots, N, \tag{4}$$

as basic point inequality measures from which he derived the definition of his well-known synthetic inequality index

$$R := \frac{\sum_{i=1}^{N-1} R_i p_i}{\sum_{i=1}^{N-1} p_i} = \frac{\sum_{i=1}^{N-1} (p_i - q_i)}{\sum_{i=1}^{N-1} p_i}. \tag{5}$$

Thereafter he showed that R is linked to the graph with the Lorenz curve (Lorenz 1905) in the sense that R is equal to the ratio between the "concentration area" and the area of the triangle with vertices in (0, 0), $((N - 1)/N, 0)$ and (1, 1) (sometimes called the "maximum concentration area").

Starting from the observation that the mean income

$$M_i^-(Y) := \frac{1}{i} \sum_{\nu=1}^{i} y_\nu, \quad i = 1, 2, \ldots, N, \tag{6}$$

of the i "poorest" population members cannot exceed the mean income

$$M(Y) := M_N^-(Y) = \frac{1}{N} \sum_{\nu=1}^{N} y_\nu \tag{7}$$

of the whole population, Bonferroni (1930) proposed the inequality index

$$B := \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{M(Y) - M_i^-(Y)}{M(Y)}. \tag{8}$$

As pointed out by DeVergottini (1940), B can also be viewed as unweighted average of the point inequality measures R_i proposed by Gini (1914). In fact,

$$\frac{M(Y) - M_i^-(Y)}{M(Y)} = \frac{\frac{\sum_{\nu=1}^{N} y_\nu}{N} - \frac{\sum_{\nu=1}^{i} y_\nu}{i}}{\frac{\sum_{\nu=1}^{N} y_\nu}{N}} = \frac{p_i - q_i}{p_i} =: R_i.$$

More recently, Zenga (1984, 2007a) introduced two new types of point inequality measures and put forward corresponding synthetic inequality indexes. In the present article we shall consider only the latter proposal. It is based on the point inequality measures given by

$$I_i := \frac{M_i^+(Y) - M_i^-(Y)}{M_i^+(Y)}, \quad i = N_1, N_2, \ldots, N_k, \tag{9}$$

where

$$M_i^+(Y) := \begin{cases} \dfrac{1}{N-i}\sum_{\nu=i+1}^{N} y_i & \text{if } i = 1, 2, \ldots, N-1 \\ y_N & \text{if } i = N \end{cases} \tag{10}$$

and where $N_1 < N_2 < \cdots < N_k = N$ are the cumulative frequencies corresponding to the k *different* values taken on by total income Y. Using the point inequality measures I_i, Zenga (2007a) defined the synthetic inequality index

$$I := \frac{1}{N}\sum_{s=1}^{k} I_{N_s}\, n_s \tag{11}$$

where $n_1, n_2, \ldots, n_k$ denote the absolute frequencies of the k different values observed for total income Y.

Notice that as opposed to the indexes proposed by Gini and Bonferroni, Zenga's synthetic inequality index I involves only the point inequality measures at $i = N_1, N_2, \ldots, N_k$, which, as will be seen in the next section, makes it easier to apply the approach to factor component decomposition based on point inequality measures.

Before moving on to factor component decomposition, we provide a brief list of references regarding the synthetic Zenga index I. Applications to real distributions may be found in Zenga (2007b), Zenga (2008), and Greselin et al. (2013). Polisicchio (2008), Polisicchio and Porro (2009), Porro (2008), and Porro (2011) deal with properties of the curve defined by the point inequality measures I_i and its relation with the Lorenz curve. Inferential problems related to the I index have been analyzed in Greselin and Pasquazzi (2009), Greselin et al. (2010), Langel and Tillé (2012), Antal et al. (2011), and Greselin et al. (2014). As for decomposition rules, Radaelli (2008a) proposed a subgroups decomposition for the point inequality indexes I_i and the synthetic I index that has been applied to income data in Radaelli (2007), Radaelli (2008b), and Greselin et al. (2009) and that has been compared with a subgroups decomposition rule for Gini's index in Radaelli (2010). Finally, as already mentioned above, the decomposition rule considered in the present work has been originally proposed in Zenga et al. (2012) and has been extended to the Gini and Bonferroni indexes in Zenga (2013).

3. Factor Component Contributions to Inequality

Assume

$$y_i := x_{i,1} + x_{i,2} + \cdots + x_{i,c}, \quad i = 1, 2, \ldots, N, \tag{12}$$

where $x_{i,j}$ denotes the income from factor component X_j of the ith individual or household. Obviously,

$$\sum_{\nu=1}^{i} y_\nu = \sum_{\nu=1}^{i} x_{\nu,1} + \sum_{\nu=1}^{i} x_{\nu,2} + \cdots + \sum_{\nu=1}^{i} x_{\nu,c}$$

so that

$$M(Y) = M(X_1) + M(X_2) + \cdots + M(X_c), \quad i = 1, 2, \ldots, N, \tag{13}$$

$$M_i^-(Y) = M_i^-(X_1) + M_i^-(X_2) + \cdots + M_i^-(X_c), \quad i = 1, 2, \ldots, N, \tag{14}$$

and

$$M_i^+(Y) = M_i^+(X_1) + M_i^+(X_2) + \cdots + M_i^+(X_c), \quad i = 1, 2, \ldots, N, \tag{15}$$

where $M(X_j)$, $M_i^-(X_j)$ and $M_i^+(X_j)$ are defined as $M(Y)$, $M_i^-(Y)$ and $M_i^+(Y)$, respectively, with $x_{i,j}$ in place of y_i. It is important to note that while $M_i^-(Y)$ is the mean of the i smallest values observed for total income Y, this is usually not the case for $M_i^-(X_j)$. In fact, $M_i^-(X_j)$ is the mean of the i smallest values observed for factor component X_j only if Y and X_j are perfectly rank correlated (the situation is analogous for $M_i^+(Y)$ and $M_i^+(X_j)$).

Using relations (13), (14), and (15) yields simple decomposition rules for the point inequality indexes R_i and I_i. In fact, it is easily seen that

$$R_i := \frac{M(Y) - M_i^-(Y)}{M(Y)} = \sum_{j=1}^{c} \frac{M(X_j) - M_i^-(X_j)}{M(Y)}, \quad i = 1, 2, \ldots, N,$$

and

$$I_i := \frac{M_i^+(Y) - M_i^-(Y)}{M(Y)} = \sum_{j=1}^{c} \frac{M^+(X_j) - M_i^-(X_j)}{M(Y)}, \quad i = 1, 2, \ldots, N,$$

so that

$$\mathcal{R}_i(X_j) := \frac{M(X_j) - M_i^-(X_j)}{M(Y)}, \quad j = 1, 2, \ldots, c \tag{16}$$

and

$$\mathcal{I}_i(X_j) := \frac{M_i^+(X_j) - M_i^-(X_j)}{M_i^+(Y)}, \quad j = 1, 2, \ldots, c \tag{17}$$

can be interpreted as contributions from the factor components X_j to R_i and I_i, respectively. The corresponding relative contributions have a very neat interpretation:

$$\rho_i(X_j) := \frac{\mathcal{R}_i(X_j)}{R_i} = \frac{M(X_j) - M_i^-(X_j)}{M(Y) - M_i^-(Y)}, \quad i = 1, 2, \ldots, N-1, \tag{18}$$

and

$$\zeta_i(X_j) := \frac{\mathcal{I}_i(X_j)}{I_i} = \frac{M_i^+(X_j) - M_i^-(X_j)}{M_i^+(Y) - M_i^-(Y)}, \quad i = 1, 2, \ldots, N, \tag{19}$$

are simply the contributions from factor component X_j to $M(Y) - M_i^-(Y)$ and to $M_i^+(Y) - M_i^-(Y)$, respectively (observe that ρ_N is not defined because $R_N = 0$). The interpretations of $\rho_i(X_j)$ and $\zeta_i(X_j)$ can actually be interchanged since for $i = 1, 2, \ldots, N-1$ these relative contributions are always the same. This, perhaps

unexpected result follows immediately from the fact that

$$\frac{N-i}{N}\left(M_i^+(\cdot)-M_i^-(\cdot)\right)=M(\cdot)-M_i^-(\cdot),\quad i=1,2,\ldots,N-1. \tag{20}$$

Based on the contributions $\mathcal{R}_i(X_j)$ and $\mathcal{I}_i(X_j)$, it is straightforward to define contributions to the corresponding synthetic inequality indexes as well. In fact,

$$R:=\frac{\sum_{i=1}^{N-1}R_i\,p_i}{\sum_{i=1}^{N-1}p_i}=\frac{\sum_{i=1}^{N-1}\sum_{j=1}^{c}\mathcal{R}_i(X_j)\,p_i}{\sum_{i=1}^{N-1}p_i}=\sum_{j=1}^{c}\frac{\sum_{i=1}^{N-1}\mathcal{R}_i(X_j)\,p_i}{\sum_{i=1}^{N-1}p_i},$$

$$B:=\frac{1}{N-1}\sum_{i=1}^{N-1}R_i=\frac{1}{N-1}\sum_{i=1}^{N-1}\sum_{j=1}^{c}\mathcal{R}_i(X_j)=\sum_{j=1}^{c}\frac{1}{N-1}\sum_{i=1}^{N-1}\mathcal{R}_i(X_j)$$

and

$$I:=\frac{1}{N}\sum_{s=1}^{k}I_{N_s}n_s=\frac{1}{N}\sum_{s=1}^{k}\sum_{j=1}^{c}\mathcal{I}_{N_s}(X_j)\,n_s=\sum_{j=1}^{c}\frac{1}{N}\sum_{s=1}^{k}\mathcal{I}_{N_s}(X_j)\,n_s,$$

and the expressions on the far right suggest to consider

$$\mathcal{R}(X_j):=\frac{\sum_{i=1}^{N-1}\mathcal{R}_i(X_j)\,p_i}{\sum_{i=1}^{N-1}p_i},\quad j=1,2,\ldots,c, \tag{21}$$

$$\mathcal{B}(X_j):=\frac{1}{N-1}\sum_{i=1}^{N-1}\mathcal{R}_i(X_j),\quad j=1,2,\ldots,c, \tag{22}$$

and

$$\mathcal{I}(X_j):=\frac{1}{N}\sum_{s=1}^{k}I_{N_s}(X_j)\,n_s,\quad j=1,2,\ldots,c, \tag{23}$$

as contributions to the synthetic inequality indexes R, B, and I, respectively. The corresponding relative contributions are then given by

$$\rho(X_j):=\frac{\mathcal{R}(X_j)}{R}=\frac{\sum_{i=1}^{N-1}\mathcal{R}_i(X_j)\,p_i}{\sum_{i=1}^{N-1}R_i\,p_i}=\frac{\sum_{i=1}^{N-1}\rho_i(X_j)\,R_i\,p_i}{\sum_{i=1}^{N-1}R_i\,p_i} \tag{24}$$

$$\beta(X_j):=\frac{\mathcal{B}(X_j)}{B}=\frac{\sum_{i=1}^{N-1}\mathcal{R}_i(X_j)}{\sum_{i=1}^{N-1}R_i}=\frac{\sum_{i=1}^{N-1}\rho_i(X_j)\,R_i}{\sum_{i=1}^{N-1}R_i} \tag{25}$$

$$\zeta(X_j) := \frac{\mathcal{I}(X_j)}{I} = \frac{\sum_{s=1}^{k} \mathcal{I}_{N_s}(X_j)\, n_s}{\sum_{s=1}^{k} I_{N_s}\, n_s} = \frac{\sum_{i=1}^{N-1} \zeta_{N_s}(X_j)\, I_{N_s}\, n_s}{\sum_{i=1}^{N-1} I_{N_s}\, n_s}, \tag{26}$$

and are thus nothing else than weighted averages, with different sets of weights, of essentially the same relative contributions (recall $\rho_i(X_j) = \zeta_i(X_j)$ for $i = 1, 2, \ldots, N-1$ and for $j = 1, 2, \ldots, c$). $\rho(X_j)$, $\beta(X_j)$ and $\zeta(X_j)$ can thus be interpreted as average values of the contributions of the factor components X_j to the N differences $M(Y) - M_i^-(Y)$ or $M_i^+(Y) - M_i^-(Y)$.

However, there might be a nonuniqueness problem in the definitions of the contributions. The problem occurs if there are several population members with the same total income Y and with different incomes from two or more factor components X_j. In this case, the values of $M_i^-(X_j)$ and $M_i^+(X_j)$ for $i \neq N_1, N_2, \ldots, N_k$ depend on the i index assigned to the population members with same total income Y, and thus the corresponding contributions $\mathcal{R}_i(X_j)$ and $\mathcal{I}_i(X_j)$ depend on this assignment as well. It follows that $\mathcal{R}(X_j)$ and $\mathcal{B}(X_j)$ depend on the way in which the i indexes are assigned, while for $\mathcal{I}(X_j)$ this is not the case, because $\mathcal{I}(X_j)$ depends only on the contributions $\mathcal{I}_i(X_j)$ for $i = N_1, N_2, \ldots, N_k$. Even though in large populations with few repeated values for total income Y this dependence has little impact on the results, we propose an easy way to neutralize it: instead of the original definitions, one might consider modified versions of the Gini and Bonferroni indexes that are weighted averages of the point inequality measures R_i and B_i just for $i = N_1, N_2, \ldots, N_k$. A convenient modified version of the Gini index is for example given by

$$R' := \frac{\sum_{s=1}^{k} R_{N_s}\, r_s}{\sum_{s=1}^{k} r_s} \tag{27}$$

where

$$r_s := \begin{cases} N_s\,(n_s + n_{s+1}) & \text{if } 1 \leq s < k \\ N\, n_k & s = k \end{cases} \tag{28}$$

while for the Bonferroni index we suggest

$$B' := \frac{1}{N} \sum_{s=1}^{k} R_{N_s}\, n_s. \tag{29}$$

A few comments are due regarding the definitions of R' and B'. In first place we observe that in large populations with few repeated values R' and B' are close to R and B, respectively. Second, it is worth noting that the definitions of R' and B', as opposed to those of R and B, include the point inequality measure R_N even though $R_N = 0$ for every income distribution: we made this choice for ease of comparison with the Zenga index which depends on k point inequality measures as well. Finally, regarding the definition of R', it is not difficult to show that it coincides with the ratio between the "concentration area" and the area of triangle with vertices in (0, 0), (1, 0), and (1, 1) (see the proof in the Appendix).

The factor component contributions to R' and B' are obviously defined as

$$\mathcal{R}'(X_j) := \frac{\sum_{s=1}^{k} \mathcal{R}_{N_s}(X_j)\, r_s}{\sum_{s=1}^{k} r_s}, \quad j = 1, 2, \ldots, c, \tag{30}$$

and

$$\mathcal{B}'(X_j) := \frac{1}{N} \sum_{s=1}^{k} \mathcal{R}_{N_s}(X_j)\, n_s, \quad j = 1, 2, \ldots, c, \tag{31}$$

and the corresponding relative contributions are then given by

$$\rho'(X_j) := \frac{\mathcal{R}'(X_j)}{R'} = \frac{\sum_{s=1}^{k} \mathcal{R}_{N_s}(X_j) r_s}{\sum_{s=1}^{k} R_{N_s} r_s} = \frac{\sum_{s=1}^{k} \rho_{N_s}(X_j)\, R_{N_s} r_s}{\sum_{s=1}^{k} R_{N_s} r_s} \tag{32}$$

and

$$\beta'(X_j) := \frac{\mathcal{B}'(X_j)}{B'} = \frac{\sum_{s=1}^{k} \mathcal{R}_{N_s}(X_j)\, n_s}{\sum_{s=1}^{k} R_{N_s} n_s} = \frac{\sum_{s=1}^{k} \rho_{N_s}(X_j)\, R_{N_s} n_s}{\sum_{s=1}^{k} R_{N_s} n_s}. \tag{33}$$

4. Contributions to Inequality and Shares on Population Income

As suggested by Zenga et al. (2012), it is instructive to compare the relative contributions $\rho_i(X_j)$ and $\zeta_i(X_j)$ and their weighted averages $\rho(X_j)$, $\beta(X_j)$, and $\zeta(X_j)$ (as well as $\rho'(X_j)$ and $\beta'(X_j)$) with the share

$$\gamma(X_j) := \frac{\sum_{i=1}^{N} x_{i,j}}{\sum_{i=1}^{N} y_i} \tag{34}$$

of their corresponding factor component X_j on total population income. In fact, in the hypothetical case, the so-called *scale transformation hypothesis*, where

$$x_{i,j} = \gamma(X_j)\, y_i \quad \text{for every } \; i = 1, 2, \ldots, N,$$

one would have

$$M_i^-(X_j) = \gamma(X_j)\, M_i^-(Y) \quad \text{and} \quad M_i^+(X_j) = \gamma(X_j)\, M_i^+(Y)$$

for all $i = 1, 2, \ldots, N$, so that

$$\rho_i(X_j) = \zeta_i(X_j) = \gamma(X_j) \quad \text{for } i = 1, 2, \ldots, N-1 \text{ and } \zeta_N = \gamma(X_j).$$

In this case it follows that

$$\rho(X_j) = \beta(X_j) = \zeta(X_j) = \gamma(X_j).$$

In real income distributions one should obviously expect that

$$x_{i,j} \neq \gamma(X_j)\, y_i$$

for most population members i, but since the deviations $x_{i,j} - \gamma(X_j)\, y_i$ must sum (over i) to zero, the scale transformation hypothesis provides a useful benchmark against which to compare the actual distribution of the factor components. For illustrative purposes we shall next describe two types of deviations from the scale transformation hypothesis that are helpful for the interpretation of the relative contributions:

- First, consider the case where

$$x_{i,j} < \gamma(X_j)\, y_i \quad \text{for } i = 1, 2, \ldots, i^* < N$$

and

$$x_{i,j} \geq \gamma(X_j)\, y_i \quad \text{for } i = i^* + 1, i^* + 2, \ldots, N.$$

Since y_i is nondecreasing in i, we can describe this as a situation where *all* population members with total income Y below a given threshold value y_{i^*} have less income from factor component X_j than they would have under the scale transformation hypothesis, while *all* other (more fortunate) population members have at least as much income from X_j as they would have under the scale transformation hypothesis. It is not difficult to show that in this case

$$M_i^-(X_j) < \gamma(X_j)\, M_i^-(Y) \quad \text{and} \quad M_i^+(X_j) > \gamma(X_j)\, M_i^+(Y) \tag{35}$$

for all $i = 1, 2, \ldots, N$, so that

$$\rho_i(X_j) = \zeta_i(X_j) > \gamma(X_j) \text{ for } i = 1, 2, \ldots, N-1 \quad \text{and} \quad \zeta_N > \gamma(X_j).$$

From these inequalities it follows that

$$\rho(X_j) > \gamma(X_j), \quad \beta(X_j) > \gamma(X_j) \quad \text{and} \quad \zeta(X_j) > \gamma(X_j). \tag{36}$$

The first two inequalities hold also with $\rho'(X_j)$ and $\beta'(X_j)$ in place of $\rho(X_j)$ and $\beta(X_j)$, respectively.

- The second case is opposite to the first one. It occurs when

$$x_{i,j} > \gamma(X_j)\, y_i \quad \text{for } i = 1, 2, \ldots, i^* < N$$

and

$$x_{i,j} \leq \gamma(X_j)\, y_i \quad \text{for } i = i^* + 1, i^* + 2, \ldots, N.$$

In this case,

$$M_i^-(X_j) > \gamma(X_j)\, M_i^-(Y) \quad \text{and} \quad M_i^+(X_j) < \gamma(X_j)\, M_i^+(Y) \tag{37}$$

for all $i = 1, 2, \ldots, N$, so that

$$\rho_i(X_j) = \zeta_i(X_j) < \gamma(X_j) \quad \text{for } i = 1, 2, \ldots, N-1 \text{ and } \quad \zeta_N < \gamma(X_j).$$

Therefore it follows that

$$\rho(X_j) < \gamma(X_j), \quad \beta(X_j) < \gamma(X_j) \quad \text{and} \quad \zeta(X_j) < \gamma(X_j). \tag{38}$$

Also here, the first two inequalities hold also with $\rho'(X_j)$ and $\beta'(X_j)$ in place of $\rho(X_j)$ and $\beta(X_j)$, respectively.

The two cases described above are somewhat artificial in that they require that *all* population members with total income below (above) a certain threshold value have smaller (larger) income from factor component X_j than they would have under the scale transformation hypothesis. Nevertheless, we can regard the inequalities in (36) (and in (38)) as symptomatic for situations where income from a given factor component X_j tends to be more concentrated among population members with large (small) total income Y than total income Y itself. In fact, if $\gamma(X_j)$ is positive (which is usually the case), the inequalities in (35) imply that

$$\frac{\sum_{\nu=1}^{i} x_{\nu,j}}{\sum_{\nu=1}^{N} x_{\nu,j}} < \frac{\sum_{\nu=1}^{i} y_\nu}{\sum_{\nu=1}^{N} y_\nu}$$

and

$$\frac{\sum_{\nu=i+1}^{N} x_{\nu,j}}{\sum_{\nu=1}^{N} x_{\nu,j}} > \frac{\sum_{\nu=i+1}^{N} y_\nu}{\sum_{\nu=1}^{N} y_\nu}$$

for $1 \leq i \leq N-1$, while those in (37) imply that

$$\frac{\sum_{\nu=1}^{i} x_{\nu,j}}{\sum_{\nu=1}^{N} x_{\nu,j}} > \frac{\sum_{\nu=1}^{i} y_\nu}{\sum_{\nu=1}^{N} y_\nu}$$

and

$$\frac{\sum_{\nu=i+1}^{N} x_{\nu,j}}{\sum_{\nu=1}^{N} x_{\nu,j}} < \frac{\sum_{\nu=i+1}^{N} y_\nu}{\sum_{\nu=1}^{N} y_\nu}$$

for $1 \leq i \leq N-1$.

5. Estimation from Survey Data

The definitions of the Gini, Bonferroni, and Zenga indexes and the decomposition rules outlined in Section 3 can be directly applied to *population data*. In this section we propose estimators which can be applied to *survey data* and which should be reasonably well-behaved for a broad class of sample designs. So let

$$S = \{i_1, i_2, \ldots, i_d\} \tag{39}$$

denote a set of indexes corresponding to a sample of d units drawn from the population $\mathcal{U} = \{1, 2, \ldots, N\}$ and let

$$w_{i_1}, w_{i_2}, \ldots, w_{i_d} \tag{40}$$

denote survey weights corresponding to the d sample units in S. In what follows we shall assume that the survey weights w_i are strictly positive and that they are scaled so that

$$\sum_{i \in S} w_i = N. \tag{41}$$

The estimators we shall propose below do not actually depend on how the survey weights are scaled. Assumption (41) is only needed to make the estimators look more similar to their corresponding population quantities.

Now, suppose there are $\hat{k} \leq d$ different values for total income Y among the d observed values in the sample, and denote these values by

$$\tilde{y}_1 < \tilde{y}_2 < \cdots < \tilde{y}_{\hat{k}}. \tag{42}$$

For $s = 1, 2, \ldots, \hat{k}$, let

$$\hat{n}_s := \sum_{i \in S : y_i = \tilde{y}_s} w_i \tag{43}$$

denote the sum of the survey weights w_i corresponding to the sample units with total income Y equal to $\tilde{y}_s$. Moreover, let

$$\hat{N}_s := \sum_{\nu=1}^{s} \hat{n}_\nu, \quad s = 1, 2, \ldots, \hat{k}, \tag{44}$$

denote the corresponding cumulative weights. Obviously, $\hat{N}_{\hat{k}} = N$. Based on the cumulative weights define

$$\sigma(p) := \min\{s : \hat{N}_s \geq N p\}, \quad p \in [0, 1]. \tag{45}$$

Then, use $\sigma(p)$ to define

$$\hat{M}_p^-(Y) := \frac{\sum_{s=1}^{\sigma(p)} \tilde{y}_s \hat{n}_s}{\sum_{s=1}^{\sigma(p)} \hat{n}_s}, \tag{46}$$

$$\hat{M}_p^+(Y) := \begin{cases} \dfrac{\sum_{s=\sigma(p)+1}^{\hat{k}} \tilde{y}_s \hat{n}_s}{\sum_{s=\sigma(p)+1}^{\hat{k}} \hat{n}_s} & \text{if } \sigma(p) < \hat{k}, \\ \tilde{y}_{\hat{k}} & \text{if } \sigma(p) = \hat{k}, \end{cases} \tag{47}$$

and observe that $\hat{M}_p^-(Y)$ and $\hat{M}_p^+(Y)$ at $p = i/N$ can be taken as estimators for $M_i^-(Y)$ and $M_i^+(Y)$, respectively. Note, however, that the estimators $\hat{M}_p^-(Y)$ and $\hat{M}_p^+(Y)$ are defined for every $p \in [0, 1]$ and that they give rise to right continuous step functions with discontinuities at $p = \hat{N}_s/N$ for $s = 1, 2, \ldots, \hat{k}$. Obviously, $\hat{M}_p^-(Y)$ at $p = 1$ is equal to the weighted sample mean

$$\hat{M}(Y) := \frac{\sum_{s=1}^{\hat{k}} \tilde{y}_s \hat{n}_s}{\sum_{s=1}^{\hat{k}} \hat{n}_s}. \tag{48}$$

On the other hand, $\hat{M}_p^+(Y)$ at $p = 0$ is larger than the weighted sample mean $\hat{M}(Y)$, unless there are no different values for total income Y in the sample in which case $\hat{M}_p^+(Y)$ would not be defined for any $p \in [0, 1]$. The latter case is obviously not of interest in applications.

To estimate the point inequality measures, let

$$\hat{R}_p := \frac{\hat{M}_p^-(Y) - \hat{M}(Y)}{\hat{M}(Y)} \tag{49}$$

and

$$\hat{I}_p := \frac{\hat{M}_p^+(Y) - \hat{M}_p^-(Y)}{\hat{M}_{,p}^+(Y)}, \tag{50}$$

and, as before, put $p = i/N$ to get estimators for R_i and I_i, respectively.

Next, consider the synthetic inequality indexes. To define an estimator for R', let

$$\hat{r}_s := \begin{cases} \hat{N}_s\,(\hat{n}_s + \hat{n}_{s+1}) & \text{if } 1 \le s < \hat{k} \\ \hat{N}_{\hat{k}}\,\hat{n}_k & \text{if } s = \hat{k} \end{cases} \tag{51}$$

and use $\hat{r}_s$ in place of the weights r_s and $\hat{R}_p$ at $p = \hat{N}_s/N$ in place of R_{N_s} in the definition of R'. The resulting estimator is then given by

$$\hat{R}' := \frac{\sum_{s=1}^{\hat{k}} \hat{R}_{\hat{N}_s/N}\,\hat{r}_s}{\sum_{s=1}^{\hat{k}} \hat{r}_s} \tag{52}$$

and, under suitable conditions, it can be used to estimate R as well. Similar reasoning suggests that B' and B can be estimated by

$$\hat{B}' := \frac{1}{\hat{N}_{\hat{k}}} \sum_{s=1}^{\hat{k}} \hat{R}_{\hat{N}_s/N}\,\hat{n}_s \tag{53}$$

and that

$$\hat{I} := \frac{1}{\hat{N}_{\hat{k}}} \sum_{s=1}^{\hat{k}} \hat{I}_{\hat{N}_s/N}\,\hat{n}_s. \tag{54}$$

can be used to estimate I.

Now, consider the population quantities involving the factor components X_j. For their estimation we shall employ the weighted averages given by

$$\tilde{x}_{s,j} := \frac{\sum_{i \in S: y_i = \tilde{y}_s} x_{i,j}\,w_i}{\sum_{i \in S: y_i = \tilde{y}_s} w_i} = \frac{1}{\hat{n}_s} \sum_{i \in S: y_i = \tilde{y}_s} x_{i,j}\,w_i, \tag{55}$$

for $s = 1, 2, \ldots, \hat{k}$ and $j = 1, 2, \ldots, c$. Note that $\tilde{x}_{s,j}$ is the weighted average of income from factor component X_j among the sample units with total income equal to $\tilde{y}_s$. Using $\hat{M}_p^-(X_j)$ and $\hat{M}_p^+(X_j)$ to indicate $\hat{M}_p^-(Y)$ and $\hat{M}_p^+(Y)$ with $\tilde{x}_{s,j}$ in place of $\tilde{y}_s$, we define the

step functions

$$\hat{\mathcal{R}}_p(X_j) := \frac{\hat{M}_p^-(X_j) - \hat{M}(X_j)}{\hat{M}(Y)} \tag{56}$$

and

$$\hat{\mathcal{I}}_p(X_j) := \frac{\hat{M}_p^+(X_j) - \hat{M}_p^-(X_j)}{\hat{M}_p^+(Y)}, \tag{57}$$

which, at $p = i/N$, provide estimates for the contributions $\mathcal{R}_i(X_j)$ and $\mathcal{I}_i(X_j)$. Based on the step functions $\hat{\mathcal{R}}_p(X_j)$ and $\hat{\mathcal{I}}_p(X_j)$ we further construct estimators for the contributions $\mathcal{R}'(X_j)$, $\mathcal{B}'(X_j)$ and $\mathcal{I}(X_j)$. These are given by

$$\hat{\mathcal{R}}'(X_j) := \frac{\sum_{s=1}^{\hat{k}} \hat{\mathcal{R}}_{\hat{N}_s/N}(X_j)\,\hat{r}_s}{\sum_{s=1}^{\hat{k}} \hat{r}_s}, \tag{58}$$

$$\hat{\mathcal{B}}'(X_j) := \frac{1}{\hat{N}_{\hat{k}}} \sum_{s=1}^{\hat{k}} \hat{\mathcal{R}}_{\hat{N}_s/N}(X_j)\,\hat{n}_s \tag{59}$$

and

$$\hat{\mathcal{I}}(X_j) := \frac{1}{\hat{N}_{\hat{k}}} \sum_{s=1}^{\kappa} \hat{\mathcal{I}}_{\hat{N}_s/N}(X_j)\,\hat{n}_s \tag{60}$$

Also here, under suitable conditions, we can regard $\hat{\mathcal{R}}'(X_j)$ and $\hat{\mathcal{B}}'(X_j)$ as well as estimators of $\mathcal{R}(X_j)$ and $\mathcal{B}(X_j)$, respectively.

Since

$$\sum_{j=1}^{c} \hat{M}_p^-(X_j) = \hat{M}_p^-(Y) \quad \text{and} \quad \sum_{j=1}^{c} \hat{M}_p^+(X_j) = \hat{M}_p^+(Y)$$

for every $p \in [0, 1]$ (as for the corresponding population quantities), it follows that the relations

$$\sum_{j=1}^{c} \hat{\mathcal{R}}_p(X_j) = \hat{R}_p, \quad \sum_{j=1}^{c} \hat{\mathcal{R}}'(X_j) = \hat{R}', \quad \sum_{j=1}^{c} \hat{\mathcal{B}}'(X_j) = \hat{B}'$$

and

$$\sum_{j=1}^{c} \hat{\mathcal{I}}_p(X_j) = \hat{I}_p, \quad \sum_{j=1}^{c} \hat{\mathcal{I}}(X_j) = \hat{I}.$$

hold true for the estimators as well.

To estimate the relative contributions to the point inequality measures we can use the values taken on by the step functions

$$\hat{\rho}_p(X_j) := \frac{\hat{\mathcal{R}}_p(X_j)}{\hat{R}_p} = \frac{\hat{M}_p^-(X_j) - \hat{M}(X_j)}{\hat{M}_p^-(Y) - \hat{M}(Y)} \tag{61}$$

and

$$\hat{\zeta}_p(X_j) := \frac{\hat{\mathcal{I}}_p(X_j)}{\hat{I}_p} = \frac{\hat{M}_p^+(X_j) - \hat{M}_p^-(X_j)}{\hat{M}_p^+(Y) - \hat{M}_p^-(Y)} \tag{62}$$

at $p = i/N$ for $i = 1, 2, \ldots, N$. Note that, as for the corresponding population quantities, $\hat{\rho}_p(X_j)$ is not defined for $p \in (\hat{N}_{\hat{k}-1}/N, 1]$, and that for $p \in [0, \hat{N}_{\hat{k}-1}/N]$

$$\hat{\rho}_p(X_j) = \hat{\zeta}_p(X_j), \quad j = 1, 2, \ldots, c,$$

since an obvious generalization of relation (20) holds for weighted means as well. Taking appropriate averages finally yields

$$\hat{\rho}'(X_j) := \frac{\hat{\mathcal{R}}'(X_j)}{\hat{R}'} = \frac{\sum_{s=1}^{\hat{k}} \hat{\rho}_{\hat{N}_s/N}(X_j)\, \hat{R}_{\hat{N}_s/N}\, \hat{r}_s}{\sum_{s=1}^{\hat{k}} \hat{R}_{\hat{N}_s/N}\, \hat{r}_s}, \tag{63}$$

$$\hat{\beta}'(X_j) := \frac{\hat{\mathcal{B}}'(X_j)}{\hat{B}} = \frac{\sum_{s=1}^{\hat{k}} \hat{\rho}_{\hat{N}_s/N}(X_j)\, \hat{R}_{\hat{N}_s/N}\, \hat{n}_s}{\sum_{s=1}^{\hat{k}} \hat{R}_{\hat{N}_s/N}\, \hat{n}_s} \tag{64}$$

and

$$\hat{\zeta}(X_j) := \frac{\hat{\mathcal{R}}(X_j)}{\hat{R}} = \frac{\sum_{s=1}^{\hat{k}} \hat{\rho}_{\hat{N}_s/N}(X_j)\, \hat{R}_{\hat{N}_s/N}\, \hat{n}_s}{\sum_{s=1}^{\hat{k}} \hat{R}_{\hat{N}_s/N}\, \hat{n}_s} \tag{65}$$

as estimators of $\rho'(X_j)$, $\beta'(X_j)$ and $\zeta(X_j)$. Again, under suitable conditions, the former two estimators can be used to estimate $\rho(X_j)$ and $\beta(X_j)$ as well.

Finally, to estimate the shares $\gamma(X_j)$, one can simply use

$$\hat{\gamma}(X_j) := \frac{\sum_{i \in S} x_{i,j}\, w_i}{\sum_{i \in S} w_i} = \frac{\sum_{s=1}^{\hat{k}} \tilde{x}_{s,j}\, \hat{n}_s}{\sum_{s=1}^{\hat{k}} \hat{n}_s}, \quad j = 1, 2, \ldots, c. \tag{66}$$

It is not difficult to check that the relations between the relative contributions and the shares outlined in Section 4 hold for the estimates obtained from the estimators defined in the present section as well.

6. Application to ECHP Income Data

The European Community Household Panel (ECHP) is a multi-purpose annual longitudinal survey covering the time span between 1994 and 2001. Its aim is to provide comparable information from EU countries. It is centrally designed and coordinated by Eurostat and covers topics such as demographics, labor force behavior, income, health, education and training, housing, migration, and so on. The objective of the ECHP is to represent the population of the EU at individual and household level. More information about this survey may be found in the accompanying documentation (see Eurostat 1996; Eurostat 2003a; Eurostat 2003b; Eurostat 2002; Eurostat 2003c; Eurostat 2003d; Eurostat 2003e).

In the present work we analyze data about household income from the Users' Database (UDB) referring to the 2001 wave of the ECHP. Information on income is collected very detailed in the ECHP questionnaire. Some of the income components are collected at household level, while others are collected for each individual in sample households. In order to have complete information at both household and individual level, household income components are shared among its members aged over 16, and personal income components are aggregated for the whole household. To be specific, income components collected at household level are: property and rental income, social assistance and housing allowances. All other income components are collected individually among persons aged over 16 who reside in sample households. As for taxes, some of the income components are collected net and others gross of taxes. To allow for the computation of comparable net values, the survey provides net/gross ratios for each household (variable *HI020* in the Household-file of the UDB; except for the country-specific informations provided in Table 2, all other variables listed in this work are included in the Household-file).

Below we shall apply the estimators of Section 5 to evaluate the contributions from several income components to inequality in the distribution of total net household income (variable *HI100*). To avoid excessive scattering of the contributions among a large number of income components, we shall aggregate the latter into four main components:

- ***Wage and salary income*** ($X_1 :=$ variable *HI111*). This income component includes wages and salary payments and any other form of pay for work as an employee or apprentice.
- ***Self-employment income*** ($X_2 :=$ variable *HI112*). This includes any income from self-employment such as own business, professional practice or farm, working as free-lance or subcontractor, providing services or selling goods on own account.
- ***Other income components*** ($X_3 :=$ the sum of variables *HI121*, *HI122*, *HI123* and *HI140*). This includes capital income (variable *HI121*), income from property and rents (variable *HI122*), private transfers (variable *HI123*) and adjustments for within household non-response (variable *HI140*).
- ***Social transfers*** ($X_4 :=$ variable *HI130*). This includes unemployment related benefits, pension or benefit relating to old-age or retirement, survivor's pension or benefits for widows or orphans, family related benefits, benefits relating to sickness or invalidity, education related allowances and any other social benefits.

Except for the samples from France and Finland, the variables *HIxxx* in the UDB contain amounts of income net of taxes. For households where these variables are filled (the

variables referring to the income components are always filled if the net household income variable *HI100* is filled; however, for all countries, except Luxembourg, there are a few households where the value of the net household income variable is missing), the reported net values are consistent in the sense that

$$\begin{aligned}&\text{net household income } (Y := HI100) :=\\ &:= \textit{wage and salary income } (X_1 := HI111) +\\ &\quad + \textit{self employment income } (X_2 := HI112) +\\ &\quad + \textit{other income components}\\ &\qquad (X_3 := HI121 + HI122 + HI123 + HI140) +\\ &\quad + \textit{social transfers income } (X_4 := HI130).\end{aligned}$$

For households belonging to the samples from France and Finland, the variables *HI111*, *HI112*, *HI130*, *HI121*, *HI122*, and *HI123* report gross values, which must be converted into net values through multiplication by variable *HI020* (the household net/gross ratio), while all other variables *HIxxx* still contain net values. Thus, for the households included in the samples from France and Finland,

$$\begin{aligned}&\text{net household income } (Y := HI100) :=\\ &:= \textit{wage and salary income } (X_1 := HI111) +\\ &\quad + \textit{self employment income } (X_2 := HI020 \times HI112) +\\ &\quad + \textit{other income components}\\ &\quad (X_3 := HI020 \times (HI121 + HI122 + HI123) + HI140)\\ &\quad + \textit{social transfers income } (X_4 := HI020 \times HI130).\end{aligned}$$

Finally, as for the sample weights w_i, we shall follow a suggestion given in Eurostat (2003a) and use the cross-sectional household weights provided in the Household-file of the UDB (variable *HG004*). In fact, in the ECHP each household with completed household interview has its own nonnegative cross-sectional household weight *HG004*, and these weights are scaled to make sure that their sum over all interviewed households in each country equals the number d^* of interviewed households within the country. However, since for all countries except Luxembourg there are some sample households for which the net household income variable $Y := HI100$ is not filled, the final samples S we shall use for estimation comprise $d \leq d^*$ households. Table 1 reports the values of d^*, d and the relative weight θ of the sample households for which the total net household income *HI100* is missing (i.e., θ is the ratio between the sum of the cross-sectional household weights for sample households where the total net income variable *HI100* is missing and d^*). Note that there is no country for which θ exceeds two percent.

Now, consider Table 2. For each of the 15 countries included in the ECHP, Table 2 reports the population size, the number of households and the average household size as from the Country-file included in the UDB provided by Eurostat. Besides this general informations, Table 2 reports also the final sample sizes d used for estimation and some estimates regarding the distribution of net household income Y. The estimates for the

Table 1. Sample sizes in the 2001 wave of the ECHP.

Country	d^*	d	$(d^* - d)/d^*$	θ
Ireland	1,760	1,757	0.002	0.001
Denmark	2,283	2,279	0.002	0.001
Belgium	2,362	2,342	0.008	0.010
Luxembourg	2,428	2,428	0.000	0.000
Austria	2,544	2,535	0.004	0.002
Finland	3,115	3,106	0.003	0.002
Greece	3,916	3,895	0.005	0.006
Portugal	4,614	4,588	0.006	0.005
UK	4,819	4,779	0.008	0.009
Netherlands	4,851	4,824	0.006	0.005
Spain	4,966	4,950	0.003	0.003
Sweden	5,680	5,085	0.105	0.020
France	5,345	5,247	0.018	0.015
Italy	5,606	5,525	0.014	0.012
Germany	5,563	5,559	0.001	0.003

Legend: d^* is the number of interviewed households which coincides with the sum of the cross-sectional household weights *HG*004; d is the number of households used for estimation of the inequality indexes and the contributions to inequality from the four factor components, that is, number of households for which the net household income variable $Y := HI100$ is filled; θ is the ratio between the sum of the cross-sectional household weights for which Y is not filled and d^*.

median were obtained from the estimator

$$\widehat{Median}(Y) := \tilde{y}_{s^*},$$

where, in the notation of Section 5, s^* is the smallest integer s, $1 \leq s \leq \hat{k}$, such that $\hat{N}_{s^*}/N \geq 0.5$. Observe that the countries in Table 2 are ordered according to the estimates $\hat{R}'$ of the Gini index.

Next, consider the contributions in Table 3:

- ***Wage and salary income***, with shares $\hat{\gamma}(X_1)$ between 0.482 in Greece and 0.680 in Denmark, accounts for the largest share on total population income Y in all 15 countries. To understand how this factor component affects inequality, we first observe that the contributions $\hat{\rho}'(X_1)$, $\hat{\beta}'(X_1)$ and $\hat{\zeta}(X_1)$ are clearly larger than $\hat{\gamma}(X_1)$ which suggests that wage and salary income tends to be more concentrated among high income households than total income Y itself.

  To assess the impact on inequality at different levels p of the income distribution, we shall next examine the relative contributions $\hat{\rho}_p(X_1)$: we find that $\hat{\rho}_p(X_1) > \hat{\gamma}(X_1)$ for all countries for all values of p reported in Table 3, and that the trend of $\hat{\rho}_p(X_1)$ is quite similar in all countries: $\hat{\rho}_p(X_1)$ tends to increase for $0 < p \leq 0.25$ and to decrease for $p > 0.75$. For the interpretation of the relative contributions, recall that $\hat{\rho}_p(X_1)$ is the ratio between $M_p^+(X_1) - M_p^-(X_1)$ and $M_p^+(Y) - M_p^-(Y)$. In Italy, for example, $\hat{\rho}_{0.50}(X_1) = 0.661$ indicates that the difference between the means of wage and salary income among the households belonging to the upper half of the income distribution an those belonging to the lower half is equal to 0.661 times the difference between the corresponding means of total income Y.

Table 2. General information about countries included in the 2001 wave of the ECHP.

Country	Population size (in milions)	Number of households (in milions)	Average household size	Sample size d	$\widehat{Median}(Y)$	$\hat{M}(Y)$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
Denmark	5.368	2.456	2.19	2279	33561	34597	0.302	0.435	0.646
Netherlands	15.773	6.889	2.29	4824	22331	24788	0.303	0.428	0.643
Luxembourg	0.433	0.172	2.52	2428	38333	44729	0.304	0.414	0.631
Austria	7.986	3.3	2.42	2535	25058	28543	0.328	0.456	0.672
France	57.949	24.523	2.36	5247	24408	28053	0.329	0.457	0.674
Sweden	8.663	4.576	1.89	5085	20389	23651	0.331	0.459	0.677
Germany	81.569	37.711	2.16	5559	24554	28486	0.336	0.460	0.679
Italy	57.388	21.967	2.61	5525	18179	21210	0.342	0.471	0.688
Finland	5.12	2.381	2.15	3106	21067	24801	0.350	0.481	0.697
Belgium	10.263	4.278	2.4	2342	25558	30374	0.354	0.472	0.694
United Kingdom	59.063	25.564	2.31	4779	26893	32151	0.369	0.499	0.717
Greece	10.354	3.993	2.59	3895	12208	14853	0.382	0.517	0.734
Ireland	3.839	1.291	2.97	1757	25457	30685	0.388	0.524	0.740
Spain	39.137	13.281	2.95	4950	16810	21453	0.399	0.526	0.745
Portugal	10.024	3.391	2.96	4588	12362	15661	0.402	0.530	0.749

The estimates for the median and the mean of net household income Y are expressed in euros. They have been obtained using the fixed conversion rates for Germany, Denmark, Netherlands, Luxembourg, France, UK, Ireland, Italy, Greece, Spain, Portugal, and Austria and using the conversion rates for the year 2001 as given in the Country-file of the ECHP for Belgium, Finland, and Sweden.

Table 3. Contributions to inequality from income factor components.

					Austria					
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.622	0.740	0.802	0.834	0.857	0.817	0.811	0.761	0.834	0.818	0.818
X_2	0.061	0.063	0.075	0.086	0.094	0.118	0.111	0.132	0.097	0.087	0.095
X_3	0.032	0.006	0.020	0.025	0.031	0.046	0.056	0.056	0.036	0.028	0.033
X_4	0.284	0.191	0.103	0.055	0.018	0.019	0.022	0.050	0.033	0.066	0.053
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}_p$	0.810	0.751	0.630	0.460	0.280	0.147	0.091	0.328	0.456	–
	$\hat{I}_p$	0.817	0.770	0.694	0.630	0.609	0.633	0.664	–	–	0.672
					Belgium					
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.550	0.672	0.724	0.776	0.787	0.696	0.544	0.431	0.720	0.731	0.692
X_2	0.071	0.081	0.086	0.103	0.130	0.172	0.230	0.301	0.149	0.122	0.152
X_3	0.108	0.129	0.135	0.139	0.150	0.172	0.219	0.260	0.163	0.149	0.166
X_4	0.271	0.118	0.055	−0.019	−0.067	−0.040	0.007	0.008	−0.032	−0.001	−0.009
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}_p$	0.787	0.733	0.635	0.488	0.310	0.182	0.126	0.354	0.472	–
	$\hat{I}_p$	0.795	0.752	0.699	0.656	0.643	0.689	0.742	–	–	0.694
					Denmark					
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.680	0.762	0.822	0.928	1.018	0.961	0.773	0.675	0.945	0.908	0.890
X_2	0.051	0.061	0.066	0.079	0.095	0.110	0.175	0.217	0.102	0.086	0.108
X_3	0.043	0.039	0.038	0.040	0.036	0.038	0.067	0.064	0.039	0.038	0.042
X_4	0.226	0.138	0.073	−0.047	−0.149	−0.109	−0.014	0.044	−0.086	−0.032	−0.040
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}_p$	0.816	0.748	0.619	0.437	0.240	0.120	0.075	0.302	0.435	–
	$\hat{I}_p$	0.823	0.768	0.684	0.608	0.558	0.577	0.619	–	–	0.646

Table 3. Continued.

	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
					Finland						
X_1	0.612	0.676	0.725	0.808	0.814	0.809	0.707	0.608	0.791	0.772	0.758
X_2	0.070	0.078	0.084	0.094	0.110	0.137	0.172	0.211	0.119	0.103	0.117
X_3	0.053	0.053	0.060	0.065	0.075	0.099	0.161	0.231	0.090	0.075	0.098
X_4	0.266	0.193	0.132	0.034	0.000	−0.045	−0.040	−0.051	0.001	0.050	0.027
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}_p$	0.821	0.763	0.666	0.500	0.298	0.157	0.098	0.350	0.481	–
	$\hat{I}_p$	0.829	0.781	0.727	0.666	0.629	0.649	0.685	–	–	0.697
					France						
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.604	0.695	0.748	0.780	0.830	0.815	0.777	0.744	0.803	0.777	0.779
X_2	0.063	0.069	0.072	0.083	0.090	0.111	0.148	0.182	0.100	0.087	0.102
X_3	0.044	0.030	0.031	0.036	0.036	0.041	0.036	0.038	0.038	0.036	0.037
X_4	0.290	0.207	0.148	0.101	0.045	0.033	0.0390	0.036	0.060	0.100	0.082
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}'_p$	0.823	0.749	0.623	0.458	0.284	0.153	0.096	0.329	0.457	–
	$\hat{I}_p$	0.831	0.768	0.688	0.628	0.614	0.643	0.679	–	–	0.674
					Germany						
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.529	0.622	0.658	0.687	0.712	0.648	0.555	0.454	0.667	0.667	0.641
X_2	0.090	0.109	0.120	0.128	0.161	0.212	0.268	0.327	0.178	0.151	0.183
X_3	0.076	0.065	0.076	0.091	0.096	0.128	0.129	0.133	0.108	0.095	0.105
X_4	0.305	0.204	0.146	0.094	0.031	0.012	0.048	0.085	0.047	0.088	0.071
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}_p$	0.797	0.738	0.628	0.467	0.292	0.162	0.105	0.336	0.460	–
	$\hat{I}_p$	0.805	0.758	0.692	0.637	0.622	0.658	0.698	–	–	0.679

Table 3. Continued.

	Greece										
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.482	0.539	0.569	0.649	0.681	0.654	0.588	0.587	0.647	0.627	0.621
X_2	0.210	0.226	0.230	0.250	0.220	0.217	0.251	0.280	0.233	0.233	0.239
X_3	0.058	0.054	0.054	0.052	0.065	0.074	0.100	0.098	0.070	0.062	0.071
X_4	0.250	0.181	0.147	0.050	0.034	0.054	0.061	0.035	0.050	0.079	0.069
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}_p$	0.878	0.821	0.706	0.532	0.336	0.182	0.116	0.382	0.517	–
	$\hat{I}_p$	0.884	0.836	0.762	0.695	0.669	0.689	0.723	–	–	0.734
	Ireland										
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.615	0.720	0.732	0.784	0.755	0.691	0.611	0.588	0.732	0.742	0.715
X_2	0.137	0.153	0.162	0.167	0.200	0.240	0.311	0.361	0.214	0.189	0.219
X_3	0.059	0.063	0.067	0.067	0.072	0.071	0.075	0.056	0.068	0.067	0.066
X_4	0.190	0.064	0.039	−0.019	−0.028	−0.002	0.003	−0.004	−0.014	0.002	0.000
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}'_p$	0.847	0.820	0.734	0.548	0.336	0.178	0.111	0.388	0.524	–
	$\hat{I}_p$	0.853	0.835	0.786	0.708	0.669	0.683	0.713	–	–	0.740
	Italy										
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.505	0.586	0.617	0.688	0.661	0.631	0.542	0.479	0.642	0.639	0.621
X_2	0.162	0.182	0.189	0.197	0.220	0.229	0.306	0.347	0.226	0.208	0.230
X_3	0.041	0.022	0.029	0.040	0.053	0.063	0.080	0.098	0.055	0.045	0.054
X_4	0.293	0.211	0.166	0.075	0.067	0.077	0.072	0.077	0.077	0.108	0.095
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}_p$	0.831	0.769	0.644	0.480	0.297	0.159	0.098	0.342	0.471	–
	$\hat{I}_p$	0.838	0.787	0.707	0.648	0.626	0.653	0.685	–	–	0.689

Table 3. Continued.

	Luxembourg										
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.635	0.733	0.771	0.859	0.882	0.847	0.737	0.674	0.843	0.825	0.812
X_2	0.042	0.054	0.060	0.065	0.078	0.115	0.143	0.155	0.092	0.077	0.091
X_3	0.050	0.063	0.062	0.064	0.071	0.085	0.125	0.169	0.079	0.071	0.085
X_4	0.273	0.151	0.107	0.012	−0.032	−0.047	−0.005	0.002	−0.015	0.028	0.012
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}'_p$	0.714	0.662	0.563	0.423	0.269	0.144	0.089	0.304	0.414	–
	$\hat{I}_p$	0.725	0.685	0.632	0.594	0.596	0.626	0.658	–	–	0.631
	Netherlands										
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.635	0.693	0.788	0.871	0.883	0.839	0.804	0.743	0.856	0.831	0.827
X_2	0.038	0.045	0.046	0.057	0.074	0.098	0.118	0.160	0.080	0.065	0.080
X_3	0.058	0.034	0.049	0.068	0.064	0.060	0.048	0.052	0.062	0.059	0.059
X_4	0.269	0.228	0.117	0.004	−0.020	0.003	0.031	0.045	0.002	0.045	0.034
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}_p$	0.804	0.714	0.589	0.423	0.253	0.136	0.088	0.303	0.428	–
	$\hat{I}_p$	0.812	0.735	0.656	0.595	0.576	0.611	0.657	–	–	0.643
	Portugal										
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.629	0.716	0.750	0.778	0.747	0.745	0.747	0.714	0.755	0.754	0.746
X_2	0.124	0.131	0.141	0.148	0.148	0.121	0.088	0.108	0.133	0.137	0.132
X_3	0.034	0.035	0.037	0.039	0.048	0.059	0.071	0.093	0.052	0.045	0.055
X_4	0.214	0.117	0.073	0.035	0.057	0.075	0.094	0.085	0.061	0.065	0.067
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}'_p$	0.869	0.821	0.710	0.544	0.364	0.216	0.142	0.402	0.530	–
	$\hat{I}_p$	0.874	0.836	0.765	0.705	0.696	0.733	0.768	–	–	0.749

Table 3. Continued.

	Spain										
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.573	0.634	0.679	0.715	0.686	0.625	0.585	0.491	0.655	0.668	0.643
X_2	0.145	0.159	0.173	0.181	0.196	0.222	0.255	0.319	0.210	0.192	0.212
X_3	0.061	0.058	0.065	0.068	0.081	0.104	0.142	0.162	0.093	0.080	0.091
X_4	0.221	0.149	0.084	0.036	0.037	0.049	0.018	0.028	0.042	0.060	0.053
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{G}_p$	0.857	0.803	0.703	0.543	0.360	0.210	0.144	0.399	0.526	–
	$\hat{I}_p$	0.863	0.819	0.759	0.703	0.692	0.727	0.766	–	–	0.745
	Sweden										
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.609	0.658	0.718	0.839	0.829	0.900	0.865	0.827	0.845	0.803	0.817
X_2	0.018	0.003	0.000	0.013	0.013	0.007	0.016	0.024	0.011	0.010	0.012
X_3	0.050	0.055	0.058	0.070	0.083	0.105	0.139	0.156	0.092	0.077	0.091
X_4	0.323	0.284	0.224	0.078	0.075	−0.012	−0.020	−0.006	0.052	0.110	0.080
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}_p$	0.835	0.753	0.624	0.465	0.282	0.154	0.100	0.331	0.459	–
	$\hat{I}_p$	0.842	0.772	0.688	0.635	0.611	0.645	0.689	–	–	0.677
	United Kingdom										
	$\hat{\gamma}(\cdot)$	$\hat{\rho}_{0.05}(\cdot)$	$\hat{\rho}_{0.10}(\cdot)$	$\hat{\rho}_{0.25}(\cdot)$	$\hat{\rho}_{0.50}(\cdot)$	$\hat{\rho}_{0.75}(\cdot)$	$\hat{\rho}_{0.90}(\cdot)$	$\hat{\rho}_{0.95}(\cdot)$	$\hat{\rho}'(\cdot)$	$\hat{\beta}'(\cdot)$	$\zeta(\cdot)$
X_1	0.556	0.633	0.668	0.737	0.787	0.734	0.670	0.588	0.741	0.722	0.709
X_2	0.076	0.085	0.091	0.104	0.117	0.145	0.161	0.193	0.127	0.112	0.130
X_3	0.132	0.138	0.145	0.162	0.178	0.187	0.222	0.244	0.181	0.167	0.178
X_4	0.236	0.144	0.096	−0.003	−0.082	−0.066	−0.053	−0.025	−0.049	−0.001	−0.017
		$p = 0.05$	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	$p = 0.95$	$\hat{R}'$	$\hat{B}'$	$\hat{I}$
	$\hat{R}_p$	0.838	0.784	0.680	0.514	0.322	0.178	0.115	0.369	0.499	–
	$\hat{I}_p$	0.844	0.802	0.739	0.679	0.655	0.684	0.721	–	–	0.717

- ***Self-employment income***. The share $\hat{\gamma}(X_2)$ of self-employment income on total population income may vary a lot from country to country. In fact, it ranges from $\hat{\gamma}(X_2) = 0.018$ in Sweden to $\hat{\gamma}(X_2) = 0.210$ in Greece. Apart from Greece, the group of countries with large shares $\hat{\gamma}(X_2)$ includes Italy ($\hat{\gamma}(X_2) = 0.162$), Spain ($\hat{\gamma}(X_2) = 0.145$), Ireland ($\hat{\gamma}(X_2) = 0.137$) and Portugal ($\hat{\gamma}(X_2) = 0.124$). The contributions $\hat{\rho}'(X_2)$, $\hat{\beta}'(X_2)$ and $\hat{\zeta}(X_2)$ do clearly exceed $\hat{\gamma}(X_2)$ in all countries except for Sweden, indicating that also this factor component tends to be more concentrated among high income households than total income Y. The relative contributions $\hat{\rho}_p(X_2)$ are, except for Sweden, clearly larger than $\hat{\gamma}(X_2)$ at all levels of p reported in Table 3, and they tend to increase as p gets larger. In many countries the increasing trend is quite marked starting from $p = 0.5$.
- ***Other income components***. The share of income from this component is about $\hat{\gamma}(X_3) = 0.050$ in all countries except for Belgium and the United Kingdom, where $\hat{\gamma}(X_3) = 0.108$ and $\hat{\gamma}(X_3) = 0.132$, respectively. The contributions $\hat{\rho}'(X_3)$, $\hat{\beta}'(X_3)$, and $\hat{\zeta}(X_3)$ do slightly exceed $\hat{\gamma}(X_3)$ in most countries, indicating that, like for the former two factor components, the distribution of the other income components X_3 tends to exacerbate inequality in total income Y as well. The largest contributions $\hat{\rho}'(X_3)$, $\hat{\beta}'(X_3)$, and $\hat{\zeta}(X_3)$ are observed in those countries where the share $\hat{\gamma}(X_3)$ is also largest, that is, Belgium and the United Kingdom. Inspection of the relative contributions $\hat{\rho}_p(X_3)$ reveals an increasing trend in most countries. In some countries like Belgium, Finland, Sweden, and the United Kingdom the increasing trend is quite marked in the final part of the income distribution (i.e., for $p \geq 0.75$).
- ***Social transfers***, with shares $\hat{\gamma}(X_4)$ between 0.190 in Ireland, and 0.323 in Sweden, is the second largest factor component in all considered countries. As expected, the relative contributions $\hat{\rho}'(X_4)$, $\hat{\beta}'(X_4)$, and $\hat{\zeta}(X_4)$ are clearly smaller than $\hat{\gamma}(X_4)$, confirming that the distribution of this income component has an offsetting impact on inequality. In Belgium, Denmark, Ireland, Luxembourg, and the United Kingdom some of the relative contributions $\hat{\rho}'(X_4)$, $\hat{\beta}'(X_4)$, and/or $\hat{\zeta}(X_4)$ are even negative. As for the relative contributions $\hat{\rho}_p(X_4)$, they are for all countries smaller than $\hat{\gamma}(X_4)$ at all levels of p reported in Table 3, and they exhibit a decreasing trend in the initial part of the income distribution up to $p = 0.50$, and are thereafter almost constant, except for Sweden, where the decreasing trend holds on up to $p = 0.75$, and for Denmark, where $\hat{\rho}_p(X_4)$ increases after $p = 0.500$.

Appendix

In this appendix we prove that R' as defined in (27) and (28) is the ratio between the concentration area (i.e., the area between the Lorenz curve and the straight line which joins the origin (0, 0) with the point (1, 1)) and the area of the triangle with vertices in (0, 0), (1, 0) and (1, 1).

So let $P_s := p_{N_s}$ and $Q_s := q_{N_s}$, $s = 1, 2, \ldots, k-1$, be the abscissa and ordinate values of the points at which the slope of the Lorenz curve changes. It is not difficult to see that the conentration area is given by the sum of

- the area of the triangle with vertices in (0, 0), (P_1, P_1) and (P_1, Q_1), which is given by

$$\begin{aligned} A_1 &= \frac{(P_1 - Q_1)\,P_1}{2} \\ &= R_{N_1}\,\frac{P_1^2}{2} \end{aligned}$$

- the sum of areas of the $k-2$ trapezoids with vertices in (P_{s-1}, Q_{s-1}), (P_{s-1}, P_{s-1}), (P_s, Q_s) and (P_s, P_s), $s = 2, 3, \ldots, k-1$, which are given by

$$\begin{aligned} A_s &= \frac{[(P_{s-1} - Q_{s-1}) + (P_s - Q_s)]\,(P_s - P_{s-1})}{2} \\ &= R_{N_{s-1}}\,\frac{P_{s-1}\,(P_s - P_{s-1})}{2} + R_{N_s}\,\frac{P_s\,(P_s - P_{s-1})}{2} \end{aligned}$$

- the area of the triangle with vertices in (P_{k-1}, Q_{k-1}), (P_{k-1}, P_{k-1}) and (1, 1), which is given by

$$\begin{aligned} A_k &= \frac{(P_{k-1} - Q_{k-1})\,(1 - P_{k-1})}{2} \\ &= R_{N_{k-1}}\,\frac{P_{k-1}\,(1 - P_{k-1})}{2} \end{aligned}$$

Thus, the concentration area is given by

$$\begin{aligned} \sum_{s=1}^{k} A_s = R_{N_1}\,\frac{P_1^2}{2} + \sum_{s=2}^{k-1} R_{N_{s-1}}\,\frac{P_{s-1}\,(P_s - P_{s-1})}{2} + \\ + \sum_{s=2}^{k-1} R_{N_s}\,\frac{P_s\,(P_s - P_{s-1})}{2} + R_{N_{k-1}}\,\frac{P_{k-1}\,(1 - P_{k-1})}{2}. \end{aligned}$$

Setting $P_0 := 0$ and $P_k := 1$, the concentration area can also be written as

$$\sum_{s=1}^{k} A_s = \sum_{s=2}^{k} R_{N_{s-1}}\,\frac{P_{s-1}\,(P_s - P_{s-1})}{2} + \sum_{s=1}^{k-1} R_{N_s}\,\frac{P_s\,(P_s - P_{s-1})}{2}.$$

Using the fact that

$$\sum_{s=2}^{k} R_{N_{s-1}} \frac{P_{s-1}\,(P_s - P_{s-1})}{2} = \sum_{s=1}^{k-1} R_{N_s} \frac{P_s\,(P_{s+1} - P_s)}{2},$$

it is easily seen that

$$\sum_{s=1}^{k} A_s = \sum_{s=1}^{k-1} R_{N_s}\, r_s^*, \tag{67}$$

with

$$r_s^* := \frac{P_s(P_{s+1} - P_{s-1})}{2}, \quad s = 1, 2, \ldots, k-1,$$

Next, consider the hypothetical case where

$$Q_1 = Q_2 = \cdots = Q_{k-1} = 0.$$

In this case the concentration area would be given by the area of the triangle with vertices in (0, 0), (P_{k-1}, 0) and (1, 1), which is

$$\sum_{s=1}^{k} A_s = \frac{P_{k-1}}{2}, \tag{68}$$

and since we would have

$$R_{N_1} = R_{N_2} = \cdots = R_{N_{k-1}} = 1,$$

it follows from (67) and (68) that

$$\sum_{s=1}^{k-1} r_s^* = \frac{P_{k-1}}{2}.$$

Thus, if we set

$$r_k^* := \frac{1 - P_{k-1}}{2},$$

we get

$$\sum_{s=1}^{k} r_s^* = \frac{1}{2}, \tag{69}$$

and since $R_{N_k} = R_N = 0$ for every income distribution, it follows that the ratio between the concentration area and the area of triangle with vertices in (0, 0), (1, 0) and (1, 1) is given by (use (67) and (69))

$$2\sum_{s=1}^{k} A_s = 2\sum_{s=1}^{k} R_{N_s}\, r_s^* = \frac{\sum_{s=1}^{k} R_{N_s}\, r_s^*}{\sum_{s=1}^{k} r_s^*}.$$

Rescaling the weights r_s^* through multiplication by $2N^2$ yields finally the definition of R' in (27) and (28).

List of Notations

Symbol	Equation	Meaning
N	(1)	Number of population members
y_i for $i = 1, 2, \ldots, N$	(1)	Total incomes of the population members
Y	(1)	Symbol to indicate the total income variable
p_i for $i = 1, 2, \ldots, N$	(2)	Cumulative population shares
q_i for $i = 1, 2, \ldots, N$	(3)	Cumulative income shares
R_i for $i = 1, 2, \ldots, N$	(4)	Gini's point inequality measures
R	(5)	Gini's synthetic inequality index
$M_i^-(Y)$ for $i = 1, 2, \ldots, N$	(6)	Mean income of the i "poorest" population members, i.e., the i population members with smallest total income Y
$M(Y)$	(7)	Mean income of the whole population
B	(8)	Bonferroni's synthetic inequality index
I_i for $i = 1, 2, \ldots, N$	(9)	Zenga's point inequality indexes
k	(9)	Number of different values among $y_1, y_2, \ldots, y_N$
N_j for $j = 1, 2, \ldots, k$	(9)	Cumulative frequencies corresponding to different values among $y_1, y_2, \ldots, y_N$
$M_i^+(Y)$ for $i = 1, 2, \ldots, N$	(10)	Mean income of the $n - i$ "richest" population members, i.e., the $n - i$ population members with largest total income Y
I	(11)	Zenga's synthetic inequality index
n_j for $j = 1, 2, \ldots, k$	(11)	Absolute frequencies corresponding to different values among $y_1, y_2, \ldots, y_N$
c	(12)	Number of factor components
$x_{i,j}$ for $i = 1, 2, \ldots, N$ and for $j = 1, 2, \ldots, c$	(12)	Incomes from the c factor components
X_j for $j = 1, 2, \ldots, c$	(13)	Symbols to indicate factor components
$M(X_j)$ for $j = 1, 2, \ldots, c$	(13)	Population means of the factor components
$M_i^-(X_j)$ for $i = 1, 2, \ldots, N$ and for $j = 1, 2, \ldots, c$	(14)	Mean incomes from the factor components among the i "poorest" population members, that is, among the i population members with smallest total income Y
$M_i^+(X_j)$ for $i = 1, 2, \ldots, N$ and for $j = 1, 2, \ldots, c$	(15)	Mean incomes from the factor components among the $n - i$ "richest" population members, that is, among the $n - i$ population members with largest total income Y
$\mathcal{R}_i(X_j)$ for $i = 1, 2, \ldots, N$ and for $j = 1, 2, \ldots, c$	(16)	Contribution to the Gini point inequality index R_i from factor component X_j
$\mathcal{I}_i(X_j)$ for $i = 1, 2, \ldots, N$ and for $j = 1, 2, \ldots, c$	(17)	Contribution to the Zenga point inequality index I_i from factor component X_j
$\rho_i(X_j)$ for $i = 1, 2, \ldots, N$ and for $j = 1, 2, \ldots, c$	(18)	Relative contribution to the Gini point inequality index R_i from factor component X_j
$\zeta_i(X_j)$ for $i = 1, 2, \ldots, N$ and for $j = 1, 2, \ldots, c$	(19)	Relative contribution to the Zenga point inequality index I_i from factor component X_j
$\mathcal{R}(X_j)$ for $j = 1, 2, \ldots, c$	(21)	Contribution to Gini's synthetic inequality index R from factor component X_j
$\mathcal{B}(X_j)$ for $j = 1, 2, \ldots, c$	(22)	Contribution to Bonferroni's synthetic inequality index B from factor component X_j
$\mathcal{I}(X_j)$ for $j = 1, 2, \ldots, c$	(23)	Contribution to Zenga's synthetic inequality index I from factor component X_j

Symbol	Equation	Meaning
$\rho(X_j)$ for $j = 1, 2, \ldots, c$	(24)	Relative contribution to Gini's synthetic inequality index R from factor component X_j
$\beta(X_j)$ for $j = 1, 2, \ldots, c$	(25)	Relative contribution to Bonferroni's synthetic inequality index B from factor component X_j
$\zeta(X_j)$ for $j = 1, 2, \ldots, c$	(26)	Relative contribution to Zenga's synthetic inequality index I from factor component X_j
R'	(27)	Modified version of Gini's synthetic inequality index
r_s	(28)	Weights in Gini's synthetic inequality index
B'	(29)	Modified version of Bonferroni's synthetic inequality index
$\mathcal{R}'(X_j)$ for $j = 1, 2, \ldots, c$	(30)	Contribution to the modified version R' of Gini's synthetic inequality index from factor component X_j
$\mathcal{B}'(X_j)$ for $j = 1, 2, \ldots, c$	(31)	Contribution to the modified version B' of Bonferroni's synthetic inequality index from factor component X_j
$\rho'(X_j)$ for $j = 1, 2, \ldots, c$	(32)	Relative contribution from factor component X_j to the modified version of Gin's synthetic inequality index
$\beta'(X_j)$ for $j = 1, 2, \ldots, c$	(33)	Relative contribution from factor component X_j to the modified version of Bonferroni's synthetic inequality index
$\gamma(X_j)$ for $j = 1, 2, \ldots, c$	(34)	Share of factor component X_j on total population income
S	(39)	Set of indexes i identifying population units belonging to a sample
d	(39)	Sample size, i.e., number of indexes i in S. Note that in the application of Section 6 we considered for estimation only sample households for which the net household income variable $Y := HI100$ is filled. Thus, the samples S used for estimation do not comprise all interviewed households: in fact, for every country there are some interviewed households for which the net household income variable $Y := HI100$ is not filled (see Table 1).
w_i for $i \in S$	(40)	Survey weights corresponding to the sample units $i \in S$
$\hat{k}$	(42)	Number of sample units with different total income Y
$\tilde{y}_1 < \tilde{y}_2 < \cdots < \tilde{y}_{\hat{k}}$	(42)	Different values of total income Y among sample units
$\hat{n}_s$ for $s = 1, 2, \ldots, \hat{k}$	(43)	Sum of survey weights corresponding to the sample units with total income Y equal to $\tilde{y}_s$
$\hat{N}_s$ for $s = 1, 2, \ldots, \hat{k}$	(44)	Cumulative survey weights corresponding to different values of total income Y in the sample
$\sigma(p)$ for $p \in [0, 1]$	(45)	$\sigma(p) := \min\{s : \hat{N}_s \geq N p\}$, that is, number of different values $\tilde{y}_s$ of total income Y among sample units with total income not larger the pth sample quantile of total income
$\hat{M}_p^-(Y)$ for $p \in [0, 1]$	(46)	Weighted mean of total income Y among sample units with total income not larger than the pth sample quantile $\tilde{y}_{\sigma(p)}$
$\hat{M}_p^+(Y)$ for $p \in [0, 1]$	(47)	Weighted mean of total income Y among sample units with total income larger than the pth sample quantile $\tilde{y}_{\sigma(p)}$
$\hat{M}(Y)$ for $p \in [0, 1]$	(48)	Weighted sample mean of total income Y
$\hat{R}_p$ for $p \in [0, 1]$	(49)	Estimates for Gini's point inequality measures
$\hat{I}_p$ for $p \in [0, 1]$	(50)	Estimates for Zenga's point inequality measures

Symbol	Equation	Meaning
$\hat{r}_s$ for $s = 1, 2, \ldots, \hat{k}$	(51)	Estimates for the weights in the modified version of Gini's synthetic inequality index R'
$\hat{R}'$	(52)	Estimate for the modified version R' of Gini's synthetic inequality
$\hat{B}'$	(53)	Estimate for the modified version B' of Bonferroni's synthetic inequality
$\hat{I}$	(54)	Estimate for Zenga's synthetic inequality index I
$\tilde{x}_{s,j}$ for $s = 1, 2, \ldots, \hat{k}$ and for $j = 1, 2, \ldots, c$	(55)	Weighted average of income from factor component X_j among the sample units with total income equal to $\tilde{y}_s$
$\hat{\mathcal{R}}_p(X_j)$ for $p \in [0, 1]$ and for $j = 1, 2, \ldots, c$	(56)	Sample estimate for the contribution $\mathcal{R}_i(X_j)$ at $i = \lceil N p \rceil$
$\hat{\mathcal{I}}_p(X_j)$ for $p \in [0, 1]$ and for $j = 1, 2, \ldots, c$	(57)	Sample estimate for the contribution $\mathcal{I}_p(X_j)$ at $i = \lceil N p \rceil$
$\hat{\mathcal{R}}(X_j)$ for $j = 1, 2, \ldots, c$	(58)	Sample estimate for the contribution $\mathcal{R}'(X_j)$
$\hat{\mathcal{B}}(X_j)$ for $j = 1, 2, \ldots, c$	(59)	Sample estimate for the contribution $\mathcal{B}'(X_j)$
$\hat{\mathcal{I}}(X_j)$ for $j = 1, 2, \ldots, c$	(60)	Sample estimate for the contribution $\mathcal{I}(X_j)$
$\hat{\rho}_p(X_j)$ for $p \in [0, \hat{N}_{\hat{k}-1}/N]$ and for $j = 1, 2, \ldots, c$	(61)	Sample estimate for the relative contribution $\rho_i(X_j)$ at $i = \lceil N p \rceil$
$\hat{\zeta}_p(X_j)$ for $p \in [0, \hat{N}_{\hat{k}-1}/N]$ and for $j = 1, 2, \ldots, c$	(62)	Sample estimate for the relative contribution $\zeta_i(X_j)$ at $i = \lceil N p \rceil$
$\hat{\rho}(X_j)$ for $j = 1, 2, \ldots, c$	(63)	Sample estimate for the relative contribution $\rho'(X_j)$
$\hat{\beta}(X_j)$ for $j = 1, 2, \ldots, c$	(64)	Sample estimate for the relative contribution $\beta'(X_j)$
$\hat{\zeta}(X_j)$ for $j = 1, 2, \ldots, c$	(65)	Sample estimate for the relative contribution $\zeta(X_j)$
$\hat{\gamma}(X_j)$ for $j = 1, 2, \ldots, c$	(66)	Sample estimate for the share of factor component X_j on total population income

7. References

Antal, E., M. Langel, and Y. Tillé. 2011. "Variance Estimation of Inequality Indices in Complex Sample Designs." In Bulletin of the International Statistical Institute Proceedings of the 58th World Statistics Congress, 21st-26th August 2011, Dublin Convention Centre, 1036–1045. Available at: http://2011.isiproceedings.org/papers/450008.pdf (accessed April 2017). ISBN: 978-90-73592-33-9.

Bonferroni, C.E. 1930. *Elementi di Statistica Generale*. Firenze: Seeber.

DeVergottini, M. 1940. "Sul Signicato di Alcuni Indici di Concentrazione." *Giornale degli Economisti e Annuali di Economia* 2: 317–347.

Eurostat. 1996. *European Community Household Panel (ECHP): Volume 1 – Survey methodology and implementation*. Luxembourg: Office for Official Publications of the European Communities. ISBN: 92-827-8928-4.

Eurostat. 2002. *Imputation of Income in the ECHP, PAN 164/2002-12*. Eurostat.

Eurostat. 2003a. *ECHP UDB Manual: Waves 1 to 8 (Survey Years 1994 to 2001), PAN 168/2003-12*. Eurostat.

Eurostat. 2003b. *Anonymization Criteria Applied to the Users' Database, PAN105/2003-05*. Eurostat.

Eurostat. 2003c. *Construction of Weights in the ECHP, PAN 165/2003-06*. Eurostat.

Eurostat. 2003d. *Description of Variables: Data Dictionary, Codebook and Differences Between Countries and Waves, PAN 166/2003-12*. Eurostat.

Eurostat. 2003e. *Construction of Variables: from ECHP Questions to UDB Variables, PAN 167/2003-12*. Eurostat.

Gini, C. 1914. "Sulla Misura Della Concentrazione e Della Variabilità dei Caratteri." In *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti. Anno Accademico 1913-1914, Tomo LXXIII – Parte Seconda*.

Greselin, F., L. Pasquazzi, and R. Zitikis. 2013. "Contrasting the Gini and Zenga Indices of Economic Inequality." *Journal of Applied Statistics* 40(2) : 282–297. Doi: http://dx.doi.org/10.1080/02664763.2012.740627.

Greselin, F., L. Pasquazzi, and R. Zitikis. 2014. "Heavy Tailed Capital Incomes: Zenga Index, Statistical Inference, and ECHP Data Analysis." *Extremes* 17(1): 127–155. Doi: http://dx.doi.org/10.1007/s10687-013-0177-2.

Greselin, F. and L. Pasquazzi. 2009. "Asymptotic Confidence Intervals for a New Inequality Measure." *Communications in Statistics-Simulation and Computation* 38(8): 1742–1756. Doi: http://dx.doi.org/10.1080/03610910903121974.

Greselin, F., L. Pasquazzi, and R. Zitikis. 2010. "Zenga's New Index of Economic Inequality, its Estimation, and an Analysis of Incomes in Italy." *Journal of Probability and Statistics* (special issue on "Actuarial and Financial Risks: Models, Statistical Inference, and Case Studies") . Article ID 718905, 26 pages. Available at: http://www.emis.de/journals/HOA/JPS/Volume2010/718905.pdf (accessed April 2017).

Greselin, F., M. Puri, and R. Zitikis. 2009. "L-Functions, Processes, and Statistics in Measuring Economic Inequality and Actuarial Risks." *Statistics and Its Interface* 2(2): 227–245. Doi: http://dx.doi.org/10.4310/SII.2009.v2.n2.a13.

Langel, M. and Y. Tillé. 2012. "Inference by Linearization for Zenga's New Inequality Index: A Comparison with the Gini Index." *Metrika* 75(8): 1093–1110. Doi: http://dx.doi.org/10.1007/s00184-011-0369-1.

Lerman, R. and S. Yitzhaki. 1984. "A Note on the Calculation and Interpretation of the Gini Index." *Economics Letters* 15(3): 363–368. Doi: https://doi.org/10.1016/0165-1765(84)90126-5.

Lerman, R. and S. Yitzhaki. 1985. "Income Inequality Effects by Income Source: A New Approach and Applications to the United States." *Review of Economics and Statistics* 67(1): 151–156. Available at: https://www.researchgate.net/profile/Shlomo_Yitzhaki/publication/24094305_Income_Inequality_Effects_by_Income/links/02e7e5274ff3fce713000000.pdf (accessed April 2017).

Lorenz, M.O. 1905. "Methods of Measuring the Concentration of Wealth." *Publications of the American Statistical Association* 9(70): 209–219. Doi: http://dx.doi.org/10.2307/2276207.

Polisicchio, M. 2008. "The Continuous Random Variable with Uniform Point Inequality Measure I(p)." *Statistica & Applicazioni* 6(2): 137–151.

Polisicchio, M. and F. Porro. 2009. "A Comparison Between the Lorenz L(p) Curve and the Zenga I(p) Curve." *Italian Journal of Applied Statistics* 21(3–4): 289–301.

Porro, F. 2008. "Equivalence Between the Partial Order Based on L(p) Curve and Partial Order Based on I(p) Curve." In *Atti della XLIV Riunione Scientica*: Università della Calabria 25–27 giugno 2008. Sessione plenaria, Sessioni specializzate, Sessioni spontanee (cd), edited by CLEUP – Padova. ISBN: 9788861292284.

Porro, F. 2011. "The Distribution Model with Linear Inequality Curve I(p)." *Statistica & Applicazioni* 9(1): 47–61.

Radaelli, P. 2007. "A Subgroup Decomposition of a New Inequality Index Proposed by Zenga." In *Bulletin of the ISI 56th World Statistics Congress of the International Statistical Institute, 22nd-29th August 2007, Lisboa Congress Centre (CCL)*, 5151–5154. Available at: http://isi.cbs.nl/iamamember/CD7-Lisboa2007/Bulletin-of-the-ISI-Volume-LXII-2007.pdf (accessed April 2017), ISBN: 978-972-673-992-0.

Radaelli, P. 2008a. "A Subgroups Decomposition of Zenga's Uniformity and Inequality Indexes." *Statistica & Applicazioni* 6(2): 117–136.

Radaelli, P. 2008b. "Decomposition of Zenga's Inequality Measure by Subgroups." In *Atti della XLIV Riunione Scientica: Università della Calabria 25–27 giugno 2008. Sessione plenaria, Sessioni specializzate, Sessioni spontanee (cd)*, edited by CLEUP – Padova. isbn: 9788861292284.

Radaelli, P. 2010. "On the Decomposition by Subgroups of the Gini Index and Zenga's Uniformity Index and Inequality Indexes." *International Statistical Review* 78(1): 81–101. Doi: http://dx.doi.org/10.1111/j.1751-5823.2010.00100.x.

Radaelli, P. and M.M. Zenga. 2005. "On the Decomposition of the Gini's Mean Difference and Concentration Ratio." *Statistica & Applicazioni* 3(2): 5–24.

Rao, V. 1969. "Two Decompositions of Concentration Ratio." *Journal of the Royal Statistical Society;* Series A 132(3): 418–425. Doi: http://dx.doi.org/10.2307/2344120.

Shorrocks, A.F. 1982. "Inequality Decomposition by Factor Components." *Econometrica* 5(1): 193–212. Available at: http://www.ophi.org.uk/wp-content/uploads/ssShorrocks-1982.pdf (accessed April 2017), Doi: http://dx.doi.org/10.2307/1912537.

Shorrocks, A.F. 1983. "The Impact of Income Components on the Distribution of Family Incomes." *The Quarterly Journal of Economics* 98(2): 311–326. Doi: http://dx.doi.org/10.2307/1885627.

Zenga, M. 2008. "An Extension of Inequality *I* and *I(p)* Curve to Non-Economic Variables." In *Atti della XLIV Riunione Scientica: Università della Calabria 25–27 giugno 2008. Sessione plenaria, Sessioni specializzate, Sessioni spontanee* (*cd*), edited by CLEUP – Padova. ISBN: 9788861292284.

Zenga, M.M. 1984. "Proposta per un Indice di Concentrazione Basato sui Rapport fra Quantili di Popolazione e Quantili di Reddito." *Giornale Degli Economisti e Annali di Economia* 43(5–6): 301–326.

Zenga, M.M. 2007a. "Inequality Curve and Inequality Index Based on the Ratios Between Lower and Upper Arithmetic Means." *Statistica & Applicazioni* 5(1): 3–27.

Zenga, M.M. 2007b. "Applications of a New Inequality curve and Inequality Index Based on Ratios Between Lower and Upper Arithmetic Means." In Bulletin of the ISI 56th World Statistics Congress of the International Statistical Institute, 22nd-29th August 2007, Lisboa Congress Centre (CCL), 5167–5170. Available at: http://isi.cbs.nl/iamamember/CD7-Lisboa2007/Bulletin-of-the-ISI-Volume-LXII-2007.pdf (accessed April 2017). ISBN: 978-972-673-992-0.

Zenga, M.M. 2013. "Decomposition by Sources of the Gini, Bonferroni and Zenga Inequality Indexes." *Statistica & Applicazioni* 11(2): 133–161.

Zenga, M.M., P. Radaelli, and M. Zenga. 2012. "Decomposition of Zenga's Inequality Index by Sources." *Statistica & Applicazioni* 10(1): 3–31.

Received October 2014
Revised April 2017
Accepted September 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 181–209, http://dx.doi.org/10.1515/JOS-2018-0009

Using Social Network Information for Survey Estimation

Thomas Suesse[1] *and Ray Chambers*[1]

Model-based and model-assisted methods of survey estimation aim to improve the precision of estimators of the population total or mean relative to methods based on the nonparametric Horvitz-Thompson estimator. These methods often use a linear regression model defined in terms of auxiliary variables whose values are assumed known for all population units. Information on networks represents another form of auxiliary information that might increase the precision of these estimators, particularly if it is reasonable to assume that networked population units have similar values of the survey variable. Linear models that use networks as a source of auxiliary information include autocorrelation, disturbance, and contextual models. In this article we focus on social networks, and investigate how much of the population structure of the network needs to be known for estimation methods based on these models to be useful. In particular, we use simulation to compare the performance of the best linear unbiased predictor under a model that ignores the network with model-based estimators that incorporate network information. Our results show that incorporating network information via a contextual model seems to be the most appropriate approach. We also show that one does not need to know the full population network, but that knowledge of the partial network linking the sampled population units to the non-sampled population units is necessary. Finally, we also provide an estimator for the mean-squared error to make an informed decision about using the contextual information, as well as the results showing that this adaptive strategy leads to higher precision.

Key words: BLUP; social network models; linear models; model-based survey estimation.

1. Introduction

Survey estimation typically focuses on estimating the total $T_Y = \sum_{i \in U} Y_i$ of the values of a variable Y defined over a finite population U. Here $i \in U$ denotes the N units making up the population U. Given a sample s of n units from U, T_Y is usually estimated by $\hat{T}_Y = \sum_{i \in s} w_i Y_i$, where the w_i are sample weights and $i \in s$ denotes the n units in the sample. Traditionally, these weights are expansion weights, that is w_i is the inverse of the selection probability of the ith population unit. However, expansion weights can be quite inefficient, and alternative weighting methods derived from model-based and model-assisted methods of survey estimation, see Chambers and Clark (2012) and Särndal et al. (1992), are used to increase the precision of $\hat{T}_Y$. In most cases this is done by defining the sample weights so that $\hat{T}_Y$ is an efficient unbiased predictor of T_Y under a linear regression model for Y in terms of a multivariate auxiliary variable $\mathbf{X}$.

[1] National Institute for Applied Statistics Research Australia and University of Wollongong, Northfield Avenue, Wollongong, New South Wales 2522, Wollongong, Australia. Emails: tsuesse@uow.edu.au and ray@uow.edu.au
Acknowledgments: We would like to thank NIASRA for providing the High Performance Cluster to run the simulations and the referees for their helpful comments that greatly improved this paper. We would also like to acknowledge financial support from Australian Research Council (ARC) (Grant number LX0883143).

Population regression models that link an individual's value of Y to auxiliary variables corresponding to that individual's geographic location, gender, and age are commonly used in survey estimation. However, auxiliary information can be more complex than this. In particular, information about other individuals in the population that are 'linked' to a particular individual also constitutes auxiliary information about that individual. This is sometimes referred to as *network* information, and typically indicates between individual correlation in the population values of Y. In this article we describe model-based survey estimation methods that exploit auxiliary information about population networks. In particular, we describe how the specification of the Best Linear Unbiased Predictor (BLUP) of T_Y can be tailored to allow for between individual correlation induced by the presence of a population network. Such correlation or association between individuals with similar characteristics is often referred to as *homophily* in the network literature.

In order to motivate the use of network information in survey estimation, consider the case of the British Household Panel Study (BHPS, https://www.iser.essex.ac.uk/bhps/). This is an annual longitudinal survey of British households that has been conducted since 1991. It is based on a sample of approximately 5,500 households, covering more than 10,000 individuals. The main objective of the survey is to further the understanding of social and economic change at the individual and household level in Britain. However, in addition to information about the surveyed individual, the BHPS also provides information about a person's three closest friends. Variables collected on the three closest friends are: age, sex, ethnicity, distance to friend (<1 mile, between 1 and <5 miles, between 5 and 50 miles, >50 miles), and unemployment status. This information is available in seven waves, corresponding to the even-numbered years 1992–2004.

Because friends tend to share common characteristics, it is plausible that the BHPS information on friendship ties may be of value when modelling the other survey variables, in the same way as the ties between household members are typically viewed as influential in determining the outcomes of many social and economic variables. For example, a person whose friends are older than the norm might have a higher than average income, even after adjusting for that person's age and gender. As a consequence, one might think of also controlling for the average age of friends when predicting a person's income. A model of this type is referred to as a *contextual* model in what follows since it controls for contextual effects, such as the average age of friends. Clearly, since the BHPS collects information on a person's three best friends, there is scope for applying a contextual model when estimating using BHPS data. This might lead to more precise survey estimates, as a contextual variable represents an additional source of information.

The friendship data collected in the BHPS are a special case of a general type of auxiliary data whose availability is becoming increasingly widespread, especially with the rapid uptake of modern telecommunications technology. This is network data, defined by the existence, direction and strength of relationships between individuals in a population of interest. Statistical modelling of networks is now reasonably well established, see, for example Frank and Strauss (1986), Snijders (2002), Hunter and Handcock (2006), though applications to very large networks (e.g., defined by populations similar in size to those covered by a survey like the BHPS) are still rare, with data on very large networks now considered to be part of the ubiquitous Big Data concept. Furthermore, we are not aware of any attempt to use the information in a network defined on a population of interest to

improve survey estimation for that population, although, as the argument put forward in the previous paragraph indicates, there may be value in doing so.

In order to use network information in a model linking a survey variable Y to the auxiliary variable $\mathbf{X}$ we need to characterise the population network as the outcome of a random process. In this context, we focus in this article on a network that identifies the existence and direction of a relationship between individuals in a population of size N. It is standard to represent such a network by a matrix of zeros and ones, $\mathbf{Z} = (Z_{ij})_{i,j=1}^{N}$ with $Z_{ii} = 0$ by convention. If a relationship exists between two individuals i and j, then $Z_{ij} = 1$ and we refer to i and j as being linked; otherwise $Z_{ij} = 0$. Such a network is said to be undirected if $\mathbf{Z} = \mathbf{Z}^{\top}$, otherwise it is a directed network.

Networks are most useful when characteristics of the individuals that make up the population covered by the network are also known. In such networks one not only knows the characteristics of a particular individual, but also the characteristics of the other individuals in the population linked to that individual via the network. This *external* auxiliary information may be useful in discriminating between individuals, and hence may be useful in prediction, the ultimate goal of survey estimation. For example, the BHPS collects information about the three best friends of a surveyed individual, without identifying the friends. Given that the links corresponding to being 'one of three best friends' define a network, this information can be treated as auxiliary data for the surveyed individual, and, combined with a model for the network, may help with formulating a more efficient prediction model for the population.

Linear models that use a social network as additional information to model the expected value of a response variable include contextual network (CN) models (Friedkin 1990). However, this information can also be used to model between unit correlation in the population values of the response variable. Such second order models include network effects models, also known as autocorrelation (AR) models, and network disturbance (ND) models (Ord 1975; Doreian et al. 1984; Duke 1993; Marsden and Friedkin 1993; Leenders 2002).

When the network defined by $\mathbf{Z}$ is known for all N individuals in the population, the CN, AR and ND population models can be used for survey estimation. However, in practice it is extremely unlikely that $\mathbf{Z}$ will be fully known, and a more realistic scenario is one where one or more components of this matrix will be known. The most obvious is where only the component $\mathbf{Z}_{ss}$ corresponding to the sub-network of relationships between the n sampled individuals in s is known. Unless the sampling fraction is large, or the sample is highly clustered, it is unlikely that this sub-network will contain much useful information. Of more use, perhaps, is the component $\mathbf{Z}_{sr}$, defined by the links between the sampled individuals and the remaining $N - n$ non-sampled individuals in the population, denoted collectively by r. Clearly, if the network is an undirected one, the links from the non-sampled individuals to the sampled individuals will then also be known since, under symmetry, $\mathbf{Z}_{rs} = \mathbf{Z}_{sr}^{\top}$. The remaining component of $\mathbf{Z}$ is $\mathbf{Z}_{rr}$, which corresponds to the sub-network defined by the links between the $N - n$ non-sampled individuals in the population. This will generally be unknown. Using network information in a survey sampling context therefore implies that one has to deal with situations where partial network information is observed. This inevitably means that one needs to either use more complicated modelling methods or that one needs to somehow impute the missing network components.

The main focus of this article is on the potential use of network information in survey estimation. In particular, we aim to address three questions: (i) Is embedding network information useful for survey estimation based on linear models? (ii) If the answer to (i) is yes, then which network models are potentially useful? and (iii) How much network data needs to be collected in order to obtain potentially higher precision for survey estimation? In Section 2 we provide some context for these questions by defining a standard linear model that is often used for survey estimation. This linear model does not incorporate network information, so we then describe three widely used linear models that allow for the availability of network information in addition to standard covariate information.

In Section 3 we briefly discuss estimation of the population mean of a survey variable using the empirical version of the BLUP (typically referred to as the empirical best linear unbiased predictor or EBLUP) based on a linear model for this variable, and its application under the network models introduced in the previous Section. In Section 4, the Exponential Random Graph Model (ERGM) for a network is introduced and a simple imputation of missing network information is described, with the aim of using this imputed information in the network model-based estimators introduced in Chapter 3. These ideas are then brought together in Section 5 where we describe a simulation study that investigates the performances of the imputation-based EBLUPs defined by these different network models. In particular, we compare these estimators with the standard linear estimators that ignore network information. In Section 6 we use data from "wave N" (year 2004) of the BHPS to illustrate age by sex by region estimation of population means based on a model that includes age by sex effects and a contextual variable corresponding to the maleness proportion of an individual's three best friends. Section 7 completes the article with a discussion of our findings as they relate to the three questions raised above.

2. Linear Models on Networks

In this section we describe a number of population level linear models that use network information. Throughout, we use a friendship social network structure for simplicity of exposition. In order to develop our notation, the starting point is the linear model that assumes uncorrelated errors.

2.1. The Standard Model

The classical linear model for a population of N individuals can be written in matrix form as

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}), \tag{1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_N)^\top$ is a population vector of responses, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^\top$ with $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ is the model design matrix for the population with p columns defined by a set of covariates that depend on auxiliary population information, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N)^\top$ is the vector of population model residuals with $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2)$ and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is the vector of regression coefficients. The population mean vector and population covariance matrix of $\mathbf{Y}$ are then $\mu = \mathbf{X}\beta$ and $\mathbf{V} = \sigma^2\mathbf{I}_N$. Here $\mathbf{I}_N$ denotes the identity matrix of order N.

It is assumed that the matrix $\mathbf{X}$ defined by the auxiliary population information does not include variables related to social networks, so (1) does not use social network information.

Survey populations are often hierarchical and can be characterised as grouped into clusters, with each cluster j accounted for by a cluster-specific random effect u_j in the model. This can be modelled by a linear mixed model to incorporate dependence of units in the same cluster. See Chambers and Clark (2012, Chapter 6) for more details.

2.2. *The Contextual Network (CN) Model*

Consider an educational modelling exercise where Academic Performance (AP) is the response variable and Socioeconomic Status (SES) of the student is the explanatory variable. A classical contextual approach might then lead one to include the average SES of the student's school as another explanatory variable. Friedkin (1990) adapts this idea to network data by considering models where the response for a particular subject also depends on the characteristics of other subjects that are linked to the one of interest. In our example this would correspond to modelling AP in terms of both the student's SES as well as the SES values of the student's friends. Since a student will generally have several friends, a student's AP could then be modelled in terms of his/her SES as well as the average SES of his/her friends.

In general, such a CN model can be written in matrix form as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}\gamma + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{Y}$ and $\mathbf{X}$ have the same meaning as for Model (1), but the columns of $\mathbf{U}$ correspond to statistics derived from the variables that are measured on the network. In particular, the ith row of $\mathbf{U}$ contains appropriate summary characteristics of those other individuals on the network that are linked to individual i. Thus, in the preceding example, assuming that SES is the only covariate measured on the network, then $\mathbf{U}$ is the column vector of length N whose ith value is $\overline{SES}_i$, the average SES of all friends of student i. More generally, let $\tilde{\mathbf{X}}$ denote the population matrix of covariates measured on the network. The matrix $\tilde{\mathbf{X}}$ can be a subset of $\mathbf{X}$ but can also include other variables that are not in $\mathbf{X}$.

Then one way of defining $\mathbf{U}$ is via the identity

$$\mathbf{U} = \mathbf{W}\tilde{\mathbf{X}}, \tag{3}$$

where $\mathbf{W} = \mathbf{Z}/\mathbf{Z}\mathbf{1}_{N,N}$ is a row-normalised version of $\mathbf{Z}$, that is the rows of $\mathbf{W}$ sum to one. In general, $\mathbf{U} = g(\mathbf{Z}, \tilde{\mathbf{X}})$ is a function of the network $\mathbf{Z}$ and $\tilde{\mathbf{X}}$. A contextual variable for person i often includes the value for this person, for example a household contextual effect is computed over all household members including person i. However, the contextual value for person i defined by (3) excludes person i, because $Z_{ii} = 0$ by definition.

2.3. *The Autocorrelation (AR) Model*

The matrix $\tilde{\mathbf{X}}$ introduced in the preceding description of the CN model can be any set of measurements on the individuals in the network. In particular, it can be $\mathbf{Y}$. This leads to another class of models, called Autocorrelation (AR) models, and also known as network effects models, that incorporate network information into a linear structure. See, for

example, Doreian et al. (1984), Duke (1993), Marsden and Friedkin (1993), and Leenders (2002), and in the context of spatial models, Ord (1975). Under an AR model,

$$\mathbf{Y} = \theta\bar{\mathbf{Y}} + \mathbf{X}\beta + \boldsymbol{\epsilon}, \tag{4}$$

where $\bar{\mathbf{Y}} = (\bar{Y}_1, \ldots, \bar{Y}_N)^\top$ and $\bar{Y}_i$ is the average response of the individuals in the network that are linked to individual i, so $\bar{\mathbf{Y}} = \mathbf{WY}$, with $\mathbf{W}$ defined in the previous subsection. The conditional (on $\mathbf{X}$) mean and variance of $\mathbf{Y}$ under (4) are $\mu = \mathbf{D}^{-1}\mathbf{X}\beta$ and $\mathbf{V} = \sigma^2(\mathbf{D}^\top\mathbf{D})^{-1}$, where $\mathbf{D} = \mathbf{I}_N - \theta\mathbf{W}$. Note that $\mathbf{W}$ can be defined in a variety of ways, see Leenders (2002), though typically it is defined as the row-normalised version of $\mathbf{Z}$, that is, $\sum_{j=1}^{N} W_{ij} = 1$. The parameter θ is restricted $\theta \notin \{\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_1}\}$ as a necessary condition for $\mathbf{V}$ to exist, where λ_i are the eigenvalues of the row-normalised $\mathbf{W}$. Often θ is restricted to $(-1, 1)$.

In the context of the academic performance example introduced in the previous subsection we see that (4) implies that a student's AP score now depends on his/her SES value as well as the average AP scores of his/her friends.

2.4. *The Network Disturbance (ND) Model*

Models of this type have been considered by Ord (1975) and Leenders (2002) among others, and correspond to imposing an AR structure on the error term in the standard linear model (1). They are referred to as Network Disturbance (ND) models and are specified by

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \theta\bar{\boldsymbol{\epsilon}} + \mathbf{v}, \mathbf{v} \sim N(0, \sigma^2\mathbf{I}_N). \tag{5}$$

Here $\bar{\boldsymbol{\epsilon}} = (\bar{\epsilon}_1, \ldots, \bar{\epsilon}_N)$ where $\bar{\epsilon}_i$ is the average error of those individuals in the network linked to individual i. Returning to the academic performance example introduced in Subsection 2.2, the model can be interpreted as implying that if a student's friends have a below/above average AP value (as predicted by their SES values), then the student is more likely to have an AP value that is also below/above average.

Note that the Model (5) can be rewritten as

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \sigma^2(\mathbf{D}^\top\mathbf{D})^{-1}), \tag{6}$$

where $\mathbf{D}$ was defined in Subsection 2.3, with the same restrictions in place on θ as for the AR model. The parameter θ is an indicator of the strength of the between individual correlations generated by the network. For $\theta = 0$, the correlation between the Y values of any two individuals in the network is zero after one adjusts for their respective values of $\mathbf{X}$. Under (6), the conditional (on $\mathbf{X}$) mean and variance of $\mathbf{Y}$ are $\mu = \mathbf{X}\beta$ and $\mathbf{V} = \sigma^2(\mathbf{D}^\top\mathbf{D})^{-1}$ respectively.

It is worth pointing out that under the ND model, $\mu = \mathbf{X}\beta$ is unaffected by the social network, whereas under the AR model (4), $\mu = \mathbf{D}^{-1}\mathbf{X}\beta$ depends on the network through $\mathbf{D}$. That is, under the ND model, the expected value of Y for an individual only depends on the values of that individual's covariates. Unbiased prediction of Y can therefore ignore the network. Of course efficient prediction depends on the second order moments of (6), and so requires network information – as does prediction variance and mean squared error estimation. This is analogous to estimation under a multi-level model, where one can

ignore the multi-level structure of the data if unbiased estimation is the aim, but one needs to take this structure into account for efficient inference.

3. Prediction of Population Totals Using Network Models

The models discussed in the previous section are predictive models, that is, when second order moments are known, they can be used to compute efficient predictions of unknown values of the response variable. We now describe how these models can be fitted, and how predicted values derived from them can be used to estimate the population total $T_Y = \sum_{i \in U} Y_i$ given the sample values $\{Y_i \; i \in s\}$, the population matrix of model covariates $\mathbf{X}$ and either part of or all of the network matrix $\mathbf{Z}$. Throughout we assume that inclusion in sample does not depend on $\mathbf{Z}$ and that there is non-informative sampling given $\mathbf{X}$, see Section 1.4 in Chambers and Clark (2012). Consequently, all unknown parameter values for the standard model (1) can be estimated from the sample data and predicted values of Y for the non-sampled population individuals can be computed. We start by summarising known results from finite population estimation theory.

3.1. The Empirical Best Linear Unbiased Predictor

Let $E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{H}\boldsymbol{\lambda}$, where $\mathbf{H}$ is a known matrix with N rows and q columns and $\boldsymbol{\lambda}$ is an unknown parameter vector of length q. Also, suppose that $\mathrm{Var}(\mathbf{Y}) = \mathbf{V}$ is a positive definite matrix of order N whose value is known up to a constant of proportionality. Examples of $\mathbf{H}$ and $\mathbf{V}$ are given in the following subsection. The best linear unbiased predictor or BLUP of the population total $T_Y = \sum_{i \in U} Y_i$ is then an efficient estimator of this quantity, see Royall (1976). In order to specify the BLUP, let s and r denote the n sampled and $N - n$ non-sampled population individuals respectively, and put $\mathbf{H} = \left(\mathbf{H}_s^\top, \mathbf{H}_r^\top\right)^\top$ and $\mathbf{Y} = \left(\mathbf{Y}_s^\top, \mathbf{Y}_r^\top\right)^\top$. The matrix $\mathbf{V}$ can then be partitioned conformably as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{pmatrix}.$$

A standard expression for the BLUP is its so-called predictive form

$$\hat{T}_Y^{BLUP} = \sum_{i \in s} Y_i + \sum_{i \in r} \mathbf{H}_i \hat{\boldsymbol{\lambda}} + \sum_{i \in s} \tau_i (Y_i - \mathbf{H}_i \hat{\boldsymbol{\lambda}}), \tag{7}$$

where $\mathbf{H}_i$ is the ith row of $\mathbf{H}$, $\hat{\boldsymbol{\lambda}} = (\mathbf{H}_s^\top \mathbf{V}_{ss}^{-1} \mathbf{H}_s)^{-1} \mathbf{H}_s^\top \mathbf{V}_{ss}^{-1} \mathbf{Y}_s$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\lambda}$, and τ_i is the ith element of the vector $\mathbf{V}_{ss}^{-1} \mathbf{V}_{sr} \mathbf{1}_{N-n}$, with $\mathbf{1}_{N-n}$ denoting a vector of ones of size $N - n$.

Note that the BLUP can also be expressed as a weighted sum $\hat{T}_Y^{BLUP} = \sum_{i \in s} w_i Y_i = \mathbf{w}_s^\top \mathbf{Y}_s$ of the sample values of Y, where

$$\mathbf{w}_s = \mathbf{1}_n + \mathbf{M}^\top \left(\mathbf{H}^\top \mathbf{1}_N - \mathbf{H}_s^\top \mathbf{1}_n\right) + \left(\mathbf{I}_n - \mathbf{M}^\top \mathbf{H}_s^\top\right) \mathbf{V}_{ss}^{-1} \mathbf{V}_{sr} \mathbf{1}_{N-n} \tag{8}$$

is the vector of BLUP weights. Here $\mathbf{1}_n$ is a vector of ones of size n and matrix $\mathbf{M}$ is defined as $\mathbf{M} = \left(\mathbf{H}_s^\top \mathbf{V}_{ss}^{-1} \mathbf{H}_s\right)^{-1} \mathbf{H}_s^\top \mathbf{V}_{ss}^{-1}$.

A key assumption of the BLUP is that the variance matrix $\mathbf{V}$ is known up to a constant of proportionality. This is often unrealistic, since $\mathbf{V}$ can depend on unknown parameters,

which must then be estimated. Methods for doing this are described in the next section. Substituting these estimates into $\mathbf{V}$ defines its plug-in estimator $\hat{\mathbf{V}}$, which can be used in (8) instead of $\mathbf{V}$. The resulting estimator of the population total is called the empirical BLUP or EBLUP.

3.2. Calculating the EBLUP under Network Models

In order to use the EBLUP with the different network models defined in the previous section, we need to specify $\mathbf{H}$ and $\mathbf{V}$ as well as estimators of the unknown parameters that underpin these matrices. These are defined as follows:

Standard Model: Here $\mathbf{H} = \mathbf{X}$ and $\mathbf{V} = \sigma^2\mathbf{I}_N$. The residual mean squared error defines an unbiased estimator of σ^2.

CN Model: For this model $\mathbf{H} = [\mathbf{X}, \mathbf{U}]$ and $\mathbf{V} = \sigma^2\mathbf{I}_N$. We can unbiasedly estimate σ^2 using the residual mean squared error.

AR Model: In this case $\mathbf{H} = \mathbf{D}^{-1}\mathbf{X}$ with $\mathbf{D} = \mathbf{I}_N - \theta\mathbf{W}$ and $\mathbf{V} = \sigma^2(\mathbf{D}^\top\mathbf{D})^{-1}$. Estimates of σ^2 and θ can be obtained by Maximum Likelihood (ML). Restricted ML (REML) is often used to obtain unbiased variance estimates but it cannot be applied here, because both the mean and variance depend on the parameter θ. The EBLUP uses the plug-in estimates of $\mathbf{H}$ and $\mathbf{V}$ defined by the ML estimates of σ^2 and θ.

ND Model: Here $\mathbf{H} = \mathbf{X}$ and $\mathbf{V} = \sigma^2(\mathbf{D}^\top\mathbf{D})^{-1}$. ML estimation of σ^2 and θ can be carried out, and the resulting plug-in estimate of $\mathbf{V}$ is used to calculate the EBLUP.

ML estimation of σ^2 and θ for the AR and ND models is not straightforward. Both models are not reproducible, that is, they do not share the property that the model for a subset of units of the population has the same form as the model for the whole population. To see this, note that the variance of the population response vector $\mathbf{Y}$ under both models is $\sigma^2(\mathbf{D}^\top\mathbf{D})^{-1}$ so that the variance for the sample response vector $\mathbf{Y}_s$ is $\sigma^2[(\mathbf{D}^\top\mathbf{D})^{-1}]_{ss}$. In general, this will not equal $\sigma^2(\mathbf{D}_{ss}^\top\mathbf{D}_{ss})^{-1}$, which is the assumed variance if the model is fitted via ML at the sample level. This misspecification can lead to biased estimates of the model parameters. A modified approach that yields unbiased estimates of the fixed effects in the model is described in Suesse (2012a). However this is computationally intensive. An alternative approach replaces $\mathbf{D}^{-1}$ by a fourth order Taylor series approximation. This speeds up computation considerably since it effectively replaces matrices of dimension $N \times N$ by matrices of dimension $n \times n$. See Suesse (2012a) where it is shown that ML estimates based on this approximation are essentially identical to those obtained using the modified ML method. An alternative exact ML method that is computationally more demanding was considered by Suesse and Zammit Mangion (2017).

3.3. Variance Estimation for the EBLUP

The prediction variance of the BLUP is

$$\mathrm{Var}(\hat{T}^{BLUP} - T) = \tilde{\mathbf{w}}^\top\mathbf{V}\tilde{\mathbf{w}} \tag{9}$$

with $\tilde{\mathbf{w}}^\top = \left(\mathbf{w}_s^\top - \mathbf{1}_n^\top, -\mathbf{1}_{N-n}^\top\right)^\top$. This formula assumes that the vector of survey weights $\mathbf{w}_s$ is fixed. We can use the same formula for the EBLUP, although from (8) it is clear that

the EBLUP weights are not fixed in general because the plug-in estimates of **H** and **V** used to calculate them will depend on estimated parameters. However, the increase in the prediction variance due to ML estimation of these parameters will be small for large sample sizes, and can be ignored, see Chambers et al. (2011).

Using (9) to estimate the prediction variance of the EBLUP depends on correct specification of the second order moments of *Y*. For the standard model and the CN model, we can avoid this by using an alternative prediction variance estimator that does not rely on specification of these second order moments, see Section 9.2 of Chambers and Clark (2012). This estimator is given by

$$\widehat{\mathrm{Var}}(\hat{t}_{BLUP} - t) = \sum_{i \in s} (w_{is} - 1)^2 (Y_i - \hat{\mu}_i)^2 + (N - n)\hat{\sigma}^2 \tag{10}$$

where $\hat{\mu}_i$ is the estimated mean for $i \in s$, that is $\hat{\mu}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}}$ for the standard model and $\hat{\mu}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{U}_i\hat{\gamma}$ for the CN model, with $\hat{\sigma}^2$ corresponding to the usual unbiased estimator of σ^2 under each model.

For the AR and ND models we use Equation (9) with a plug-in estimator $\hat{\mathbf{V}}$. In this context, we note that ML estimates of variance parameters are known to be biased, which could therefore lead to a bias in $\hat{\mathbf{V}}$ and in the resulting plug-in estimator defined by (9). The standard approach to dealing with this issue is to apply REML instead of ML. Unfortunately, the AR model does not allow the application of REML, and furthermore REML is computationally more complex when fitting these population models. Consequently a bias-corrected version of ML was applied, based on the approach set out in Goldstein (1989), which adjusts Iterative Generalized Least Squares (IGLS) to obtain estimates that are equivalent to REML. The details of this are outlined in the Appendix of Suesse and Chambers (2014).

4. Modelling of Networks

Our EBLUP development in the previous section assumed that the matrix **Z** defining the network is known. This is rather unlikely to be the case. It is far more likely that we will know either just that part of the network defined by the sampled individuals (i.e., $\mathbf{Z}_{ss}$) or that part of the network defined by the sampled individuals and their corresponding network links (i.e., $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$).

An implementation of a 'network-based' EBLUP in this situation must therefore take account of this incomplete network data. In this section we describe simple model-based imputation methods that can be used to approximate the impact of the unknown full network (i.e., **Z**) on this EBLUP. In turn, this requires that we have a way of modelling **Z**, given that we see only a part of this matrix. We start with a brief overview of models for networks.

4.1. Exponential Random Graph Models

A popular class of models that is able to describe dependencies in a network **Z** is the class of (curved) exponential random graph models (ERGMs), these are discussed in Wasserman and Faust (1994) and Carrington et al. (2005). Under an ERGM, the

distribution of $\mathbf{Z}$ is characterised by

$$\Pr(\mathbf{Z} = \mathbf{z}) = \exp\big(\eta(\zeta)'\mathbf{G}(\mathbf{z}) - \kappa(\zeta)\big), \tag{11}$$

where ζ is the vector of model parameters, $\eta(\zeta)$ is a mapping from the p-dimensional to the q-dimensional space with $p \leq q$, and $\kappa(\zeta)$ is the normalising constant. Here $\mathbf{G}(\mathbf{z})$ is a vector of q 'network statistics' which, together with ζ, completely characterises the distribution of $\mathbf{Z}$. Simple examples of network statistics are the number of 'edges' in the network (i.e., the number of observed links, usually expressed as a fraction of the total number $N(N-1)$ of possible links) and the number of triangles (a triangle is said to exist between individuals i, j and k, if $Z_{ij} = Z_{jk} = Z_{ik} = 1$). A more complicated, but widely used network statistic is GWESP, or the geometrically weighted edgewise shared partner statistic. Roughly speaking, this corresponds to a weighted sum, over possible values of m, of counts of the number of links 'connecting' any two individuals in the network who are themselves linked to exactly m other individuals. Like interaction terms in regression, such statistics allow one to model networks whose 'connectivety' structure is extremely complicated.

Fitting an ERGM via ML is usually not possible, mainly because direct calculation of the normalising constant $\kappa(\zeta)$ is infeasible. One way of circumventing this problem is to sample from the network distribution (11) using a Markov-Chain-Monte-Carlo (MCMC) algorithm in order to obtain a stochastic approximation to the maximum likelihood estimate of ζ. Such estimates are called MCMC ML estimates (Hunter and Handcock 2006). Describing the network distribution via simple network statistics, such as the number of triangles then becomes problematic, because such specifications often lead to degenerate MCMC samples. Some authors (Snijders 2002; Snijders et al. 2006) have therefore proposed the use of more complex network statistics, such as the family of GWESP statistics, for which degeneracy seems less of a problem. For more details of network modelling, see Strauss and Ikeda (1990), Hunter and Handcock (2006), Hunter (2007), Hunter et al. (2008a), and Butts (2008).

4.2. Types of Partially Observed Networks

In the first case, denoted by SS in what follows, only $\mathbf{Z}_{ss}$ is observed and so $\mathbf{Z}_{sr}$, $\mathbf{Z}_{rs}$ and $\mathbf{Z}_{rr}$ are missing. In the second case, denoted by SS+SR in what follows, $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ are observed but $\mathbf{Z}_{rs}$ and $\mathbf{Z}_{rr}$ are missing. This might appear strange, because for an undirected network $\mathbf{Z}_{sr} = \mathbf{Z}_{rs}^{\top}$. This situation is motivated by the BHPS data set for which contextual information $\mathbf{U}_s$ is available for the sample but not for the non-sample that is, $\mathbf{U}_r$ is unavailable. This corresponds to knowing $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$, but not knowing $\mathbf{Z}_{rs}$ and $\mathbf{Z}_{rr}$, because $\mathbf{U}_s$ is function of $\mathbf{Z}_{ss}$, $\mathbf{Z}_{sr}$ and $\tilde{\mathbf{X}}$ and $\mathbf{U}_r$ is function of $\mathbf{Z}_{rs}$, $\mathbf{Z}_{rr}$ and $\tilde{\mathbf{X}}$. See Appendix C for more details on the relationship between $\mathbf{U}$ and $\mathbf{Z}$ and $\tilde{\mathbf{X}}$ and simple estimators.

The third case, denoted SS+SR+RS is where $\mathbf{Z}_{ss}$, $\mathbf{Z}_{sr}$ and $\mathbf{Z}_{rs}$ are observed, with $\mathbf{Z}_{rr}$ missing. The second and third cases are more realistic from the viewpoint of having usable network information, since here we at least have complete network information for all sampled individuals. In this context, we note that the second case provides a scenario which is related to the situation of the BHPS, where the network is not directly available but where contextual variables defined by the sample, that is $\mathbf{U}_s$, are known.

4.3. Imputation of Partly Observed Networks

An estimate $\hat{\mathbf{Z}}$ of the full network is necessary for calculation of the EBLUP under the network models considered in this article. However, in practice only part of the network will be observed, say $\mathbf{Z}^{obs}$, and another part will be missing, say $\mathbf{Z}^{mis}$. For example, for the scenario SS the observed network $\mathbf{Z}^{obs}$ is $\mathbf{Z}_{ss}$ and the missing network $\mathbf{Z}^{mis}$ is $\mathbf{Z}_{sr} \cup \mathbf{Z}_{rs} \cup \mathbf{Z}_{rr}$. Note that here we focus on single-value imputation of $\mathbf{Z}^{mis}$. Our approach can be extended to multiple imputation.

We apply a simple, robust and computationally feasible approach for imputation. Standard ERGMs, for example the ERGM with EDGES and GWESP, imply a fixed marginal probability of the form $P(Z_{ij} = 1) = p$. However, this probability cannot in general be analytically determined from the model parameters, and must instead be estimated separately, for example either via simulation from the underlying ERGM using plug-in estimates of the ERGM parameters, or more simply by the moment estimator $\hat{p} = \frac{1}{|\mathbf{Z}^{obs}|} \sum_{i,j,i \neq j} Z_{ij}^{obs}$. In what follows, we apply the latter approach as this is a standard estimator for a proportion, replacing all Z_{ij}^{mis} by $\hat{p}$. This approach is simple and clearly can be improved upon as it uses the unconditional expectation $E(\mathbf{Z}^{mis})$ as an estimator of the conditional expectation $E(\mathbf{Z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs})$. However, estimating the latter expectation requires first fitting an ERGM to the incomplete network data obtained from the sample, calculating model parameter estimates $\hat{\theta}$, and then applying model-based imputation methods, for example multiple imputation. This approach was infeasible for the simulation study reported next, as fitting an ERGM with a large portion of the network missing took more than four hours on a single core of type Intel Xeon E5-2620 v2 – 2.10GHz for a single data set with the latest available version of `ergm` (Hunter et al. 2008b). In comparison, when no missing data are present, this fitting process took only a few seconds. Note that such considerations may not be relevant in an application where imputation of a single network is required. In this case, adopting a more sophisticated imputation method may be advisable, for example by applying the approach of Pattison et al. (2013). However even if its estimation could be improved, sampling a large number of networks is needed to obtain an estimate of $E(\mathbf{Z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs})$ or even only $E(\mathbf{Z})$. But when N is large, for example $N = 100{,}000$, this also may not be feasible since sampling one network took 31 hours, meaning that for large N using a more sophisticated method is impractical even if estimation is not an issue. That is, in practice, for large N the simple method mentioned above appears to be the only feasible method.

5. Mean Squared Error Estimators under the Standard and Contextual Linear Model

The AR and the ND models are difficult to fit for large N when the networks are imputed, because then the contiguity matrices are not sparse any more, making it near impossible to calculate the log-likelihood. From a practical perspective, only the contextual model is feasible to fit for large N when the network is imputed. We consider now mean squared error estimation under the standard linear model and under the contextual model when the network is only partially available (situation SS+SR) but when $\tilde{\mathbf{X}}$ is available for the whole population. For this situation, the population total $\mathbf{T}_{\mathbf{U}} = 1_N^\top \mathbf{U}$ with $\mathbf{U} = \left(\mathbf{U}_s^\top, \mathbf{U}_r^\top\right)^\top$ referring to the total of the contextual variable must be estimated, that is $\hat{\mathbf{U}} = g(\hat{\mathbf{Z}}, \tilde{\mathbf{X}})$.

In Appendix C, we show under SS+SR ($\mathbf{U}_s$ known) that $\hat{\mathbf{U}}_r$ has a simple form that does not require the network to be observed, but only requires an estimate of the constant $P(Z_{ij} = 1) = p$, $\tilde{\mathbf{X}}$ and the parameters N and n. Hence even for the BHPS for which p is known, this estimator $\hat{\mathbf{U}}$ can be calculated relatively easily and the estimated total $\hat{\mathbf{T}}_{\mathbf{U}}$ obtained. Appendix C shows the estimator $\hat{\mathbf{T}}_{\mathbf{U}}$ along with a (co)variance estimator denoted by $\hat{\mathbf{V}}_{\hat{\mathbf{T}}_{\mathbf{U}}}$.

Let the mean squared error (MSE) of the EBLUP $\hat{T}$ be denoted by $MSE[\hat{T}]$ and let the EBLUP for the linear model without contextual information be denoted by $\hat{T}_I$ and the corresponding EBLUP defined by the contextual model be denoted $\hat{T}_C$.

Put $\Delta = MSE[\hat{T}_C] - MSE[\hat{T}_I]$. See Appendix B for an expression for Δ depending on $\hat{\mathbf{T}}_{\mathbf{U}}$ and $\hat{\mathbf{V}}_{\hat{\mathbf{T}}_{\mathbf{U}}}$ along with an estimator $\hat{\Delta}$. In practice we propose to use these estimators to choose between the models without and with contextual information. When $\hat{\Delta} < 0$ we propose to use the contextual model and when $\hat{\Delta} \geq 0$, we propose to use the standard model. Appendix B also shows expressions for Δ and an estimator for $\hat{\Delta}$ for the unrealistic situation that $\mathbf{U}$ is fully known, using the estimator $\hat{\Delta}$ proposed by Clark and Chambers (2008).

6. Simulation Study

6.1. Study Design

This section contains the results from a simulation study whose aim was to investigate the effect of using networks as an additional source of information when estimating the population total T_Y of a survey variable Y. A networked population of size $N = 1{,}000$ was independently simulated 2,000 times, balancing computation time against the number of different scenarios that were explored in the study, and independent simple random samples of size $n = 100$ and $n = 200$ were independently selected without replacement from each simulated population. This study comprises all four models under investigation. We also consider larger N, but then estimation of the AR and ND models becomes infeasible. To illustrate the effect of large N, only the contextual model is compared with the standard linear model.

6.1.1. Network Generation

We mainly consider undirected networks in our simulations. The literature on network analysis suggests that such networks are often well characterised by an ERGM defined in terms of an EDGES (number of edges) statistic and a GWESP statistic (Hunter et al. 2008b). Consequently, $\mathbf{Z}$ was generated as a random draw from an ERGM with an EDGES statistic equal to θ on the logit scale and a weight parameter of 1.0 for the GWESP statistic. In what follows we use ERGM(m) to denote such an ERGM, where m is the network density, that is the average number of links per individual. The values of θ were then chosen in order to generate a network with a density of about $m = 3, 15, 50$ network links respectively for each individual, that is $m \approx P(Z_{ij} = 1) \times N$ with $P(Z_{ij} = 1) \approx \text{expit}(\theta)$. Note that with this specification the number of network links for an individual is random, with only the approximate population average number of links fixed.

In this study we consider the three types of partially observed networks mentioned before, namely SS, SS+SR and SS+SR+RS. Finally, we also considered the situation

where no network data are used (the standard model) and also the case where the population network is fully known.

6.1.2. Parameter Specification for Linear Network Models

We generated data under the CN, AR and ND linear network models, see Section 2. Population data were simulated assuming $\sigma^2 = 3^2 = 9$, $\beta_0 = 1$ and $\beta_1 = 2$. Furthermore, the auxiliary variable X was defined so that it took values randomly in the set $\{1, \ldots, 9\}$. This model has medium to high predictive power, since the Standard model implies a theoretical value of approximately $R^2 = 0.75$.

CN Model:

$$Y_i = \beta_0 + X_i\beta_1 + U_i\gamma + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Here $\gamma = 2$ and the contextual variable U_i is defined as the average value of X for all other individuals in the network that individual i has links with, that is $\mathbf{U} = \mathbf{WX}$, where $\mathbf{W}$ is the row-normalised version of $\mathbf{Z}$ and $\mathbf{X}$ denotes the vector of population values of X.

AR Model:

$$\mathbf{Y} = \theta\mathbf{WY} + \beta_0 + \mathbf{X}\beta_1 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_N)$$

with $\theta = 0.5$.

ND Model:

$$\mathbf{Y} = \beta_0 + \mathbf{X}\beta_1 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} = \theta\mathbf{W}\boldsymbol{\epsilon} + \mathbf{v}, \quad \mathbf{v} \sim N(0, \sigma^2\mathbf{I}_N)$$

with $\theta = 0.5$.

6.2. Simulation Results

Results for the $n = 100$ case are presented. Table 1 show the Monte Carlo relative mean squared errors of the estimates of T when the network is generated under an ERGM where the total number of friends is random, with expectations three and ten respectively. Corresponding simulation results for $X \sim N(0, 25)$ can be found in Suesse and Chambers (2014). Results for the $n = 200$ case are similar. Note that we do not show Monte Carlo bias, since these values were effectively zero for all methods. The results displayed in each table include the two cases where the network is ignored (the 'standard' model) and when the population network matrix $\boldsymbol{Z}$ is fully known ('True Model and $\mathbf{Z}_U$ known'). For partially observed network data we show results for the SS case (only $\mathbf{Z}_{ss}$ known), the SS+SR case ($\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ known, simple imputation) and the SS+SR+RS case ($\mathbf{Z}_{ss}$, $\mathbf{Z}_{sr}$ and $\mathbf{Z}_{rs}$ known, simple imputation). All results are shown relative to those for the BLUP, which uses complete network information as well as knowledge of θ. Although the level of knowledge required to compute the BLUP is unrealistic in practice, its performance provides us with a benchmark against which to gauge the relative benefit of putting more effort into collecting more network information and in carrying out more intensive network modelling for imputation of the unknown parts of the network. Furthermore, comparisons with the 'Standard' case allow us to assess how much efficiency is lost by ignoring network information.

Table 1. n = 100: Undirected ERGM(3) and ERGM(10) networks with X drawn randomly from {1, . . .,9}. Ratio of MSE(EBLUP) to MSE(BLUP).

	Population data generated under model					
	ERGM(3)			ERGM(10)		
	CN	AR	ND	CN	AR	ND
Actual MSE	86, 474	101, 148	101, 082	86, 537	89, 483	89, 705
Relative EBLUP to actual MSE based on						
True model and $\mathbf{Z}_U$ known	1.00	1.00	1.01	1.00	0.99	1.02
CN						
SS+SR+RS	1.11	1.09	1.04	1.03	1.02	1.00
SS+SR	1.11	1.09	1.04	1.03	1.02	1.00
SS	2.33	1.37	1.03	1.32	1.07	1.01
AR						
SS+SR+RS	1.34	1.05	1.05	1.09	1.01	1.00
SS+SR	1.20	1.06	1.05	1.07	1.01	1.00
SS	2.42	1.39	1.03	1.33	1.06	1.01
ND						
SS+SR+RS	2.21	1.31	1.03	1.36	1.06	1.00
SS+SR	2.36	1.35	1.01	1.34	1.07	1.01
SS	2.42	1.40	1.02	1.34	1.07	1.01
Standard	2.40	1.40	1.03	1.33	1.07	1.00

It is clear from the results shown in Table 1 that ignoring the network (i.e., using the 'Standard' model for estimation) can lead to a large loss in efficiency if in fact either the AR or the CN models are true. Interestingly, our results also seem to indicate that adopting the CN model when in fact the AR model is true seems as good as using the correctly specified AR model when the number of friends is not small. Note that when the ND model is true, ignoring the network information in the data only leads to a marginal loss in efficiency. In fact, the EBLUPs based on the different network models are all almost fully efficient in this case, irrespective of whether the assumed network model is true.

When $\mathbf{Z}$ is known, but not θ, we see a loss of efficiency under the AR model, mainly because the pseudo-design matrix $\mathbf{D}^{-1}(\theta)\mathbf{X}$ for this model depends on the estimated value of θ. As the number of friends increases, this loss of efficiency associated with having to estimate θ from the sample data decreases in importance. This problem is much less of an issue for the ND model because in this case the design matrix does not depend on θ. Obviously, there is no impact under the CN model.

In order to see why the CN model yields similar results as the AR model when in fact the AR model holds, we note that the mean of the AR model is $\mu = \mathbf{D}(\theta)^{-1}\mathbf{X}\beta$. If we approximate $\mathbf{D}(\theta)^{-1}$ by a first order Taylor series around zero, that is

$\mathbf{D}(\theta)^{-1} = (\mathbf{I}_N - \theta\mathbf{W})^{-1} \approx \mathbf{I}_N + \theta\mathbf{W}$, then

$$\mu \approx \mathbf{X}\beta + \theta\mathbf{W}\mathbf{X}\beta = \mathbf{X}\beta + \mathbf{U}\gamma$$

with $\gamma = \theta\beta$ and $\mathbf{U} = \mathbf{W}\mathbf{X}$. That is, the implied mean structure under the AR model is approximately the same as that under a CN model.

When $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ are observed, the EBLUP based on the CN model appears to perform well generally. This is because the EBLUP under this model does not depend on either $\mathbf{Z}_{rs}$ or $\mathbf{Z}_{rr}$ and hence is unaffected by imputation of this part of the network. This is in contrast to the performance of this EBLUP when only $\mathbf{Z}_{ss}$ is observed. Here we see that the need to impute $\mathbf{Z}_{sr}$ leads to a significant loss of efficiency. Since estimation of θ in the pseudo-design matrix $\mathbf{D}(\theta)^{-1}\mathbf{X}$ under the AR model has a larger negative effect than the approximation of the AR model by the CN model, we conclude that the EBLUP based on the CN model seems a generally more robust method for estimating the population total than the EBLUP based on the AR model.

It is interesting to also observe that the EBLUP based on the AR model and SS+SR network data performs generally better than the same EBLUP with access to more extensive SS+SR+RS network data when the CN model holds, reflecting the interaction of model misspecification and imputation biases. However, this effect is reversed when a ND-based EBLUP is used and a CN model underpins the network.

When we focus on where the expected number of friends per subject is small, here equal to three (see Table 1), we note that there is only a small gain associated with using imputation method SS compared to ignoring the network information and basing estimation on the 'Standard' model. In this situation network imputation based on SS+SR or SS+SR+RS provides the largest gains relative to ignoring the network when the contextual CN or AR model is fitted.

To investigate the effect for larger N only the contextual model and the standard model are compared, as the AR and ND models are infeasible to fit. Table 2 shows the MSE of the BLUP and various EBLUPs including the adaptive strategy that chooses either model depending on the sign of $\hat{\Delta}$ for an undirected ERGM network with approximately ten friends per person. Table 3 shows results for a directed ERGM network with approximately three friends per person. The tables also show the empirical mean of $\hat{\Delta}$ denoted by $E(\hat{\Delta})$ and Δ. Note that the estimator $\hat{\Delta}$ has only a small bias.

The results also show that the adaptive strategy is effective and that this strategy also works for large N and small sampling fractions. Surprisingly the efficiency gains even increase as the sampling ratio increases. This might be due to the fact that the simple estimate $\hat{p}$, the proportion of links in the observed network, is more precise with increasing N because the number of dyads in $\mathbf{Z}^{obs}$ is $n(N-1)$ and increases with N.

Average lengths and associated coverages for nominal 95% Gaussian confidence intervals generated by the estimates of the mean squared errors of the different estimators are set out in Table 6, see the Appendix. Results for $X \sim N(0, 25)$ can be found in Suesse and Chambers (2014). Monte Carlo coverages in all cases are close to the nominal level. However, the average confidence interval length in the SS+SR/SS+SR+RS case is considerably shorter than that for the SS case when estimation is carried out under the AR

Table 2. Undirected ERGM(10) network with X drawn randomly from {1, . . . , 9}. Ratio of MSE(EBLUP) to MSE(BLUP) and estimated and true MSE-difference Δ.

	Settings of N and n					
N	1,000	10,000	10,000	100,000	100,000	100,000
n	100	100	200	100	200	1,000
BLUP -actual MSE multiplied by $100/N^2$	8.653	9.095	8.940	9.310	9.160	9.064
Relative MSE of EBLUP based on						
$\mathbf{Z}_U$ known	1.000	1.000	1.000	1.000	1.000	1.000
SS+SR+RS	1.092	1.004	1.006	1.000	1.001	1.001
SS+SR	1.075	1.004	1.006	1.000	1.001	1.001
SS	1.334	22.69	1.334	1658	734.1	1.321
Standard	1.337	1.318	1.340	1.335	1.317	1.323
Adaptive strategy	1.078	1.004	1.006	1.000	1.001	1.001
$-\Delta$ relative to BLUP	0.249	0.314	0.328	0.334	0.316	0.322
$E(-\hat{\Delta})$ relative to BLUP	0.253	0.316	0.322	0.318	0.317	0.318

and CN models. This provides further support for the conclusion reached above, that basing an EBLUP on a CN model seems a generally robust approach to using network data when estimating a population total, even though one must keep in mind that the simple linearization-based prediction variance estimator (10) used with the CN model slightly

Table 3. Directed ERGM(3) network with X drawn randomly from {1, . . . , 9}. Ratio of MSE(EBLUP) to MSE(BLUP) and estimated and true MSE-difference Δ.

	Settings of N and n					
N	1,000	10,000	10,000	100,000	100,000	100,000
n	100	100	200	100	200	1,000
BLUP -actual MSE multiplied by $100/N^2$	8.401	9.062	8.931	9.293	9.156	9.051
Relative MSE of EBLUP based on						
$\mathbf{Z}_U$ known	1.000	1.000	1.000	1.000	1.000	1.000
SS+SR	1.122	1.013	1.024	1.001	1.002	1.011
SS	2.202	529.1	21.59	2349	2936	2.300
Standard	2.201	2.259	2.274	2.267	2.262	2.310
Adaptive strategy	1.122	1.013	1.024	1.001	1.002	1.011
$-\Delta$ relative to BLUP	1.079	1.246	1.250	1.266	1.261	1.299
$E(-\hat{\Delta})$ relative to BLUP	1.095	1.288	1.265	1.252	1.241	1.262

underestimates random variation due to its assumption of fixed sample weights, resulting in too narrow confidence intervals with slight undercoverage.

7. Illustrative Example

The British Household Panel Study (BHPS) is an annual multi-purpose household panel survey in the United Kingdom that focuses on gaining insight into the social and economic change at the individual and household level in Britain and the UK, see https://www.iser.essex.ac.uk/bhps/ for more details.

We focus on an individual's annual income (in pounds sterling) as the variable of interest. Our aim is to investigate how the use of network information available in BHPS impacts on average income estimates for the cross-classification age by gender by region, using six categories for age 15−18 (1), 19−21 (2), 22−30 (3), 31−50 (4), 51−64 (5), 65+ (6) (in years), two for gender (1: male, 0: female) and five regions defined as: (1) 'E/North' consisting of East Midlands, West Midlands Conurbation, Rest of West Midlands, Greater Manchester, Merseyside, Rest of North West, South Yorkshire, West Yorkshire, Rest of Yorks & Humberside, Tyne & Wear, Rest of North; (2) 'E/South' containing Rest of South East, South West and East Anglia; (3) 'London' includes inner and outer London; and finally (4) 'Scotland' and (5) 'Wales'. Northern Ireland is excluded from our analysis because BHPS sample sizes in Northern Ireland were too small to cross-classify by age and gender. We also exclude persons who did not report a positive income.

Incomes estimates for the cross-classification age by gender by region based on the linear model with two-way interaction effects age by sex are shown in Table 4. The BHPS also collects information from a respondent on his/her three best friends, consisting of the genders and ages of these friends, duration of friendships, frequency of contact, distances to the friends, their job/employment statuses, and their ethnicities. A contextual model can take a contextual effect based on the three friends and the collected variables into account.

Table 4. Gender by age by region cross-classification of estimated mean annual income in pounds sterling, using BHPS data and with weighting based on model with age by sex interactions.

		Region				
Gender	*Age*	London	E/North	E/South	Scotland	Wales
Female	≤18	2,668	3,666	2,279	1,678	4,120
	19−21	6,989	7,134	6,818	7,036	8,976
	22−30	17,068	12,759	13,403	14,147	12,774
	31−50	20,266	14,881	16,514	16,043	14,565
	51−64	12,129	10,725	10,931	11,822	11,279
	≥65	8,582	7,283	7,952	7,851	8,793
Male	≤18	1,257	3,896	2,180	2,578	4,897
	19−21	10,102	7,600	9,735	6,735	16,270
	22−30	15,617	15,617	18,294	20,719	15,960
	31−50	23,884	23,884	29,216	29,216	21,947
	51−64	20,186	20,186	23,293	27,237	19,305
	≥65	11,775	11,775	14,540	12,055	12,613

Table 5. Change in estimated mean annual income when BHPS data are weighted using the CN model based on age by sex interactions plus a main effect for maleness.

		Region				
Gender	*Age*	London	E/North	E/South	Scotland	Wales
Female	≤18	162	-10^{2}	-110^{4}	458^{4}	391^{4}
	19−21	324^{1}	47^{4}	0	-127^{4}	-372^{4}
	22−30	22^{2}	2^{4}	-3^{4}	-10^{4}	-7^{4}
	31−50	77^{2}	-2^{4}	-6^{4}	-3^{4}	-29^{4}
	51−64	−8	-1^{4}	5^{4}	-6^{4}	-2^{4}
	≥65	−7	5^{4}	−1	4^{4}	4^{4}
Male	≤18	-176^{3}	-20^{4}	99^{4}	106^{4}	-771^{4}
	19−21	139^{1}	-5^{4}	-15^{4}	71^{4}	-710^{4}
	22−30	−23	41^{4}	-47^{4}	-89^{4}	20^{4}
	31−50	-134^{1}	30^{4}	-60^{4}	114^{4}	-14^{4}
	51−64	-214^{1}	40^{4}	-50^{4}	209^{4}	56^{4}
	≥65	119^{1}	-5^{4}	-2^{3}	-74^{4}	12^{4}

[1, 2, 3, 4] Difference larger than 1, 2, 3, 4 estimated standard errors.

This situation of having contextual variables $\mathbf{U}_s$ available from the sample s corresponds to the SS+SR case, as $\mathbf{U}_s$ is defined by $\mathbf{U}_s = \mathbf{W}_s\tilde{\mathbf{X}}$ with $\mathbf{W}_s = \text{Diag}(\mathbf{Z}_s 1_N)^{-1}\mathbf{Z}_s$ and $\mathbf{Z}_s = (\mathbf{Z}_{ss}, \mathbf{Z}_{sr})$, see Equation (3). As the number of friends is fixed, $P(Y_{ij} = 1) = p$ does not need to be estimated, but is known to be $p = 3/(N-1)$, as there are $N-1$ candidate friends.

Table 5 shows the difference in estimates to Table 4 when estimation is based on a contextual model with effects age by sex and a contextual gender effect, where gender is a binary variable indicating whether a friend is male. In this context the model implies that a person's income is not only predicted by age by sex but also by the average income of the person's three best friends. For further details on how these estimates were obtained, see Suesse and Chambers (2012).

Table 5 shows that the application of the contextual model leads to substantially different results, as many differences are larger than 1, 2, 3, and often 4 standard errors.

We calculated the value of Δ for all cells presented in Tables 4 and 5. For most of the cells, $\hat{\Delta} > 0$ and hence the standard model should be used, however there are some cells for which indeed $\hat{\Delta} < 0$ and the contextual model is deemed as better in terms of a lower predicted MSE. For the London region the cells with $\hat{\Delta} < 0$ are with the estimated improvement of the MSE relative to the standard model in brackets: female and 19–21 years (6%), female and 31–50 years (0.1%), and male and ≥65 years (1%). These results are based on reconstructed data using the publicly available survey weights, so these statements are to be treated with caution, as the real data might yield different results.

8. Discussion

At the end of Section 1, we stated that our aim in this article is to address the questions: (i) Is embedding network information useful for survey estimation? (ii) If the answer to (i) is

yes, then which models are potentially useful? and (iii) How much network data needs to be collected in order to obtain potentially higher precision for survey estimation? Given the simulation results that we present in Section 6, our tentative answer to (i) is yes, and our corresponding answer to (ii) is the CN and AR models when either model is true, because in both cases the mean of the response depends on the network. Our simulation results provide some evidence that this conclusion may hold more generally.

However when the mean does not depend on the network, as is the case under the ND model, our results suggest that ignoring the network does not result in a significant loss of efficiency. We have also investigated this for other 'network covariance' models, where the mean structure is unaffected by the network, and we have observed similar results, see Suesse and Chambers (2014). In effect, ignoring the network under the CN and AR models leads to a misspecification of the mean model, but this does not apply for the ND (and similar) models. Finally, our answer to (iii) is that in realistic applications it will usually be impossible to collect the full network, and our simulation results are some evidence that when either the CN model or the AR model is true then both $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ must be collected or alternatively the contextual sample information $\mathbf{U}_s$ along with an estimate of $p = P(Z_{ij} = 1)$ must be available, as for the BHPS data set, in order to obtain efficiency gains. Knowledge of $\mathbf{Z}_{ss}$ alone is not enough.

In practice, we suggest a careful model fitting exercise be carried out before attempting to use either the CN model or the AR model for survey estimation. Given the numerical difficulties with fitting the AR model, see Suesse (2012a), we recommend that the CN model be used if it is a reasonable fit to the data and when $\hat{\Delta} < 0$, otherwise caution is warranted and ignoring the network might be the best option.

Clearly, more extensive information on networks needs to be collected in conjunction with standard survey data to gain further insight into the usefulness of network models for survey estimation. However, it is extremely unlikely that in practical applications complete network data will be available, in which case the issue of imputation for missing network data arises. In this article we base this imputation on the fact that the sample proportion of links per individual is a simple nonparametric estimator of the marginal probability of an unobserved link. A reasonable question to ask then is whether it is better to use an imputation method based on $\mathbb{E}(\mathbf{Z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs})$? The numerical intensity of the MCMC methods used to fit network models like the ERGM when population sizes are large meant that we could not fully explore this issue here. There is current research that tries to address some of these issues (Koskinen et al. 2010), but more is required, because even if the time issue of fitting a partially observed network is solved, simulating many ERGM networks to obtain $\mathbb{E}(\mathbf{Z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs})$ or even just $E(\mathbf{Z})$ for large $N > 100,000$ still appears infeasible.

Suesse and Chambers (2014) considered the case of known ERGM parameters (without estimating them) and then applied a more sophistcated method, however even this method was consistently worse than the simple method, despite having full knowledge of the ERGM parameters. Based on these results we anticipate that more sophisticated imputation methods are unlikely to lead to substantial efficiency gains in most cases. Our simulation results indicate marginal differences between the SS+SR case and where $\mathbf{Z}_U$ is known. We therefore hypothesize that more sophisticated imputation methods will also only provide marginal gains for estimation of a population total, and not alter the

conclusions of this article. Efficient imputation methods when the target of inference is more complex require further research.

All network models considered in this article assume that the value of the response variable Y for an individual in the study population depends on a linear combination of the values of this variable for the other individuals in the population that are linked to this person in the network. If there is an implicit ordering in the strength of these links, then this can be allowed for in the network model for Y. For example, in the case of a 'best friend' network, where the friendships are ordered by their strength, one can modify the CN model so that there is a separate parameter for each level of 'best friend', see Friedkin (1990) for similar examples. To illustrate, in the BHPS application, when this extended contextual model is fitted, a Wald test for equality of these effects supports the assumption of a common effect.

The use of ERGMs to model the network and the use of the three main regression models in Section 2 using the network as additional information might be restrictive. There are many other approaches to model networks, for example adding latent variables to a logistic regression model, as proposed by Handcock et al. (2007), but also many extensions to include network information in a regression model, see Leenders (2002) for some model extensions and the various options to define the weight matrix $\mathbf{W}$ based on the network $\mathbf{Z}$. The use of particular models might be beneficial or detrimental and exploring the use of alternative models could be useful. The same holds for using particular network structures in the modelling approaches. Conducting other simulation studies to investigate the merits of different models is subject to future research.

Finally, we note that throughout this paper we have assumed that the method of sampling is independent of the network structure given the available population auxiliary information. In effect, we assume that measurement of the network is something that is done on the sample (as in our BHPS application), rather than sampling being something that is carried out on the network. However, there are important applications, see Thompson and Seber (1996), where inclusion in sample depends on being linked to another sampled individual via a network. It is clear that in these cases we cannot treat the observed network structure in $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ in the same way as we have in this article, and this 'informative' method of sampling needs to be taken into account when we attempt to impute the unknown components of $\mathbf{Z}$. Work on this problem is continuing.

Appendix

A. *Further Simulation Results*

Table 6. n = 100: ERGM(3) and ERGM(10) network with X drawn randomly from {1, . . .,9}. Ratio of average lengths of nominal 95% Gaussian CIs (EBLUP/BLUP), with % actual coverage in subscript.

	Population Data Generated Under Model					
	ERGM(3)			ERGM(10)		
	CN	AR	ND	CN	AR	ND
Actual BLUP av(length)	$1,128_{94.2}$	$1,273_{95.0}$	$1,273_{95.0}$	$1,127_{94.1}$	$1,160_{95.2}$	$1,160_{95.2}$
Relative av(length) EBLUP based on						
True Model $\mathbf{Z}_U$ known	$0.98_{93.2}$	$0.98_{94.2}$	$0.98_{94.1}$	$0.98_{93.4}$	$0.98_{94.4}$	$0.98_{94.3}$
CN						
SS+SR+RS	$0.98_{91.7}$	$1.02_{94.1}$	$1.01_{94.9}$	$0.98_{93.2}$	$0.99_{94.2}$	$0.99_{94.7}$
SS+SR	$0.98_{91.9}$	$1.02_{94.1}$	$1.01_{94.9}$	$0.98_{93.1}$	$0.99_{94.4}$	$0.99_{94.6}$
SS	$1.47_{92.9}$	$1.16_{94.8}$	$1.01_{94.6}$	$1.11_{93.4}$	$1.03_{94.9}$	$0.99_{94.9}$
AR						
SS+SR+RS	$1.08_{92.5}$	$0.99_{93.6}$	$1.00_{94.2}$	$0.98_{92.8}$	$0.98_{94.0}$	$0.98_{94.5}$
SS+SR	$1.01_{92.0}$	$0.99_{92.9}$	$1.00_{94.4}$	$0.98_{93.1}$	$0.98_{94.0}$	$0.98_{94.5}$
SS	$1.46_{92.4}$	$1.13_{94.2}$	$0.99_{94.8}$	$1.11_{93.3}$	$1.02_{94.5}$	$0.98_{94.1}$
ND						
SS+SR+RS	$1.08_{83.9}$	$0.99_{91.5}$	$1.00_{94.6}$	$0.98_{89.1}$	$0.98_{94.1}$	$0.98_{94.3}$
SS+SR	$1.42_{92.4}$	$1.12_{94.0}$	$0.98_{94.2}$	$1.11_{93.2}$	$1.02_{94.7}$	$0.98_{94.5}$
SS	$1.49_{92.8}$	$1.14_{93.7}$	$0.98_{94.2}$	$1.12_{93.7}$	$1.02_{94.7}$	$0.98_{94.5}$
Standard	$1.49_{93.6}$	$1.17_{94.5}$	$1.01_{94.8}$	$1.13_{93.7}$	$1.03_{95.2}$	$0.99_{94.6}$

B. *Derivations*

Suppose the contextual model (CN Model) holds $Y_i = \mathbf{X}_i\beta + \mathbf{U}_i\gamma + \epsilon_i$ such that $E(Y_i) = \mathbf{X}_i\beta + \mathbf{U}_i\gamma$, $\text{Var}(Y_i) = \sigma_i^2 = v_i\sigma^2$ and $\text{Cov}(Y_i, Y_j) = 0$ for $i \neq j$. The column vector β is of length p and γ of length q. When not using the contextual information (standard model), then $\gamma = 0$ and $E(Y_i) = \mathbf{X}_i\beta$.

Define $\mathbf{H}_i = (\mathbf{X}_i, \mathbf{U}_i)$ and $\lambda = (\beta^\top, \gamma^\top)^\top$. Under independence, the weights of the BLUP $\hat{T} = \mathbf{w}_H^\top \mathbf{Y}_s$ given by (8) simplify to $\mathbf{w}_H$

$$\mathbf{w}_H = \mathbf{1}_s + \mathbf{D}_\mathbf{v}^{-1}\mathbf{H}_s\left\{\mathbf{H}_s^\top \mathbf{D}_\mathbf{v}^{-1}\mathbf{H}_s\right\}^{-1}\mathbf{T}_{\mathbf{H}r}^\top, \tag{12}$$

where $\mathbf{T}_{\mathbf{H}r} \equiv \sum_{i\in r}\mathbf{H}_i = \mathbf{H}_r\mathbf{1}_r$ is a row vector of length $q + p$, similarly the population totals are defined, for example $\mathbf{T}_\mathbf{H} \equiv \sum_{i\in U}\mathbf{H}_i = \mathbf{H1}$.

We aim at comparing the MSE of the CN and the standard model, but also under the situation SS+SR, that is when contextual information needs to be estimated for the non-sample.

$$MSE[\hat{T}] = E[(\hat{T} - T_Y)]^2$$

$$= E\left\{\sum_{i\in s} w_{Hi}Y_i - \sum_{i\in U} Y_i\right\}^2 + \mathrm{Var}\left[\sum_{i\in s}(w_{Hi}-1)Y_i - \sum_{i\in r} Y_i\right] \tag{13}$$

$$= E\left\{\mathbf{w}_H^\top \mathbf{Y}_s - \mathbf{Y}^\top \mathbf{1}\right\}^2 + \mathrm{Var}\left[(\mathbf{w}_H - \mathbf{1}_s)^\top \mathbf{Y}_s - \mathbf{Y}_r^\top \mathbf{1}_r\right]$$

Then according to Clark and Chambers (2008), $MSE[\hat{T}]$ can be re-written as

$$MSE[\hat{T}] = \left\{\left(\sum_{i\in s} w_{Hi}\mathbf{H}_i - \sum_{i\in U}\mathbf{H}_i\right)\boldsymbol{\lambda}\right\}^2 + \sum_{i\in s}(w_{Hi}-1)^2\mathrm{Var}(Y_i) + \sum_{i\in r}\mathrm{Var}(Y_i) \tag{14}$$

$$= \mathbf{d}_\mathbf{H}(\boldsymbol{\lambda}\boldsymbol{\lambda}^\top)\mathbf{d}_\mathbf{H}^\top + \sum_{i\in s}(w_{Hi}-1)^2\sigma_i^2 + \sum_{i\in r}\sigma_i^2 \tag{15}$$

where $\mathbf{d}_\mathbf{H} = \sum_{i\in s} w_{Hi}\mathbf{H}_i - \mathbf{T}_\mathbf{H}$. The term $\boldsymbol{\lambda}\boldsymbol{\lambda}^\top$ can be estimated by $\hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\lambda}}^\top - \widehat{\mathrm{Var}}(\hat{\boldsymbol{\lambda}})$.

Let the EBLUP under the CN model using $\mathbf{H}_i$ be denoted by $\hat{T}_C$ and that of standard model using only $\mathbf{X}_i$ for $i \in U$ by $\hat{T}_I$.

Then the difference $\Delta \equiv MSE[\hat{T}_C] - MSE[\hat{T}_I]$ can be estimated by

$$\hat{\Delta} = \mathbf{d}_\mathbf{H}(\hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\lambda}}^\top - \widehat{\mathrm{Var}}(\hat{\boldsymbol{\lambda}}))\mathbf{d}_\mathbf{H}^\top - \mathbf{d}_\mathbf{X}(\hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\lambda}}^\mathrm{T} - \widehat{\mathrm{Var}}(\hat{\boldsymbol{\lambda}}))\mathbf{d}_\mathbf{X}^\top + \sum_{i\in s}(w_{Hi}-1)^2\hat{\sigma}_i^2 - \sum_{i\in s}(w_{Xi}-1)^2\hat{\sigma}_i^2.$$

Since we assume that the contextual model holds $\mathbf{d}_\mathbf{H} = 0$ and

$$\hat{\Delta} = -\mathbf{d}_\mathbf{X}(\hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\lambda}}^\top - \widehat{\mathrm{Var}}(\hat{\boldsymbol{\lambda}}))\mathbf{d}_\mathbf{X}^\top + \sum_{i\in s}(w_{Hi}-1)^2\hat{\sigma}_i^2 - \sum_{i\in s}(w_{Xi}-1)^2\hat{\sigma}_i^2.$$

According to Clark and Chambers (2008) the contextual model is chosen when $\hat{\Delta} < 0$.

For the simple case of one contextual variable U_i the result of Clark and Chambers (2008) applies and $\hat{\Delta}$ simplifies to

$$\hat{\Delta} = T_{Ur}^2\left(2\hat{\sigma}^2 S_u^{-1} - \hat{\gamma}^2\right), \tag{16}$$

and the contextual model is chosen when $\hat{\gamma}^2 > 2\hat{\sigma}^2 S_u^{-1}$, where $S_c \equiv \sum_{i\in s} c_i$ and $c_i \equiv U_i - C^\top \mathbf{X}_i^\top$ with $C \equiv \left(\sum_{i\in s}\mathbf{X}_i^\top\mathbf{X}_i\right)^{-1}\mathbf{X}_i^\top U_i$.

Now suppose that $\mathbf{X}_i$ is known for all units $i \in U$ and $\mathbf{U}_i$ is only known for $i \in s$, that is the population totals $\mathbf{T}_\mathbf{X} \equiv \sum_{i\in U}\mathbf{X}_i$ of the covariates $\mathbf{X}_i$ are known, whereas $\mathbf{T}_{\mathbf{U}r} \equiv \sum_{i\in r}\mathbf{U}_i$ is unknown and must be estimated and its estimate is denoted by $\hat{\mathbf{T}}_{\mathbf{U}r}$.

The weights w_i are now a function of the sample values of $\mathbf{X}_i$ and $\mathbf{U}_i$ but also of the known population total $\mathbf{T}_\mathbf{X}$ and the estimated non-sample total of the contexual variables $\hat{\mathbf{T}}_{\mathbf{U}r}$. The weights depending on $\hat{\mathbf{T}}_{\mathbf{U}r}$ are denoted by $\hat{w}_i$. The EBLUP based on the estimated $\hat{\mathbf{T}}_{\mathbf{U}r}$ is denoted by $\hat{T}$.

Now the MSE can be expressed as follows

$$\begin{aligned} MSE[\hat{T}] &= E[(\hat{T} - T_Y)]^2 \\ &= (E(\hat{T} - T_Y))^2 + \mathrm{Var}(\hat{T} - T_Y) \\ &= (E[E(\hat{T} - T_Y|\hat{\mathbf{T}}_{\mathbf{U}r})])^2 + \mathrm{Var}[E(\hat{T} - T_Y|\hat{\mathbf{T}}_{\mathbf{U}r})] + E[\mathrm{Var}(\hat{T} - T_Y|\hat{\mathbf{T}}_{\mathbf{U}r})] \end{aligned} \tag{17}$$

Previously with known $\mathbf{T}_{\mathbf{U}r}$

$$\begin{aligned} E(\hat{T} - T_Y) &= \left(\sum_{i\in s} w_i\mathbf{X}_i - \sum_{i\in U}\mathbf{X}_i\right)\beta + \left(\sum_{i\in s} w_i\mathbf{U}_i - \sum_{i\in U}\mathbf{U}_i\right)\gamma \\ &= \left(\sum_{i\in s} w_i\mathbf{H}_i - \sum_{i\in U}\mathbf{H}_i\right)\boldsymbol{\lambda}. \end{aligned}$$

Now the conditional expectation given $\hat{\mathbf{T}}_{\mathbf{U}r}$ gives

$$E(\hat{T} - T_Y|\hat{\mathbf{T}}_{\mathbf{U}r}) = \left(\sum_{i\in s} \hat{w}_i\mathbf{X}_i - \sum_{i\in U}\mathbf{X}_i\right)\beta + \left(\sum_{i\in s} (\hat{w}_i - 1)\mathbf{U}_i - \hat{\mathbf{T}}_{\mathbf{U}r}\right)\gamma$$

The outer expectations/variances are always with respect to the distribution of $\hat{\mathbf{T}}_{\mathbf{U}r}$ and are usually suppressed, unless necessary. Now assuming that $\hat{\mathbf{T}}_{\mathbf{U}r}$ is an unbiased estimate of $\mathbf{T}_{\mathbf{U}r}$, equivalently $\hat{\mathbf{T}}_\mathbf{U}$ is unbiased estimate of $\mathbf{T}_\mathbf{U}$ with estimated (co)variance matrix $\mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}}$. It follows that the estimated weights $\hat{w}_i$ depending on $\hat{\mathbf{T}}_{\mathbf{U}r}$, see Equation (12), are also unbiased, that is $E(\hat{w}_i) = w_i$. First we obtain a variance estimate for $\hat{\mathbf{w}}$ using $\hat{\mathbf{T}}_{\mathbf{H}r} = (\mathbf{T}_{\mathbf{X}r}, \hat{\mathbf{T}}_{\mathbf{U}r})$

$$\begin{aligned} \mathrm{Var}(\hat{\mathbf{w}}) &= \mathrm{Var}\left(\mathbf{1}_s + \mathbf{D}_\mathbf{v}^{-1}\mathbf{H}_s\left\{\mathbf{H}_s^\top\mathbf{D}_\mathbf{v}^{-1}\mathbf{H}_s\right\}^{-1}\hat{\mathbf{T}}_{\mathbf{H}r}^\top\right) \\ &= \mathbf{D}_\mathbf{v}^{-1}\mathbf{H}_s\left\{\mathbf{H}_s^\top\mathbf{D}_\mathbf{v}^{-1}\mathbf{H}_s\right\}^{-1}\mathbf{V}_{\hat{\mathbf{T}}_\mathbf{H}}\left\{\mathbf{H}_s^\top\mathbf{D}_\mathbf{v}^{-1}\mathbf{H}_s\right\}^{-1}\mathbf{H}_s^\top\mathbf{D}_\mathbf{v}^{-1} \end{aligned}$$

with

$$\mathbf{V}_{\hat{\mathbf{T}}_\mathbf{H}} = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \end{pmatrix}.$$

Some blocks are zero because $\mathbf{T}_\mathbf{X}$ is known, hence there is no variability with respect to the distribution of $\hat{\mathbf{T}}_{\mathbf{U}r}$. By defining the $n \times (p+q)$ matrix $\mathbf{B} = \mathbf{D}_\mathbf{v}^{-1}\mathbf{H}_s\left\{\mathbf{H}_s^\top\mathbf{D}_\mathbf{v}^{-1}\mathbf{H}_s\right\}^{-1}$ and partitioning as

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_p & \mathbf{B}_q \end{pmatrix},$$

such that $\mathbf{B}_p$ is of dimension $n \times p$ and $\mathbf{B}_q$ of dimension $n \times q$. We can write now

$$\mathrm{Var}(\hat{\mathbf{w}}) = \mathbf{B}_q \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{B}_q^\top. \tag{18}$$

We obtain

$$\begin{aligned}
E(E(\hat{T} - T_Y | \hat{\mathbf{T}}_{\mathbf{U}r})) &= E\left\{ \left(\sum_{i \in s} \hat{w}_i \mathbf{X}_i - \sum_{i \in U} \mathbf{X}_i \right) \beta + \left(\sum_{i \in s} (\hat{w}_i - 1)\mathbf{U}_i - \hat{\mathbf{T}}_{\mathbf{U}r} \right) \gamma \right\} \\
&= \left(\sum_{i \in s} w_i \mathbf{X}_i - \sum_{i \in U} \mathbf{X}_i \right) \beta + \left(\sum_{i \in s} (w_i - 1)\mathbf{U}_i - \sum_{i \in r} \mathbf{U}_i \right) \gamma \\
&= \left(\sum_{i \in s} w_i \mathbf{H}_i - \sum_{i \in U} \mathbf{H}_i \right) \boldsymbol{\lambda}.
\end{aligned}$$

By ignoring terms that do not vary with respect to $\hat{\mathbf{T}}_{\mathbf{U}r}$ and collecting the remaining terms we obtain

$$\begin{aligned}
\mathrm{Var}(E(\hat{T} - T_Y | \hat{\mathbf{T}}_{\mathbf{U}r})) &= \mathrm{Var}\left\{ \left(\sum_{i \in s} \hat{w}_i \mathbf{X}_i - \sum_{i \in U} \mathbf{X}_i \right) \beta + \left(\sum_{i \in s} (\hat{w}_i - 1)\mathbf{U}_i - \hat{\mathbf{T}}_{\mathbf{U}r} \right) \gamma \right\} \\
&= \mathrm{Var}\left\{ \hat{\mathbf{w}}^\top \mathbf{H}_s \boldsymbol{\lambda} - \hat{\mathbf{T}}_{\mathbf{U}r} \gamma \right\} \\
&= (\mathbf{H}_s \boldsymbol{\lambda})^\top \mathrm{Var}(\hat{\mathbf{w}}) \mathbf{H}_s \boldsymbol{\lambda} + \gamma^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \gamma - 2\mathrm{Cov}(\hat{\mathbf{w}}^\top \mathbf{H}_s \boldsymbol{\lambda}, \hat{\mathbf{T}}_{\mathbf{U}r} \gamma) \\
&= (\mathbf{H}_s \boldsymbol{\lambda})^\top \mathbf{B}_q \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{B}_q^\top \mathbf{H}_s \boldsymbol{\lambda} + \gamma^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \gamma - 2\mathrm{Cov}(\hat{\mathbf{w}}^\top \mathbf{H}_s \boldsymbol{\lambda}, \hat{\mathbf{T}}_{\mathbf{U}r} \gamma) \\
&= \left(\mathbf{B}_q^\top \mathbf{H}_s \boldsymbol{\lambda} \right)^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{B}_q^\top \mathbf{H}_s \boldsymbol{\lambda} + \gamma^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \gamma - 2(\mathbf{H}_s \boldsymbol{\lambda})^\top \mathrm{Cov}(\hat{\mathbf{w}}, \hat{\mathbf{T}}_{\mathbf{U}r}) \gamma \\
&= \left(\mathbf{B}_q^\top \mathbf{H}_s \boldsymbol{\lambda} \right)^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{B}_q^\top \mathbf{H}_s \boldsymbol{\lambda} + \gamma^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \gamma - 2(\mathbf{H}_s \boldsymbol{\lambda})^\top \mathbf{B}_q \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \gamma \\
&= \left(\mathbf{B}_q^\top \mathbf{H}_s \boldsymbol{\lambda} \right)^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{B}_q^\top \mathbf{H}_s \boldsymbol{\lambda} + \gamma^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \gamma - 2\left(\mathbf{B}_q^\top \mathbf{H}_s \boldsymbol{\lambda} \right)^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \gamma \\
&= \left(\mathbf{B}_q^\top \mathbf{H}_s \boldsymbol{\lambda} - \gamma \right)^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \left(\mathbf{B}_q^\top \mathbf{H}_s \boldsymbol{\lambda} - \gamma \right)
\end{aligned}$$

Using $E(X-1)^2 a_i = (E(X)-1)^2 a_i + \mathrm{Var}(X) a_i$ for any r.v. X and constant a_i, we obtain

$$\begin{aligned}
E(\mathrm{Var}(\hat{T} - T_Y | \hat{\mathbf{T}}_{\mathbf{U}r})) &= E\left(\sum_{i \in s} (\hat{w}_i - 1)^2 \sigma_i^2 + \sum_{i \in r} \sigma_i^2 \right) \\
&= \sum_{i \in s} (w_i - 1)^2 \sigma_i^2 + \sum_{i \in r} \sigma_i^2 + \sum_{i \in s} \sigma_i^2 \mathrm{Var}(\hat{w}_i) \\
&= \sum_{i \in s} (w_i - 1)^2 \sigma_i^2 + \sum_{i \in r} \sigma_i^2 + (\sigma^2)^\top \mathrm{Diag}\left(\mathbf{B}_q \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{B}_q^\top \right),
\end{aligned}$$

where $\sigma^2 = \left(\sigma_1^2, \ldots, \sigma_n^2 \right)^\top$ and $\mathrm{Diag}(\mathbf{A})$ gives the vector on the diagonal of matrix $\mathbf{A}$.

Finally using that under the CN model $\mathbf{d}_\mathbf{H} = 0$ and using $\mathbf{b} = \mathbf{B}_q^\top \mathbf{H}_s \boldsymbol{\lambda} - \gamma = \left[(\mathbf{H}_s^\top \mathbf{H}_s)^{-1} \mathbf{H}_s^\top \mathbf{H}_s \right]_q \lambda - \gamma = \gamma - \gamma = 0$ ($[\mathbf{A}]_q$ refers to the last q rows of matrix $\mathbf{A}$,

we obtain

$$\begin{aligned} MSE[\hat{T}] &= \left\{ \left(\sum_{i \in s} w_i \mathbf{H}_i - \sum_{i \in U} \mathbf{H}_i \right) \boldsymbol{\lambda} \right\}^2 + \sum_{i \in s} (w_i - 1)^2 \sigma_i^2 + \sum_{i \in r} \sigma_i^2 \\ &\quad + \mathbf{b}^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{b} + (\sigma^2)^\top \mathrm{Diag}(\mathbf{B}_q \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{B}_q^\top) \\ &= \mathbf{d}_\mathbf{H} (\boldsymbol{\lambda} \boldsymbol{\lambda}^\top) \mathbf{d}_\mathbf{H}^\top + \sum_{i \in s} (w_i - 1)^2 \sigma_i^2 + \sum_{i \in r} \sigma_i^2 \\ &\quad + 0^\top \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} 0 + (\sigma^2)^\top \mathrm{Diag}(\mathbf{B}_q \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{B}_q^\top) \\ &= \sum_{i \in s} (w_i - 1)^2 \sigma_i^2 + \sum_{i \in r} \sigma_i^2 + (\sigma^2)^\top \mathrm{Diag}(\mathbf{B}_q \mathbf{V}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{B}_q^\top), \end{aligned} \tag{19}$$

which differs from (14) by the additional last term.

In practice $MSE[\hat{T}]$ can be estimated by

$$\widehat{MSE}[\hat{T}] = \sum_{i \in s} (\hat{w}_{Hi} - 1)^2 \sigma_i^2 + \sum_{i \in r} \sigma_i^2 + (\sigma^2)^\top \mathrm{Diag}\left(\mathbf{B}_q \hat{\mathbf{V}}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{B}_q^\top \right), \tag{20}$$

Now Δ can be estimated by

$$\begin{aligned} \hat{\Delta} &= -\mathbf{d}_\mathbf{X} (\hat{\boldsymbol{\lambda}} \hat{\boldsymbol{\lambda}}^\top - \widehat{\mathrm{Var}}(\boldsymbol{\lambda})) \mathbf{d}_\mathbf{X}^\top + \sum_{i \in s} (w_{Hi} - 1)^2 \hat{\sigma}_i^2 - \sum_{i \in s} (w_{Xi} - 1)^2 \hat{\sigma}_i^2 \\ &\quad + (\sigma^2)^\top \mathrm{Diag}\left(\mathbf{B}_q \hat{\mathbf{V}}_{\hat{\mathbf{T}}_\mathbf{U}} \mathbf{B}_q^\top \right). \end{aligned}$$

We propose to use the contextual model when $\hat{\Delta} < 0$ and the model without the contextual effects otherwise. When only one contextual variable is provided and $\sigma_i^2 = \sigma^2$, then the first three terms are replaced as in Clark and Chambers (2008) and we obtain

$$\hat{\Delta} = \hat{T}_{Ur}^2 \left(2\hat{\sigma}^2 S_u^{-1} - \hat{\gamma}^2 \right) + \hat{\sigma}^2 \mathbf{1}_s^\top \mathrm{Diag}\left(\mathbf{B}_q \mathbf{B}_q^\top \right) \hat{V}_{\hat{T}_U},$$

where the last term disappears when the total T_U is known, as then $\hat{V}_{\hat{T}_U} = 0$, and then the formula coincides with (16).

C. Estimation of Contextual Population Information

For notational convenience, we use 1_s as the vector of length n, $\mathbf{1}_r$ as the vector of length $N - n$, $\mathbf{1}_N$ as the vector of length N and similarly the matrices of ones $\mathbf{1}_{s,N}$, $\mathbf{1}_{r,N}$ $\mathbf{1}_{N,N}$ with appropriate sizes.

The contextual population information is often not available. As defined before $\tilde{\mathbf{X}}$ is a vector of population covariates, then the vector of contextual population information can be obtained by the formula

$$\mathbf{U} = \mathbf{W} \tilde{\mathbf{X}} \tag{21}$$

$$\mathbf{W} = \mathbf{Z} / \mathbf{Z} \mathbf{1}_{N,N}, \tag{22}$$

where the division refers to element-wise division of the matrices, it is an expression for dividing each row of $\mathbf{Z}$ by the number of friends of person i (the sum of row i). Often the number of friends is fixed, for example for the BHPS, each person has exactly three friends, then $\mathbf{W} = \frac{1}{3}\mathbf{Z}$.

Suppose that $\mathbf{U}$ is available for the sample, that is $\mathbf{U}_s$, and that the full $\mathbf{Z}$ is not available. We outline here a simple method to obtain $\hat{\mathbf{T}}_{\mathbf{U}}$ and $\hat{\mathbf{V}}_{\mathbf{T}_{\mathbf{U}}}$ under a simple ERGM and a simple estimator for $\hat{\mathbf{Z}}$.

The ERGM under consideration is outlined in Sections 4 and 6 and has an edges statistic and a GWESP statistic. One can show that (e.g., when simulating under this ERGM) that $P(Z_{ij} = 1) = p$ is a constant (irrespective of whether the network is undirected or directed), that is for each dyad the same marginal probability applies. However dyads are usually not independent.

Under the situation SS+SR we can estimate p by

$$\hat{p} = \frac{1}{n_s(N-1)} \sum_{i \in s} \sum_{j \neq i, j \in s, r} Z_{ij} = \mathbf{1}_s^\top \mathbf{Z}_s \mathbf{1}_N / (n_s(N-1)).$$

Under independence and from general properties of a Binomial random variable $\text{Var}(\hat{p}) = \hat{p}(1-\hat{p})/(n_s(N-1))$. Due to dependence the real variance will be larger (as the covariances between dyads are usually positive) and might be estimated by simulating under an ERGM. However from simulations, see for example Suesse (2012b) where a correlation matrix has to be estimated by simulating a large number of ERGM networks to fit a certain class of network models, it can also be shown that most correlations between dyads are near zero and that only some correlations for dyads sharing a node, for example Y_{ij} and Y_{ik}, are positive but small in magnitude, for example 0.02. Hence the correlation structure under the independence assumption (meaning that all correlations are exactly zero) does not deviate much from the true correlation structure and we expect the estimated variance derived under the binomial distribution to hold approximately.

Now we use $\hat{p}$ and the simple estimator $\text{Var}(\hat{p})$ to estimate $\mathbf{T}_{\mathbf{U}}$ and $\mathbf{V}_{\hat{\mathbf{T}}_{\mathbf{U}}}$. Using (21) and (22) we can write under $SS + SR$ using $\mathbf{Z}_r = (\mathbf{Z}_{rs}, \mathbf{Z}_{rr})$

$$\mathbf{T}_{\mathbf{U}} = \mathbf{1}_N^\top \mathbf{U} = \mathbf{1}_s^\top \mathbf{U}_s + \mathbf{1}_r^\top \mathbf{U}_r \tag{23}$$

and

$$\mathbf{U}_r = \frac{\mathbf{a}_1}{\mathbf{a}_2} = \frac{\mathbf{Z}_r \tilde{\mathbf{X}}}{\mathbf{Z}_r \mathbf{1}_N}.$$

Now $\mathbf{Z}_r$ is not observed and must be estimated, here by $\mathbf{P}_r = E(\mathbf{Z}_r)$. Note that $\mathbf{P}_r$ is a $(N-n) \times N$ matrix, but has zeros along the off-diagonal that correspond to the diagonal of $\mathbf{Z}$ because $\mathbf{Z}$ has zero diagonal entries. The other entries are constant and equal $\hat{p}$.

Now

$$\mathbf{a}_1 = \hat{\mathbf{P}}_r \tilde{\mathbf{X}} = \hat{p}\{N \mathbf{1}_r \bar{\tilde{\mathbf{X}}} - \tilde{\mathbf{X}}_r\}$$

with the population average of $\tilde{\mathbf{X}}$ denoted by $\bar{\tilde{\mathbf{X}}}$ and

$$\mathbf{a}_2 = \hat{\mathbf{P}}_r \mathbf{1}_N = \hat{p}(N-1)\mathbf{1}_r,$$

provided the number of friends is not fixed and unknown. If fixed, as for the BHPS, the network is not required to estimate p, because p is known and equals $\frac{3}{N-1}$, as each person has three friends in the population with $N-1$ possible candidates. Then $\mathbf{a}_2$ has simple structure, it is a vector with elements equal to this number, that is 3.

The estimator of $\mathbf{U}_r$ is

$$\hat{\mathbf{U}}_r = \frac{\hat{p}\{N\mathbf{1}_r\bar{\tilde{\mathbf{X}}} - \tilde{\mathbf{X}}_r\}}{\hat{p}(N-1)\mathbf{1}_r} = \frac{N\mathbf{1}_r\bar{\tilde{\mathbf{X}}} - \tilde{\mathbf{X}}_r}{(N-1)\mathbf{1}_r}$$

and hence an estimator $\hat{\mathbf{T}}_{\mathbf{U}}$ of $\mathbf{T}_{\mathbf{U}}$ is obtained by replacing $\mathbf{U}_r$ by $\hat{\mathbf{U}}_r$ in (23).

Assume for simplicity the denominator is fixed (as for the BHPS) and we have only one contextual variable, then we approximate $\mathbf{V}_{\hat{T}_U}$ by

$$\hat{\mathbf{V}}_{\hat{T}_U} = \frac{\{N\mathbf{1}_r\bar{\tilde{\mathbf{X}}} - \tilde{\mathbf{X}}_r\}^\top\{N\mathbf{1}_r\bar{\tilde{\mathbf{X}}} - \tilde{\mathbf{X}}_r\}}{\hat{p}^2(N-1)^2}\widehat{\mathrm{Var}}(\hat{p}) \tag{24}$$

Otherwise when the denominator is not fixed, then the variance of the ratio might be obtained by the delta method or by simulation under a network model or re-sampling methods, as the parametric bootstrap method. In the following we ignore the variability of the denominator and apply naively Equation (24).

9. References

Butts, C. 2008. "Network: A Package for Managing Relational Data in R." *Journal of Statistical Software* 24(2): 1–36. Doi: http://dx.doi.org/10.18637/jss.v024.i06.

Carrington, P., J. Scott, and S. Wasserman. 2005. *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.

Chambers, R., H. Chandra, and N. Tzavidis. 2011. "On Bias-Robust Mean Squared Error Estimation for Pseudo-Linear Small Area Estimators." *Survey Methodology* 37: 153–170. Available at: http://www5.statcan.gc.ca/olc-cel/olc.action?ObjId=12-001-X201100211604&ObjType=47&lang=en (accessed September 2017).

Chambers, R.L. and R.G. Clark. 2012. *An Introduction to Model-Based Survey Sampling with Applications*. Oxford: Oxford University Press.

Clark, R.G. and R.L. Chambers. 2008. "Adaptive Calibration for Prediction of Finite Population Totals." *Survey Methodology* 34: 163–172. Available at: http://www.statcan.gc.ca/pub/12-001-x/2008002/article/10757-eng.pdf (accessed September 2017).

Doreian, P., K. Teuter, and C.H. Wang. 1984. "Network Auto-Correlation Models – some Monte-Carlo Results." *Sociological Methods & Research* 13(2): 155–200. Doi: https://doi.org/10.1177/0049124184013002001.

Duke, J.B. 1993. "Estimation of the Network Effects Model in a Large Data Set." *Sociological Methods & Research* 21(4): 465–481. Doi: https://doi.org/10.1177/0049124193021004003.

Frank, O. and D. Strauss. 1986. "Markov Graphs." *Journal of the American Statistical Association* 81(395): 832–842. Doi: http://dx.doi.org/10.1080/01621459.1986.10478342.

Friedkin, N.E. 1990. "Social Networks in Structural Equation Models." *Social Psychology Quarterly* 53(4): 316–328.

Goldstein, H. 1989. "Restricted Unbiased Iterative Generalized Least-Squares Estimation." *Biometrika* 76(3): 622–623. Doi: https://doi.org/10.1093/biomet/76.3.622.

Handcock, M.S., A.E. Raftery, and J.M. Tantrum. 2007. "Model-Based Clustering for Social Networks." *Journal of the Royal Statistical Society* Series A 170: 301–322. Doi: http://dx.doi.org/10.1111/j.1467-985X.2007.00471.x.

Hunter, D., M. Handcock, C. Butts, S. Goodreau, and M. Morris. 2008b. "ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of Statistical Software* 24(2): 1–29. Doi: http://dx.doi.org/10.18637/jss.v024.i03.

Hunter, D.R. 2007. "Curved Exponential Family Models for Social Networks." *Social Networks* 29(2): 216–230. Doi: http://dx.doi.org/10.1198/106186006X133069.

Hunter, D.R., S.M. Goodreau, and M.S. Handcock. 2008a. "Goodness of Fit of Social Network Models." *Journal of the American Statistical Association* 103(481): 248–258. Doi: http://dx.doi.org/10.1198/016214507000000446.

Hunter, D.R. and M.S. Handcock. 2006. "Inference in Curved Exponential Family Models for Networks." *Journal of Computational and Graphical Statistics* 15(3): 565–583. Doi: http://dx.doi.org/10.1198/106186006X133069.

Koskinen, J., G. Robins, and P. Pattison. 2010. "Analysing Exponential Random Graph (p-star) Models with Missing Data Using Bayesian Data Augmentation." *Statistical Methodology* 7(3): 366–384. Doi: https://doi.org/10.1016/j.stamet.2009.09.007.

Leenders, R. 2002. "Modeling Social Influence Through Network Autocorrelation: Constructing the Weight Matrix." *Social Networks* 24(1): 21–47. Doi: https://doi.org/10.1016/S0378-8733(0100049-1).

Marsden, P.V. and N.E. Friedkin. 1993. "Network Studies of Social-Influence." *Sociological Methods & Research* 22(1): 127–151. Doi: https://doi.org/10.1177/ 00491 24193022001006.

Ord, K. 1975. "Estimation Methods for Models of Spatial Interaction." *Journal of the American Statistical Association* 70(349): 120–126. Doi: http://amstat.tandfonline.com/ https://doi/abs/10.1080/01621459.1975.10480272.

Pattison, P., G. Robins, T. Snijders, and P. Wang. 2013. "Conditional Estimation of Exponential Random Graph Models from Snowball Sampling Designs." *Journal of Mathematical Psychology* 57(6): 284–296. Doi: https://doi.org/10.1016/j.jmp.2013.05.004.

Royall, R.M. 1976. "Linear Least-Squares Prediction Approach to 2-Stage Sampling." *Journal of the American Statistical Association* 71(355): 657–664. Doi: http://dx.doi.org/10.2307/2285596.

Särndal, C., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer series in statistics. New York: Springer-Verlag.

Snijders, T. 2002. "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models." *Journal of Social Structure* 1–40. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.20.5323&rep=rep1&type=pdf (accessed September 2017).

Snijders, T., P. Pattison, G. Robins, and M. Handcock. 2006. "New Specifications for Exponential Random Graph Models." *Sociological Methodology* 36: 99–153. Doi: http://dx.doi.org/10.1111/j.1467-9531.2006.00176.x.

Strauss, D. and M. Ikeda. 1990. "Pseudolikelihood Estimation for Social Networks." *Journal of the American Statistical Association* 85(409): 204–212. Doi: http://dx.doi.org/10.2307/2289546.

Suesse, T. 2012a. "Estimation in Autoregressive Population Models." In Proceedings of Fifth Annual ASEARC Research Conference, 11–14. Applied Statistics and Research Collaboration (ASEARC). 2–3 February 2012, Wollongong, Australia. Available at: http://eis.uow.edu.au/asearc/5thAnnResCon/index.html (accessed September 2017).

Suesse, T. 2012b. "Marginalized Exponential Random Graph Models." *Journal of Computational and Graphical Statistics* 21(4): 883–900. Doi: http://dx.doi.org/10.1080/10618600.2012.694750.

Suesse, T. and R. Chambers. 2012. *Using Social Network Information for Survey Estimation*. Report 11/12, National Institute of Applied Statistics Research Australia, University of Wollongong. Available at: http://niasra.uow.edu.au/content/groups/public/@web/@inf/@math/documents/doc/uow137689.pdf (accessed September 2017).

Suesse, T. and R. Chambers. 2014. *Using Social Network Information for Survey Estimation*. Report 13/14, National Institute of Applied Statistics Research Australia, University of Wollongong. Available at: http://niasra.uow.edu.au/content/groups/public/@web/@inf/@math/documents/mm/uow182447.pdf (accessed September 2017).

Suesse, T. and A. Zammit Mangion. 2017. "Computational Aspects of the em Algorithm for Spatial Econometric Models with Missing Data." *Journal of Statistical Computation and Simulation* 87: 1767–1786. Doi: http://dx.doi.org/10.1080/00949655.2017.1286495.

Thompson, S. and G. Seber. 1996. *Adaptive Sampling*. Wiley Series in probability and mathematical statistics. New York: Wiley.

Wasserman, S. and K. Faust. 1994. *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

Received July 2015
Revised July 2017
Accepted September 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 211–237, http://dx.doi.org/10.1515/JOS-2018-0010

An Analysis of Interviewer Travel and Field Outcomes in Two Field Surveys

James Wagner[1] *and Kristen Olson*[2]

In this article, we investigate the relationship between interviewer travel behavior and field outcomes, such as contact rates, response rates, and contact attempts in two studies, the National Survey of Family Growth and the Health and Retirement Study. Using call record paradata that have been aggregated to interviewer-day levels, we examine two important cost drivers as measures of interviewer travel behavior: the distance that interviewers travel to segments and the number of segments visited on an interviewer-day. We explore several predictors of these measures of travel – the geographic size of the sampled areas, measures of urbanicity, and other sample and interviewer characteristics. We also explore the relationship between travel and field outcomes, such as the number of contact attempts made and response rates. We find that the number of segments that are visited on each interviewer-day has a strong association with field outcomes, but the number of miles travelled does not. These findings suggest that survey organizations should routinely monitor the number of segments that interviewers visit, and that more direct measurement of interviewer travel behavior is needed.

Key words: Interviewer travel; survey costs; nonresponse; paradata.

1. Introduction

In face-to-face surveys, survey organizations use multi-stage area probability samples with clustering of sampled housing units (i.e., 'area segments,' Kish 1965) in order to constrain travel costs, but we know very little about interviewer travel behavior in these surveys. In general, an interviewer travels from his/her home to a sampled neighborhood. Once in the sampled neighborhood, the interviewer makes contact attempts to several sampled housing units, identifying potentially eligible respondents ("screening") and conducting interviews.

[1] University of Michigan, Institute for Social Research, 426 Thompson St. Room 4050, Ann Arbor, MI 48104, U.S.A. Email: jameswag@umich.edu
[2] University of Nebraska-Lincoln, Department of Sociology, 703 Oldfather Hall, Lincoln, NE 68588, U.S.A. Email: kolson5@unl.edu

Acknowledgments: An earlier version of this article was presented at the Joint Statistical Meetings, August 2011, Miami Beach, FL. The authors thank the NSFG and the HRS projects for access to these data, and Brady West, four anonymous reviewers and the Associate Editor for comments on an earlier draft. In particular, the Associate Editor gave us the insight that traveling longer distances at higher speeds (e.g., on a freeway) might require less time than a shorter distance at a slower speed (e.g., city street). The NSFG 2006–2010 was conducted by the Centers for Disease Control and Prevention's CDC's National Center for Health Statistics (NCHS), under contract #200-2000-07001 with University of Michigan's Institute for Social Research with funding from several agencies of the U.S. Department of Health and Human Services, including CDC/NCHS, the National Institute of Child Health and Human Development (NICHD), the Office of Population Affairs (OPA), and others listed on the NSFG webpage (see http://www.cdc.gov/nchs/nsfg/). The views expressed here do not represent those of NCHS or the other funding agencies. The HRS is funded by the National Institute on Aging (U01 AG009740), with supplemental support from the Social Security Administration.

If a sampled area is quite far from the interviewer's home, large amounts of time are spent 'on the clock' for travel, not screening, interviewing, or otherwise recruiting sampled units. As a result, interviewer travel is a large component of the total budget in personal visit surveys; for many large national surveys, travel costs are 20% to 40% of the total survey budget (e.g., Judkins et al. 1990; Kalsbeek et al. 1994; Weeks et al. 1983; Sudman 1965–66; Sudman 1967). Interviewers may vary in their ability to plan efficient travel which may impact their efficiency and other field outcomes. Yet surprisingly unexplored are characteristics of interviewer travel behavior – measured either through the actual number of miles traveled or the number of area segments visited on a given interviewer workday – and the association between interviewer travel behavior and field outcomes, such as the number of contact attempts, contact and cooperation rates. These interviewer travel behaviors and field outcomes are important drivers of both response rates and costs.

In this article, we examine this important but little explored component of field surveys – interviewer travel behavior – in two large national surveys in the United States. We also investigate the relationship between the distance that interviewers travel while performing their work and field outcomes – that is, the cumulative results of an interviewer's actions when visiting sampled housing units in the field. In particular, we examine the following two questions:

1. What are predictors of interviewer travel behavior in two large national US field surveys?
2. Is interviewer travel behavior associated with field outcomes (i.e., number of contact attempts, contact rates, screening, and main interview cooperation rates)?

This article takes a first observational look at the relationship between characteristics of areas, travel-related costs, and survey errors. We will examine these relationships using cross-classified random effects models to analyze interviewer travel and field outcomes as indicated through call history and timesheet data from two large-scale national area probability sample surveys in the United States – the National Survey of Family Growth and the Health and Retirement Study. Specifically, for our first research question, we examine predictors of interviewer travel such as the geographic size of sampled areas and urbanicity. For our second research question, we examine whether interviewer travel decisions contribute to survey errors, including variability in nonresponse rates across interviewers and decisions that may lengthen the field period. We anticipate that interviewers who travel more miles will also have more contact attempts because the travel allows them to make these contact attempts, but lower contact and response rates because the time spent traveling constrains the amount of time that can be spent administering a questionnaire. Understanding the relationship between travel and field outcomes will aid survey practitioners in designing efficient surveys.

2. Background

2.1. Interviewer Travel Overview

Although all field studies require interviewers to travel to sampled housing units, there is surprisingly little empirical examination of interviewer travel behavior. Travel in sample

surveys has been specified as a constraint on sampling error, such as the variance of an estimated mean from a cluster sample (Sudman 1978; Hansen et al. 1953, 274). There are different measures of interviewer travel behavior available from paradata, including how many segments an interviewer visits on a given day or the distance traveled on these trips. Paradata include call record data (sometimes called contact history data) and timesheet and travel expense data. Cost models indicate three inputs that are needed – (1) the number of times a segment is visited, (2) the distance between the interviewer's home and the segments, and (3) the distance traveled within segments (Sudman 1967, ch. 2; Judkins et al. 1990). For example, Cochran (1977, 183) provides a cost model for a cluster sample with cost of travel between clusters as $C = c_1 n + c_t\sqrt{n} + c_2 nm$, where n is the number of clusters, m is the number of units within each cluster, c_1 is the cost of measuring (an unspecified combination of listing, screening, and interviewing) a cluster, c_t is the cost related to travel between clusters, and c_2 is the cost of travel within a cluster (see also Kalsbeek et al. 1983; Judkins et al. 1990). Despite this central role of cost models for determining optimal sample designs, empirical evaluation of what these cost inputs are in actual field studies is lacking.

Existing cost data are largely for one survey (the National Household Interview Survey, or NHIS, conducted for the US National Center for Health Statistics). Kalsbeek et al. (1983) note that empirical data are not available and, therefore, attempt to use simple geometry to model travel costs. In examining cost drivers in the 1988 NHIS, Judkins and colleagues (1990) note that there is only limited empirical data available on the number of segments visited and on distance traveled. They find that the average workload size for interviewers was 2.3 segments, but were not able to empirically evaluate mileage traveled. Chen (2012) draws on data about field outcomes from the 2004 NHIS in order to simulate travel behaviors. Thus, there is a very limited literature with almost no empirical data on travel costs as they relate to sampling error. There is even less information about how these travel behaviors are related to field outcomes.

Further, the manner in which interviewers make decisions about travel is not well understood. Little information on interviewer travel can be obtained from training materials, as interviewers receive limited training on travel; the training that they do receive generally focuses on cost containment. For example, interviewers are trained to monitor (Mayer 1968) or limit the number of trips to sampled segments to keep costs under control (Morton-Williams 1993, 141; Campanelli et al. 1997, 3–20). They are also trained to make contact attempts at times during which they are likely to reach someone at home, varying the day and time of contact attempts (e.g., Morton-Williams 1993; Stoop et al. 2010), times which may vary across sample units (e.g., Durrant and Steele 2009; Blom 2012). In practice, interviewers report visiting multiple cases once they have traveled to a particular sampled area (Peachman 1992). Interviewers may then travel between segments to reach different sampled persons at home, even revisiting area segments on the same day. The extent to which they do this is unknown.

One potential reason that interviewer travel has received limited attention is that studying interviewer travel outcomes requires a record to be kept of travel itself. In most in-person surveys, this can be obtained in three ways: (1) mileage reports from the interviewer when asking for travel reimbursement on their timesheets, (2) distances obtained from geocoding the locations of the sampled housing units as recorded in the call

records, or (3) summaries of the number of sampled neighborhoods (segments) visited by each interviewer obtained through aggregating the call records or reports by the interviewer themselves. All three sources are likely to contain errors. Nevertheless, evaluating existing travel information is necessary to establish that there are associations between travel behaviors and field outcomes such as contact and cooperation rates, even with imperfect data. If interviewers are unsuccessful at establishing contact or completing interviews, then they will be more likely to spend time in travel as they continue to seek contact with a selected housing unit. In this article, we will focus on mileage reports and a summary of the number of segments visited on a given interviewer-day obtained through the call records.

2.2. *Predictors of Interviewer Travel Behavior*

The first research question addresses predictors of interviewer travel behavior. Interviewers are not randomly assigned to segments, and their skill sets with respect to travel decisions vary. We expect that travel will differ overall for different geographic areas, interviewers with different levels of experience, and at different times during the field period.

Given higher population density in urban areas, *we anticipate that interviewers will have to travel more miles in rural areas than in suburban or urban areas because the segments are larger in rural areas*. Thus, we include urbanicity in all of the models. *We expect that larger areas and areas considered to be "more difficult" will require more travel*. Therefore, we include the area (measured in square miles) of the primary sampling units to account for the variation in area between Primary Sampling Units (PSUs). In order to account for variation between PSUs in the difficulty of obtaining an interview, we use the United States Census Bureau's "Hard-to-Count" score. This score is based upon characteristics of Census Tracts, including demographic characteristics predictive of response rates to the mailed portion of the Decennial Census (Bruce and Robinson 2006). We have created a weighted average, using the Census Tract count of housing units as the weight of this score for each PSU in the sample. We also anticipate that interviewers with less experience will travel differently than experienced interviewers (as interviewers with different experience levels have different calling times, Campanelli et al. 1997). In particular, *we expect that inexperienced interviewers will be more likely to follow travel suggestions received in training or from supervisors than experienced interviewers*. As such, in surveys where interviewers are instructed to visit all active, sampled housing units in a particular segment, *we expect inexperienced interviewers will be more likely to visit all active, sampled housing units, and thus do more within-segment traveling*. As such, we account for interviewer experience in all of the models as well.

Finally, we expect differences in travel overall for different times in the field period. In surveys where low yield segments are removed (either via two-phase sampling or by a management decision) from the active sample, *we expect that interviewers will tend to travel less at later times in the field period*.

2.3. *Association Between Interviewer Travel Behavior and Field Outcomes*

Decisions about travel affect not only costs, but also are likely to be associated with field outcomes. This may be either because of interviewers traveling in order to make more call

attempts when early attempts do not yield interviews or because the amount of time spent traveling constrains the time possible to conduct an interview. Thus, our second research question addresses whether interviewer travel behavior is associated with field outcomes such as the number of contact attempts, contact rates, and response rates.

We are aware of only two simulation studies that address this question. One of these simulation studies found that smaller PSU sizes give interviewers more time to contact households because of reduced travel time (Chen 2012). The other simulation study suggested that the length of interview may modify the effect that travel has on field outcomes in that shorter interviews yield more time for travel and increase the possibility of obtaining additional interviews (Bienias et al. 1990). Although interviewers are a key source of variability in response, contact and cooperation rates in face-to-face surveys (e.g., O'Muircheartaigh and Campanelli 1999; Campanelli et al. 1997; Purdon et al. 1999; Pickery and Loosveldt 2002; Durrant and Steele 2009; Blom et al. 2011; Blom 2012; Stoop et al. 2010), to our knowledge, no previous study has examined the association between interviewer travel behaviors and contact and cooperation rates using actual data.

The two previous simulation studies of interviewer travel (e.g., Bienias et al. 1990; Chen 2012) suggest *two competing hypotheses* about the relationship between travel behaviors and field outcomes. Both hypotheses assume that interviewers make their own decisions about which sampled segments to visit and which housing units to attempt on a given interviewer-day, that interviewers no longer visit sampled cases once an interview is completed, and that there are no constraints put on interviewers for the number of attempts that they can make on a given day or to a sampled housing unit. The first hypothesis posits a positive association between travel and the number of contact attempts made on a given day – interviewers who drive more miles or visit more segments are traveling to make more contact attempts. The second hypothesis anticipates a negative association between travel and contact attempts, arguing that interviewers who spend more time travelling – and thus have more miles driven or more segments visited – have less 'on the ground' time to make contact attempts for a fixed amount of work time in a given day.

We anticipate that the association between travel and the number of contact attempts will differ for different types of field outcomes. With the same assumptions as above, in particular, *we expect a positive association between travel behaviors and the number of contact attempts made on an interviewer day*. That is, interviewers who travel more will be doing so in order to make additional contact attempts. *We anticipate, however, that travel behaviors will be negatively associated with screening and response rates*. Interviewers who make contact with a household or obtain interviews will have less time to travel to other areas (that is, more successful field outcomes lead to less travel).

Because interviewer training on travel is generally linked to minimizing the number of trips made to individual segments (sampled neighborhoods), *we anticipate that the total number of segments visited will be more likely to be associated with field outcomes than the actual number of miles traveled*. Yet costs are directly related to the total number of miles that an interviewer drives. We thus will examine both measures of interviewer travel.

3. Data

3.1. Surveys

We examine two large-scale, national face-to-face surveys conducted by the Survey Research Center at the University of Michigan – the National Survey of Family Growth (NSFG) and the Health and Retirement Study (HRS). The two surveys differ in scope, target populations, and field periods. Additionally, the NSFG is a cross-sectional survey and the HRS is a longitudinal survey.

3.2. National Survey of Family Growth

The NSFG 2006–2010 Continuous was carried out under a contract with the US Center for Disease Control and Prevention's (CDC) National Center for Health Statistics. The NSFG collects information about fertility, childbearing, and sexual behaviors among women and men in the US aged 15 to 44. Thus, the NSFG interview process had two steps – identifying an age-eligible respondent through screening and conducting a "main" survey interview. NSFG 2006–2010 continuous released fresh samples quarterly, with a twelve-week field period, in what is called a "continuous sample design" (Lepkowski et al. 2010). Interviewers are generally assigned three neighborhoods or segments within a single, "home" PSU per quarter. The twelve-week field period is divided into two phases. At the end of ten weeks, a subsample of two segments (with additional subsampling of lines within those two segments) is selected for each interviewer. The data analyzed for this article come from the second quarter which ran from September to December of 2006. The initial sample included 5,063 housing units.

The assignment of interviewers to sampled segments is not random; in most PSUs there is only a single interviewer who is assigned a random sample of segments from within the PSU. When there is more than one interviewer, the interviewers are assigned segments near their home location to try to minimize travel. Overall, 45% of interviewers visited only one PSU over the field period, and 55% visited two or more PSUs. Additionally, 29% of PSUs had only one interviewer visit them, with the remaining 71% having at least two different interviewers visit them. Thus, interviewer-days are cross-classified between interviewers and PSUs. In some PSUs, including all of Alaska and Hawaii, interviewers are flown in for field work and do not travel from their home location. As such, we will exclude Alaska and Hawaii from these analyses.

The number of days worked by each of the 41 interviewers ranges from 2 (one interviewer) to 72, with only three interviewers working fewer than ten days, and an average of 44.6 days. On 97% of the days, interviewers remained within a single PSU.

All of the NSFG field staff record call records and timesheets electronically. Information about the housing units each interviewer visits and the corresponding field outcomes are recorded in call records. Call record information is usually filled out daily or more frequently (after each contact attempt). Travel information is recorded by the interviewer in their timesheets for purposes of mileage reimbursement in personal vehicles, and is submitted at most daily (the interviewers are paid biweekly, so payment information is not always submitted on the day that it occurred). The two systems, however, do not interface directly. That is, interviewers are not probed in their timesheets

to record time specifically related to effort recorded in the call records. As a result, there can be a "mismatch" between activities as recorded in the call records and activities as recorded in timesheets and expense reports. We will examine information from both the call records (aggregated to the number of segments visited per day) and the timesheets (reported mileage per day) in this analysis. For the data used in our models, each record represents a summary of the travel and field outcomes for a day that an interviewer worked.

3.3. Health and Retirement Study

The Health and Retirement Study (HRS) is a panel study of U.S. adults aged 50 and above, with initial data collection starting in 1992. Every six years, a new age-eligible sample is recruited to include the newly aged-in 51 to 56 year olds. In this article, we examine the new cohort added during the 2004 data collection. The HRS uses an area probability sample to identify households with newly age-eligible adults (Heeringa and Connor 1995; Health and Retirement Study 2008), with two interviewing steps (screening households for age-eligible respondents and "main" interviewing) for the new cohort. The sample in all PSUs and segments (40,120 housing units) was released at the beginning of the field period. Unlike the NSFG, the HRS did not select a second phase sample of segments. Near the end of the field period, with only 790 sample housing units still active, a subsample of half of these housing units was selected.

The panel and the new cohort of the HRS are recruited using somewhat overlapping field staff at the beginning of data collection. We examine here the effort of the interviewers who were mainly assigned to screen households and identify persons who were eligible for the new cohort and to interview them. Among these interviewers, 91.7% of trips had no contact attempts to the panel cases sample, and 5.2% had only one contact attempt to the panel cases. As such, although the HRS is a longitudinal study, the data that we examine here are analogous to a cross-sectional study.

In the HRS, multiple interviewers are (not randomly) assigned to a PSU, and each interviewer was assigned to a variable number of secondary sampling units (or segments) based on their geographic proximity to the segments, again having a cross-classified data structure for interviewer-days, nested across both interviewers and PSUs. In general, HRS field managers had more flexibility in assigning segments to interviewers than managers for the NSFG. For the HRS, there are generally more interviewers in each PSU. Segments could be assigned to the interviewer who lived close to them. Overall, each PSU had at least two interviewers work in it, ranging from two to 35 different interviewers in a single PSU. Additionally, 45% of interviewers worked in only one PSU, but the remaining 55% worked in two or more different PSUs. The field period for the HRS was about twelve months. Further, instead of having new segments assigned each quarter, HRS interviewers had all of their segments available from the beginning of the field period. There was much greater variability in the number of days worked by HRS interviewers compared to the NSFG interviewers, ranging from one day to 255 days, with an average of 48.8 days. To increase the stability of estimates (Olson and Peytchev 2007), only interviewers who had worked at least ten days were included in the analyses. This excluded 61 interviewers who had collectively worked a total of 176 days, 37 of whom had worked only one day during

the field period, with 205 interviewers remaining who had collectively worked 12,940 days, for an average of 62.2 interviewer-days per interviewer.

As with the NSFG, the data of interest for the HRS also come from timesheet information and from the call records. Both timesheets and call records are maintained electronically, and call records are filled out daily or more often. As in the NSFG, interviewers are paid biweekly. As with the NSFG, information about the number of miles traveled comes from the timesheet information; field outcomes and the number of segments visited on each interviewer day come from the call records. The models for the HRS are estimated on data for which each record represents a summary of the travel and field outcomes for a day worked by an interviewer.

4. Methods

4.1. Travel Variables

Ideally, the level of analysis for each interviewer and sampled unit for this study would be at the call attempt (sometimes called contact attempt) level. Each sampled address visited during a given trip to a sampled segment could be geocoded and the distance between the interviewer's house, the first sampled address, and among sampled addresses could be calculated. However, the timesheet data are kept at the day level for each interviewer, not at the contact attempt level. Thus, we aggregate information for each interviewer for each day of the survey period, and analyze travel at an interviewer-day level.

We use two measures of interviewer travel behavior: (1) the total number of miles reported per day in the timesheet and (2) the total number of trips made to sampled segments visited during an interviewer-day as calculated from the call records. First, we use the total number of miles traveled on a given day that the interviewer reported in their timesheet for purposes of reimbursement. In the NSFG, over 90% of days (90.9%) have mileage data reported, and in the HRS, 96.2% of days have mileage data reported. Among the days with mileage data reported, interviewers travel an average of 85.4 miles per interviewer-day in the NSFG and 53.4 miles per interviewer day in the HRS.

Second, to obtain a count of the number of trips made to segments during a day, we define trips as any visit to a segment that involves travel between the interviewer's home and the segment, or between segments. For example, if an interviewer travels to segment A from their home, then goes to segment B, and then returns to segment A, they have taken three trips to sampled segments. For parsimony, we will refer to this as a "three segment" trip. Interviewers in the NSFG and HRS visited a similar number of segments per interviewer day – 1.85 segments per interviewer-day in the NSFG and 1.91 segments per interviewer-day in the HRS. The distribution of number of segments visited per interviewer-day is remarkably similar across the two surveys. In both surveys, roughly 54% of all interviewer-days are spent visiting one segment, and just fewer than 22% of interviewer-days involve contact attempts to two segments. The balance of interviewer-days (24%) had contact attempts in more than two segments. In the NSFG, the correlation between the number of segments visited and the total number of miles traveled is 0.23 ($p < .0001$), whereas it is only 0.06 ($p < .0001$) in the HRS. These differences likely reflect the greater flexibility that HRS has in assigning segments to interviewers.

One possible limitation is the quality of the travel information. Discrepancies between the number of segments calculated from the call records and miles reported by the interviewer on a timesheet can occur for a number of reasons. For example, interviewers can record work-related travel – such as travel to a training session – that is not related to field effort, interviewers who are flown into a sampled PSU will not record mileage because they do not need to be reimbursed for rental car mileage, and interviewers who work in major metropolitan areas may not use a car to travel among sampled units, instead using public transportation. Additionally, interviewers may fail to complete a call record for some types of travel, such as driving by a house and not seeing evidence of anyone at home, or may not complete their travel reports until the end of the day, potentially forgetting a trip or misremembering where they traveled. Interviewers may also enter travel reimbursement information for the wrong date (Wang and Biemer 2010; Biemer et al. 2013; Wagner et al. 2013). Another type of error occurs when the wrong mode for a contact attempts is entered into the record. If a telephone visit is entered into the records as a face-to-face call, then incorrect assumptions about travel will be made.

To address the potential limitation of the quality of the mileage reported in the timesheets, for the NSFG, we also conducted a sensitivity analysis. We used information about the interviewer's home address to measure distance traveled on a given day starting from an interviewer's house via geocoded addresses obtained in the call records. We geocoded each interviewer's home address and the centroid of the sampled segment, and calculated the distance in miles 'as the crow flies' from the interviewer's home address to their segments and among their segments. Although other options are available for calculating distance (such as 'best routes' calculated through Google Maps), we started with this approach for simplicity. Given missing data on interviewer home addresses and limitations of the geocoding software, we were unable to geocode addresses for 24.6% of interviewer-days in the NSFG. The two sources of travel data are highly, but not perfectly correlated ($r = 0.780$, $p < .0001$). In general, the interviewer-reported timesheet data tends to be higher than the geocoded data – as would be expected given our use of the 'as the crow flies' distance for the geocoded data. We conducted all of our analyses using both metrics and came to identical conclusions. In the HRS, we do not have the interviewer's home address and thus cannot calculate a geocoded distance from the interviewer's home.

4.2. Field Outcome Variables

We examine five field outcomes (see Table 1) as the dependent variables in our models:

1. Total number of contact attempts made to sample housing units on a given interviewer-day for screening cases,
2. total number of contact attempts made for main interviews,
3. contact rates,
4. screener interview rates, and
5. main interview rates.

In both surveys, we separate contact attempts into two groups – contact attempts to screen the household for an eligible sample person and contact attempts to complete the main

Table 1. Descriptive Statistics of Travel, Field Outcomes, Urbanicity and Experience across Interviewer-Days, for the National Survey of Family Growth (NSFG) and Health and Retirement Study (HRS).

Variable	NSFG Mean (SD) or %	HRS Mean (SD) or %
Travel		
Number of miles	85.42 (74.55)	53.36 (50.08)
Number of segments	1.85 (1.18)	1.91 (1.22)
1 segment	54.9%	53.5%
2 segments	21.9%	21.5%
3 segments	12.2%	12.3%
4 segments	5.0%	6.1%
5+ segments	6.1%	6.7%
Day-level Field Outcomes		
Number of screening contact attempts	10.68 (11.58)	11.02 (11.25)
Number of main interview contact attempts	4.08 (4.03)	1.43 (2.08)
Contact rates	0.43 (0.30)	0.36 (0.29)
Screener interview rates	0.21 (0.23)	0.22 (0.25)
Main interview rates	0.30 (0.34)	0.20 (0.36)
Urbanicity		
Largest MSAs	17.2%	–
Smaller MSAs	41.2%	–
Non-MSAs	41.6%	–
Self-representing PSUs	–	34.4%
Non self-representing PSUs	–	65.6%
Interviewer Experience		
No experience	9.4%	23.9%
Any prior experience	90.6%	76.1%
Number of weeks in field period	6.51 (3.35)	28.83 (11.91)
Census Hard-to-Count Score	34.22 (12.16)	40.67 (19.40)
Area (Square miles)*	2,200 (1,464)	2,200 (3,378)

Note: Number of interviewer-days = 1,784 in the NSFG and 12,940 in the HRS. SD = Standard Deviation; MSA = Metropolitan Statistical Area; PSU = Primary Sampling Unit.
*Rounded to nearest hundreds.

interview. For both of these surveys, screening and main interviews are two separate activities that usually occur on different days (75.7% and 90.7% of completed screening interviews for the NSFG and HRS respectively required additional main contact attempts on another day). This is because the person completing the screening interview may be different than the person selected to complete the main interview and because the main interviews for both surveys take a relatively long time to complete. Therefore, interviews need to be scheduled at times that are convenient for the sampled person. In both surveys, there are fewer main interview attempts than screener interview attempts due to the length of the main interview and the eligibility criteria that had to be met to conduct a main interview. In the NSFG, there is an average of 10.68 screener attempts and 4.08 main interview attempts per interviewer-day. In the HRS, there is an average of 11.02 screener attempts and 1.43 main interview attempts per interviewer-day. While the average number of screener attempts is very similar across the two surveys, the number of main attempts is much higher for the NSFG. This is likely due to the higher eligibility rate for the NSFG

compared to the HRS, and that the HRS often attempts to interview two members of the household on the same day. The same number of screener attempts will produce more main attempts. Since these are averages across days, it also indicates that NSFG interviewers may have worked longer shifts.

The contact, screener interview, and main interview rates are calculated at an interviewer-day level. In particular, the contact rate is the total number of attempts with contact divided by the total number of attempts made on a given interviewer-day. The screener interview rates are the total number of completed screeners divided by the total number of attempts for screener interviews (those to cases of previously unknown eligibility), and the main interview rates are similarly the total number of completed main interviews divided by the total number of attempts for main interviews (those to cases of known eligibility). In the NSFG, the average contact rate across all interviewer-days is 43.2%, and the average contact rate for the HRS is 35.5%. Over all of the interviewer days, in the NSFG, the average screener interview rate is 21.3%, and the main interview rate, conditional on known eligibility is 29.5%. In the HRS, the screener rate is 21.8%, and the main interview rate among known eligible persons is 20.4%. These call-level contact, screener and main interview rates are different from the final case-level contact rate, screener and main interview rates for the survey which are calculated by the total number of cases contacted or completed by the end of the study divided by the total number of sampled (eligible) cases.

4.3. Predictor Variables for Multivariate Models

The predictors are chosen from a set that have been shown or are hypothesized to be related to interviewer travel and productivity. Since travel may be impacted by characteristics of the interviewer, the PSU, and characteristics of the available sample, we select predictors from our model from each of these areas. We model the number of segments visited and miles travelled each day, including as predictors in the models urbanicity, interviewer experience, a continuous measure of the week in the field period, a measure of how difficult the area is to enumerate based on the 2000 Decennial Census Hard-to-Count (HTC) score (Bruce and Robinson 2006), and the area of the PSU in square miles. Urbanicity is measured from characteristics of the PSU in which an interviewer's segments are located. In the NSFG, urbanicity has three levels – the largest Metropolitan Statistical Areas (MSAs, 17.2% of interviewer-days), smaller MSAs (41.2%) and non-MSAs (41.6%). A Metropolitan Statistical Area is an urban geographic area with at least 50,000 residents (United States Census Bureau 2013). In the HRS, we use an indicator for whether the PSU was self-representing or not; 34.4% of interviewer-days occurred in self-representing PSUs. In both surveys, interviewers were quite experienced – 90.6% of NSFG interviewers and 76.1% of HRS interviewers had prior interviewing experience. The week of the field period ranged from 1 to 12 for the NSFG and from 1 to 52 for the HRS. We include an indicator for whether the interviewer-day was in Phase 1(= 1) versus Phase 2(=0), and an interaction term between Phase 1 and week in the field period to account for potential nonlinearities during "close-out" periods of each survey. In the NSFG, Phase 1 was defined as all weeks before the 11th week, and in the HRS, Phase 1 was defined as all weeks before

the 48th week. The mean Census hard-to-count score was 34.22 (SD = 12.16) in the NSFG and 40.67 (SD = 19.40) in the HRS. Finally, the average PSU was about 2,200 square miles (SD = 1,464) in the NSFG and also about 2,200 square miles (SD = 3,378) in the HRS.

4.4. *Analysis Methods*

As described above in Subsections 3.2 and 3.3, interviewer-days are cross-classified within interviewers and PSUs in both the NSFG and HRS. Thus, we use cross-classified multilevel regression models, with interviewer-days nested within interviewers and PSUs, to examine the association between field effort and interviewer travel. In all models, we test whether a random effects model is needed using a likelihood ratio test that is a mixture of chi-squared distributions (West et al. 2015, 107).

4.4.1. Models for Travel

For the first research question, we examine predictors of the number of miles traveled and the number of segments visited.

Model for Miles

First, in order to predict the number of miles traveled in the NSFG we estimate a hierarchical linear regression model with a normal distribution and identity link function:

$$\begin{aligned} 1\text{pc}miles_{i(jk)} = {} & \theta_0 + \beta_1 segments_{i(jk)} + \beta_2 SmallMSA_k + \beta_3 NonMSA_k + \beta_4 CensusHTC_k \\ & + \beta_5 AreaSqMiles_k + \beta_6 AnyExp_j + \beta_7 Week_{i(jk)} + \beta_8 Phase1_{i(jk)} \\ & + \beta_9 Phase1_{i(jk)} Week_{i(jk)} + b_{00j} + c_{00k} + e_{ijk} \end{aligned}$$

where i represents interviewer-days, j represents interviewers, k represents PSUs, *miles* is the total number of miles traveled per interviewer-day, *segments* is the total number of segments visited per interviewer-day, *SmallMSA* and *NonMSA* are dichotomous indicator variables for the urbanicity of the PSU, *CensusHTC* is a continuous measure indicating the Census Hard-to-Count score (Bruce and Robinson 2006), *AreaSqMiles* is the centered number of square miles in the PSU, *AnyExp* is an interviewer-level indicator for whether the interviewer has any prior interviewing experience, *Week* is a continuous measure of the week of the field period, *Phase1* is an indicator variable for the interviewer-day being early in the field period, b_{00j} is a random effect for interviewers with a normal distribution and mean of zero, c_{00k} is a random effect for PSUs with a normal distribution and mean of zero, and e_{ijk} is a residual term (Raudenbush and Bryk 2002). The model in the HRS is identical except that the two urbanicity variables are replaced with one indicator variable for whether the segment is located in a self-representing PSU or not. SAS 9.4 PROC MIXED is used to estimate these hierarchical linear models.

Model for Segments

We estimate a hierarchical Poisson model with a log link to predict the number of segments visited:

$$\log(segments_{ij}) = \theta_0 + \beta_1 miles_{i(jk)} + \beta_2 SmallMSA_k + \beta_3 NonMSA_k + \beta_4 CensusHTC_k + \beta_5 AreaSqMiles_k + \beta_6 AnyExp_j + \beta_7 Week_{i(jk)} + \beta_8 Phase1_{i(jk)} + \beta_9 Phase1_{i(jk)} Week_{i(jk)} + b_{00j} + c_{00k}$$

where the predictors are as defined above. The number of miles driven per day is added as a predictor to this model. All Poisson models are estimated using SAS 9.4 PROC GLIMMIX.

4.4.2. Models for Field Outcomes

For the second research question, we examine whether there is a relationship between travel and field outcomes.

Model for Attempts

When examining the total number of contact attempts for the screener or the main interview, we estimate a hierarchical negative binomial regression model with a log link function:

$$\log(attempts_{ij}) = \theta_0 + \beta_1 segments_{i(jk)} + \beta_2 miles_{i(jk)} + \beta_3 SmallMSA_k + \beta_4 NonMSA_k + \beta_5 CensusHTC_k + \beta_6 AreaSqMiles_k + \beta_7 AnyExp_j + \beta_8 Week_{i(jk)} + \beta_9 Phase1_{i(jk)} + \beta_{10} Phase1_{i(jk)} Week_{i(jk)} + b_{00j} + c_{00k}$$

We initially estimated hierarchical Poisson models. These models had very poor model fit with evidence of overdispersion (Generalized Chi-Square/DF $>$ 2 for all models; Stroup 2011). Thus, hierarchical negative binomial models were estimated. The negative binomial models significantly improved model fit over the Poisson models for the total number of contact attempts (Generalized Chi-Square $\sim$ $=$ 1 in both surveys; statistically significant scale parameters). All negative binomial models are estimated using SAS 9.4 PROC GLIMMIX.

Models for Contact and Interview Rates

For the contact, screener and main interview rates, we estimate a hierarchical Poisson regression model predicting the number of contact attempts with successful contacts, screener and main interviews with the number of contact attempts, contact attempts to obtain a screener, and contact attempts to obtain a main interview as the offset variables, respectively. All Poisson models are estimated using SAS 9.4 PROC GLIMMIX. For

example, for the contact rate, we estimate:

$$\log(contacts/visits_{ij}) = \theta_0 + \beta_1 segments_{i(jk)} + \beta_2 miles_{i(jk)} + \beta_3 SmallMSA_k + \beta_4 NonMSA_k + \beta_5 CensusHTC_k + \beta_6 AreaSqMiles_k + \beta_7 AnyExp_j + \beta_8 Week_{i(jk)} + \beta_9 Phase1_{i(jk)} + \beta_{10} Phase1_{i(jk)} Week_{i(jk)} + b_{00j} + c_{00k}$$

For these models, we use the same predictors as the previous models, and include both segments visited and miles travelled per day as predictors in these models.

5. Findings

5.1. *What are Predictors of Interviewer Travel Behavior in Two Large National US Field Surveys?*

We now turn to our first research question – what predicts interviewer travel behavior in these two surveys? We start by estimating a base model predicting mileage with no predictors. In the NSFG, there is a significant variance component related to interviewers ($var(b_{00j}) = 7157.1$) and to the PSUs ($\text{var}(c_{00k}) = 5154.7$; likelihood ratio chi-square = 1691.6, $p < .0001$; see the top panel of Table 2), with an interviewer intraclass correlation coefficient of 51.3% and a PSU intraclass correlation coefficient of 36.9%. There is also a significant variance component related to interviewers in the HRS ($var(b_{00j}) = 2397.1$) and to the PSUs ($\text{var}(c_{00k}) = 61.2$; likelihood ratio chi-square = 11963.5, $p < .0001$; see the top panel Table 2), with an interviewer intraclass correlation coefficient of 71.8% and a PSU intraclass correlation coefficient of 1.8%. This means that over 50% of the variance in mileage traveled is due to interviewers overall in both surveys, but that the variance in mileage due to PSUs differs dramatically across the two surveys.

We note that interviewers vary in the characteristics of their assigned PSUs, their skill sets, as well as the proximity of their home to sampled segments. These factors may explain some or all of the variation between interviewers. To address this, we now look at predictors of the number of miles traveled each interviewer-day.

As shown in the top panel of Table 2, the total distance traveled is significantly associated with the number of segments visited in the NSFG, but not the HRS. In the NSFG, interviewers travel almost eight miles more for each additional segment visited. Counter to our expectations, there is no systematic linear association between urbanicity, the Census Hard-to-Count score, the size of the area segment, interviewer experience, or time in the field period and mileage in the NSFG.

The associations in the HRS are somewhat different than in the NSFG. In the HRS, there is no systematic association between the number of segments visited and mileage. As in the NSFG, and again counter to our expectations, there is no association between the Census Hard-to-Count score, the area of the PSU, urbanicity, interviewer experience, or the time in the field period with miles traveled in the HRS.

Table 2. Cross-Classified Random Effects Linear Regression Coefficients and Standard Errors Predicting Number of Miles Travelled Per Interviewer-Day and Cross-Classified Random Effects Poisson Regression Coefficients and Standard Errors Predicting Number of Segments Visited Per Interviewer-Day, for the National Survey of Family Growth (NSFG) and Health and Retirement Study (HRS).

	NSFG			HRS		
	Base Model	Full Model		Base Model	Full Model	
		Coefficient	Standard Error		Coefficient	Standard Error
Miles – Linear Regression						
Intercept	92.49****	55.96	88.75	53.19****	61.48	37.18
Number of segments		7.98****	0.95		0.26	0.24
Small MSA		6.69	35.01		–	
Non-MSA		44.34	44.91		–	
Non-self-representing PSU		–			−2.05	1.93
Census Hard-to-Count Score		−1.11	1.37		0.08	0.07
Area of PSU (Square Miles)		0.009	0.01		−0.0005	0.0003
Any experience		3.75	45.09		5.91	7.75
Week in the field period		4.36	5.31		−0.47	0.73
Phase 1		40.51	60.68		−2.05	36.44
Week in the field period*Phase 1		−4.00	5.34		0.06	0.73
Variance Components						
Interviewer ($var(b_{00j})$)	7157.1****	7450.8****		2397.1****	2378.5****	
Area ($var(c_{00k})$)	5154.7***	5858.6**		61.2****	70.7****	
Residual ($var(e_{ij})$)	1646.4****	1559.5****		881.9****	866.9****	
Chi-square test for variance components	1691.6****	1509.0****		11963.5****	11649.9****	
AIC	16899.9	16777.3		120822.6	120144.6	
−2 Log-Likelihood	16893.3	16771.3		120816.6	120138.6	

Table 2. Continued.

	NSFG			HRS		
	Base Model	Full Model		Base Model	Full Model	
		Coefficient	Standard Error		Coefficient	Standard Error
Segments – Poisson regression						
Intercept	0.62****	3.80**	1.10	0.61****	3.88***	0.99
Number of miles		0.002****	0.0003		0.0003	0.0002
Small MSA		−0.06	0.09		–	
Non-MSA		0.05	0.11		–	
Non-self-representing PSU		–			0.10*	0.04
Census Hard-to-Count Score		0.005	0.003		−0.002	0.001
Area of PSU (Square Miles)		−0.00007*	0.00003		−0.00001	0.000006
Any experience		−0.06	0.12		0.13**	0.04
Week in the field period		−0.30**	0.005		−0.07***	0.02
Phase 1		−3.79***	1.09		−3.47***	0.99
Week in the field period*Phase 1		0.36***	0.09		0.07***	0.02
Interviewer ($var(b_{00j})$)	0.022	0.019		0.049****	0.046****	
Area ($var(c_{00k})$)	0.031*	0.021		0.029****	0.019****	
Chi-square test for variance components	72.5****	39.4****		1216.7****	954.6****	
−2 Log Pseudo-Likelihood	3410.0	2986.5		23995.3	22952.4	
Generalized Chi-Square	1133.4	857.6		7817.6	7461.0	
Generalized Chi-Square/DF	0.64	0.53		0.61	0.60	

Note: Models account for clustering of interviewer-days within interviewer and within area, *$p < .05$, **$p < .01$, ***$p < .001$, ****$p < .0001$. AIC = Akaike Information Criterion; DF = Degrees of freedom; MSA = Metropolitan Statistical Area; PSU = Primary Sampling Unit.

Next, we look at predictors of the number of segments visited on each interviewer day (bottom panel of Table 2). Here, the predictors are surprisingly different from those of mileage. Mirroring the model predicting the number of miles driven, in the NSFG, the number of miles driven is positively associated with the number of segments visited, but in the HRS this coefficient is not significantly different from zero. The urbanicity indicators are not associated with the number of segments visited per day in the NSFG, counter to our expectations, but interviewers who work in non-self-representing PSUs visit more segments per day, on average, than interviewers who work in other PSUs in the HRS, as expected. In the NSFG, experienced and inexperienced interviewers visit the same number of segments per day, on average, but in the HRS, experienced interviewers visit more segments on average than inexperienced interviewers each day. This is likely due to constraints in the number of segments assigned to interviewers in the NSFG (only three), whereas the HRS interviewers may have larger numbers of assigned segments. In both surveys, fewer segments are visited earlier in the field period than later in the field period, indicating more cross-segment travel late in the field period, and the rate of change across the weeks for the first and second phases also changes.

5.2. Is Interviewer Travel Associated with Field Outcomes?

We now look at the relationship between these two travel measures and field behaviors. We start with the total number of contact attempts. There is significant variation across interviewers and PSUs in both studies in the number of screener interview contact attempts (base models in Table 3, NSFG: $\text{var}(b_{00j}) = 0.10$, $\text{var}(c_{00k}) = 0.06$, $p < .0001$; HRS: $\text{var}(b_{00j}) = 0.32$, $\text{var}(c_{00k}) = 0.18$, $p < .0001$) and main interview contact attempts (NSFG: $\text{var}(b_{00j}) = 0.08$, $\text{var}(c_{00k}) = 0.13$ $p < .0001$; HRS: $\text{var}(b_{00j}) = 0.32$, $\text{var}(c_{00k}) = 0.18$) per day. As shown in Table 3 in both surveys, as expected, as the number of segments visited increases, the number of screener and main contact attempts made during an interviewer-day also increases (NSFG screener $b = 0.25$, $p < .0001$; NSFG main $b = 0.20$, $p < .0001$; HRS screener $b = 0.26$, $p < .0001$; HRS main $b = 0.27$, $p < .0001$). Strikingly and surprisingly, there is no relationship ($p > 0.05$) between the number of miles traveled and the total number of main attempts in either survey, and there is a weak association ($b = 0.001$, $p < .05$) between the number of miles and screener attempts in the NSFG, but not the HRS. There is no association between urbanicity and contact attempts across the two surveys, and no association between an interviewer's prior experience, the Census Hard-to-Count score, and the area of the PSU, and the number of contact attempts. In both surveys, the number of screener attempts decreases as the field period progresses, although the decrease is stronger in Phase 2 of the survey. The number of main interview attempts does not change as the field period progresses in the HRS ($p > 0.05$), but increases late in the field period in the NSFG ($p < .01$). The NSFG has a short field period, and although the survey managers emphasize completing screening interviews early in the field period, there is a push toward completing main interviews at the end of the field period. For the HRS, the interview is much longer. In many cases, two persons within the household are interviewed. These conditions made it more difficult to schedule and complete these interviews.

Table 3. Cross-Classified Random Effects Negative Binomial Regression Coefficients and Standard Errors Predicting Number of Contact attempts Per Interviewer-Day, for the National Survey of Family Growth (NSFG) and Health and Retirement Study (HRS).

	NSFG						HRS					
	Screener			Main			Screener			Main		
	Base Model	Full Model		Base Model	Full Model		Base Model	Full Model		Base Model	Full Model	
		Coefficient	Standard Error		Coefficient	Standard Error		Coefficient	Standard Error		Coefficient	Standard Error
Intercept	2.30****	8.69****	1.49	1.37****	2.99*	1.26	2.15****	59.17****	7.34	0.069	−1.08	1.31
Distance in miles		0.001*	0.0004		0.0005	0.0004		0.0002	0.0003		−0.0007	0.0004
Number of segments visited		0.25****	0.02		0.20****	0.02		0.26****	0.007		0.27****	0.009
Small MSA		−0.26	0.16		0.20	0.16						
Non-MSA		−0.41	0.21		0.17	0.21						
Non-self-representing PSU		–						−0.11	0.06		−0.01	0.07
Census Hard-to-Count Score		0.01	0.006		0.0006	0.006		0.005	0.003		−0.003	0.002
Area of PSU (Square Miles)		−0.00003	0.00006		0.00008	0.00006		0.00001	0.00001		0.000003	0.00001
Any experience		0.03	0.25		−0.10	0.21		0.07	0.09		0.09	0.11
Week in the field period		−0.75****	0.13		−0.23*	0.11		−1.25****	0.15		−0.01	0.02
Phase 1		−6.24****	1.46		−2.39	1.22		−57.43****	7.34		−2.18	1.30
Week in the field period*Phase 1		0.60****	0.13		0.30**	0.11		1.24****	0.15		0.04	0.03

Table 3. Continued.

	NSFG						HRS					
	Screener			Main			Screener			Main		
	Base Model	Full Model		Base Model	Full Model		Base Model	Full Model		Base Model	Full Model	
		Coefficient	Standard Error		Coefficient	Standard Error		Coefficient	Standard Error		Coefficient	Standard Error
Dispersion	0.86****	0.48****		0.49****	0.22****		0.70****	0. 642****		0.82****	0.61****	
Interviewer ($var(b_{00j})$)	0.10**	0.17****		0.08**	0.10***		0.32****	0.24****		0.52****	0.37****	
Area ($var(c_{00k})$)	0.06	0.08*		0.13**	0.10**		0.18****	0.15****		0.16****	0.11****	
Chi-square test for variance components	148.0****	197.8****		218.5****	237.0****		2221.7****	1787.2****		1471.1****	1137.5****	
−2 Log-Likelihood	5098.1	4049.0		4690.0	3701.6		34836.3	152984.9		44150.6	41714.5	
Generalized Chi-square	1783.8	1659.7		1761.7	1614.2		12985.0	131668.3		12770.5	12678.1	
Generalized Chi-Square/DF	1.00	1.03		0.99	1.00		1.01	10.63		0.99	1.02	

Note: Models account for clustering of interviewer-days within interviewer, $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$. DF = Degrees of Freedom; MSA = Metropolitan Statistical Area; PSU = Primary Sampling Unit.

We now examine the relationship between travel behavior and contact rates (Table 4). As with the other field outcomes, there is significant variability across interviewers and PSUs in contact rates (NSFG: $\text{var}(b_{00j}) = 0.09$, $\text{var}(c_{00k}) = 0.04$, $p < .0001$; HRS: $\text{var}(b_{00j}) = 0.08$, $\text{var}(c_{00k}) = 0.02$, $p < .0001$). Consistent with our expectation that interviewer travel and contact rates will be negatively correlated, Table 4 shows that there is a modest, but noticeable decline in contact rates in both surveys as an interviewers' travel increases as measured by the number of segments (NSFG: $b = -0.03$, $p < .001$; HRS: $b = -0.03$, $p < .0001$). Unexpectedly, we see no relationship at all between the distance traveled in miles and contact rates ($p > 0.05$). Non-urban PSUs as represented by the non-MSAs in the NSFG ($b = 0.36$, $p < .01$) and by non-self-representing PSUs in the HRS ($n = 0.07$, $p < .0001$) have higher contact rates, on average, than urban PSUs in both surveys. We see no difference in contact rates for interviewers with some prior experience compared to those without prior experience in either survey. Contact rates do not systematically change over the field period. There is no association between the Census Hard-to-Count score and contact rates ($p > 0.05$). Larger PSUs have higher contact rates in the NSFG ($b = 0.0001$, $p < .01$), but not in the HRS ($p > 0.05$).

The next outcomes, also presented in Table 4, are screener interview rates and main interview rates. There is significant variability across interviewers in screening and main interview rates (NSFG: screening $\text{var}(b_{00j}) = 0.11$, $\text{var}(c_{00k}) = 0.08$, $p < .0001$, main $\text{var}(b_{00j}) = 0.09$, $\text{var}(c_{00k}) = 0.05$, $p < .0001$; HRS: screening $\text{var}(b_{00j}) = 0.16$, $\text{var}(c_{00k}) = 0.05$, $p < .0001$, main $\text{var}(b_{00j}) = 0.17$, $\text{var}(c_{00k}) = 0.09$, $p < .0001$). We anticipate the same association as described with contact rates – a negative association between travel (as measured by miles and number of segments visited) and screener / main interview rates, with the same rationale. As with contact rates and consistent with our expectations, there is a statistically significant negative association between screener rates and the total number of segments visited on an interviewer-day and also between main interview rates and the total number segments visited on an interviewer-day in both surveys (coefficients range from -0.05 to -0.27, $p < .001$). Neither survey displays a significant relationship ($p < .05$) between the total distance traveled in miles and any of these field outcomes. Thus, the measure of travel that predicts these important field outcomes (and error indicators) in both surveys is how many different segments are visited, not the number of miles that an interviewer drives. Non-urban PSUs tend to have higher screener and main interview rates (NSFG: $b = 0.55$, $p < .01$ screener; $b = 0.30$, $p = 0.06$ main; HRS: $b = 0.12$, $p < .0001$ screener; $b = 0.41$, $p < .0001$ main) in both surveys. There is no association between interviewer-day level screener interview and main interview rates and interviewer experience. For the HRS, interviewer-day level screener interview rates changes as the field period progresses. In both surveys, main interview rates change as the field period progresses. There is no association between the Census Hard-to-Count score and screener or main completion rates. Screener completion rates are associated with larger PSUs in the NSFG ($b = 0.005$, $p < .01$), but no association with main interview rates, and no association in the HRS ($p > 0.05$).

6. Summary and Discussion

Although costs of interviewer travel have been examined with respect to sample designs (e.g., Kalsbeek et al. 1994; Bienias et al. 1990) predictors of interviewer travel behavior

Table 4. Cross-Classified Random Effects Poisson Regression Coefficients and Standard Errors Predicting Interviewer-Day Level Contact, Screener Interview and Main Interview Rates, for the National Survey of Family Growth (NSFG) and Health and Retirement Study (HRS).

	NSFG									HRS								
	Contact			Screener			Main			Contact			Screener			Main		
	Base Model	Full Model		Base Model	Full Model		Base Model	Full Model		Base Model	Full Model		Base Model	Full Model		Base Model	Full Model	
		Coef.	SE		Coef.	SE		Coef.	SE		Coef.	SE		Coef.	SE		Coef.	SE
Intercept	−0.10****	−2.52	1.26	−1.61****	−2.29	2.80	−1.59****	−6.04*	2.48	−1.04****	−1.45****	1.20	−1.51****	−1.10****	0.10	−1.90****	−1.10****	0.17
Distance in miles		−0.0003	0.0002		0.000005	0.0004		0.0002	0.0004		−0.0002	0.0001		−0.0003	0.0002		0.0005	0.0005
Number of segments		−0.03**	0.01		−0.05**	0.02		−0.18****	0.02		−0.03****	0.004		−0.06****	0.005		−0.27****	0.02
Small MSA		0.03	0.10		0.05	0.14		0.05	0.13		–			–				
Non-MSA		0.36**	0.13		0.55**	0.17		0.30	0.16		–			–				
Non-self-representing PSU					–			–			0.07****	0.02		0.12****	0.05		0.41****	0.10
Census Hard-to-Count Score		−0.006	0.004		−0.010	0.005		−0.009	0.005		−0.002	0.001		−0.002	0.002		0.0007	0.003
Area of PSU (Square Miles)		0.0001**	0.00004		0.0001**	0.00005		0.00003	0.00004		−0.000003	0.000006		0.000005	0.000009		0.000004	0.00001
Any experience		−0.16	0.17		−0.09	0.20		−0.21	0.18		0.03	0.05		−0.04	0.07		−0.11	0.09
Week in the field period		0.15	0.11		0.04	0.25		0.42*	0.22		0.01	0.02		−0.008****	0.0006		−0.02****	0.002
Phase 1		2.03	1.24		1.35	2.79		5.55*	2.47		0.70	1.19		16.14****	4.09		4.76	3.61
Week in the field period*Phase 1		−0.16	0.11		−0.11	0.25		−0.47*	0.22		−0.02	0.02		0.32****	0.08		−0.11	0.07
Interviewer var(b_{00j})	0.09***	0.08****		0.11***	0.10****		0.09**	0.05*		0.08****	0.08****		0.16****	0.14****		0.17****	0.12****	
Area var(c_{00k})	0.04**	0.03**		0.08*	0.05*		0.05	0.04		0.02****	0.02****		0.05****	0.05****		0.09****	0.09****	
Chi-square test for variance components	702.7****	439.4****		409.2****	199.8****		104.6****	50.9****		3391.3****	3127.6****		3248.0****	2664.3****		358.6****	261.6****	
−2 Log Pseudo-Likelihood	3514.45	2974.51		3929.72	3689.6		5657.69	5029.50		27743.18	25938.50		32325.02	30392.78		31209.06	29742.38	
Generalized Chi-square	2220.44	1915.30		1801.54	1624.2		2002.76	1713.95		14966.36	14072.51		14546.67	13622.59		8632.88	7840.19	
Generalized Chi-Square/DF	1.25	1.19		1.23	1.18		1.25	1.16		1.16	1.14		1.24	1.20		1.21	1.13	

Note: Models account for clustering of interviewer-days within interviewer and area, *$p < .05$, **$p < .01$, ***$p < .001$, ****$p < .0001$; Total number of contact attempts for a contact, screener or main interview used as offset in each model. DF = Degrees of Freedom; Coef = Coefficient; SE = Standard Error; PSU = Primary Sampling Unit; MSA = Metropolitan Statistical Area.

and the association between travel decisions and field outcomes to date has not received any attention. In this article, we conducted an initial examination of variables that may be associated with interviewer travel, treating indicators of interviewer travel behavior as outcomes in some models and predictors in others. In the models, geographic characteristics such as the size in square miles of the PSU and urbanicity were not associated with travel outcomes such as the numbers of miles traveled or segments visited. In models predicting the number of segments visited, characteristics of the data collection, such as design phases or the week of data collection, were associated with these travel outcomes. In these models, interviewer variance was a significant component of the variance. In terms of field outcomes, the geographic size of the PSU did not play a significant role except for screener and contact rates in the NSFG. The week of the field period and the design phase were associated with the number of attempts made in a day.

In terms of the relationship of interviewer travel to the six different field outcomes and error indicators, we see a clear, consistent pattern for the two surveys, summarized in Table 5. The associations are clear – interviewers who visit more segments on a given day also have more contact attempts and have lower contact and response rates. This effect holds even accounting for the number of miles that the interviewers travel. In none of the analyses is the raw number of miles traveled by interviewers associated with the number of contact attempts or contact or response rates.

The replication of the findings about the association (or lack thereof) between overall distance and number of segments visited and field outcomes despite the design differences between the two surveys is encouraging. The field period in the HRS and NSFG is very different. In the NSFG, there is a limited (twelve week) field period. Interviewers are encouraged to visit every sampled housing unit as quickly as possible and visit as many segments as they can every day. In this way, the NSFG encourages interviewers to maximize their travel. In contrast, the HRS has an extended (twelve month) field period, and interviewers are encouraged to minimize their travel as the field period progresses. These differences in field period and survey management yield differences in the relationship between travel behaviors and contact attempts over the field period across the two surveys; yet there are few differences in the association between travel and field outcomes between these two surveys. The number of segments an interviewer visits has a positive association with contact attempts, but negative association with response rates. These results suggest that this finding may generalize across survey settings.

Table 5. Summary of association between travel and field outcomes.

	Is there an association between travel and field outcome?	
Field Outcome	Miles	Segments
Screener contact attempts	Weak (+ NSFG only)	Yes (+)
Main contact attempts	No	Yes (+)
Contact rates	No	Yes (−)
Screener rates	No	Yes (−)
Main interview rates	No	Yes (−)

Note: The sign in parentheses in the "segments" column indicates the direction of the association. NSFG = National Survey of Family Growth.

The association between interviewer travel behavior and contact and cooperation rates is important for survey practice for three reasons. First, most survey organizations closely monitor response rates, potentially indicating differential nonresponse bias across interviewers (West and Olson 2010). Although Groves and Peytcheva (2008) found that the response rate may not be a good indicator for the risk of bias, it still is a baseline quality measure used to assess interviewers by many survey organizations. Second, the ability to control variability between interviewers in the response rate is an important prerequisite in controlling differential response rates across subgroups, which is a strategy for minimizing the risk of nonresponse bias (Montaquila et al. 2008; Groves 2006; Schouten et al. 2016). Finally, we do not know whether different types of travel decisions are associated with contact or cooperation rates. Thus, establishing whether such an association exists is a critical first step for developing interviewer training related to travel and for understanding variability in interviewer response rates. Our findings imply that survey organizations should carefully monitor interviewer travel behavior – and in particular, between segment travel – as a way of reducing between-interviewer variability in response rates, and thus minimize the risk of nonresponse error variance due to the interviewer.

These findings also have important implications for costs and practice by survey field managers. Although interviewer mileage should be monitored for a simple cost calculation (travel costs = number of miles * reimbursement rate per mile), the number of segments visited each day should also be monitored as an additional indicator of survey error and cost. Furthermore, we recommend that the number of segments visited on a given day be used to initiate conversations with interviewers about their travel. In particular, field managers should investigate why interviewers are traveling to additional segments as the travel is an indication that calling is less productive on those trips. This may be due to the time of day and day of the week of the trip, the approach of the interviewer, or other factors which may need further investigation. Thus, the number of segments could be a useful way to monitor and provide feedback to interviewers on their behavior and obtain new insights into the reasons for nonparticipation.

This article is not without limitations. First, establishing causality from existing administrative travel data is difficult. Most travel data is reported at an aggregate interviewer-day level in timesheets; that is, travel data is not associated with individual cases or contact attempts but instead with each day that an interviewer works. Thus, we can examine associations in this analysis, but cannot determine whether different travel behaviors cause field outcomes, or field outcomes cause different travel behavior. The observed associations could also be the result of particular allocations of sample to interviewers. For example, if more experienced interviewers are allocated more difficult samples, this could lead to experience being less predictive in the estimated models while it might actually be the case that experienced interviewer produce higher response rates than less experienced interviewers when they are allocated equivalent samples. This type of allocation based on difficulty was not the usual practice in either of the surveys used here, but might be used by other surveys. Second, these two surveys are large, national surveys with screening to find particular target populations. Each of these surveys also oversamples area segments with higher proportions of black and Hispanic persons. We do not know how these results would replicate in surveys without this screening step and oversampling. Third, the segments were not randomly assigned to interviewers, and as such, all of our findings are from an observational study. We

have attempted to account for these potential differences in interviewer assignment PSUs using cross-classified random effects models and including additional area-level predictors (e.g., see Stokes and Jones 1989), but we have not explained away the PSU effect. Additionally, neither of these surveys pay interviewers per completed interview, nor require interviewers to work multiple studies at one time. Instead, NSFG and HRS pays interviewers by the hour, and both hire interviewers only for one study at a time. It is not clear whether surveys that use a different pay structure would see similar patterns.

Future evaluations of the quality of travel data are needed. Although we have initial indications that interviewer-reported mileage and mileage calculated from geocoded call records differ, we do not know which source is more accurate. Given that all of the analyses replicated with both timesheet and geocoded mileage information, we believe that our lack of association with mileage is robust to measurement errors in the timesheet data. GPS devices allow the collection of data regarding the location of interviewers. These data include latitude, longitude, speed, direction, altitude, and time and date. Collection of real time travel data through the use of GPS devices carried by interviewers would help us understand the quality of both of these sources of data (Olson and Wagner 2015; Wagner et al. 2013). Additionally, collection of real time travel data would permit the examination of the relationship between travel and field outcomes at the address level, rather than the interviewer-day level. It would also allow us to evaluate the amount of time spent to travel a certain number of miles. Interviewers may take a longer route at which they can travel with faster speed such as on a freeway, thus taking less time from their interviewing day than a shorter route at slower speeds.

Future research will incorporate information about both the distance traveled and the number of segments visited per visit into explicit cost models. Although interviewer travel is often mentioned as a constraint on the number of clusters to select and the size of the clusters, we are unaware of cost-error models that incorporate empirical data for these measures. The data presented here could provide useful inputs for future cost models related to interviewer travel in face-to-face surveys.

7. References

Biemer, P.P., P. Chen, and K. Wang. 2013. "Using Level-of-Effort Paradata in Non-Response Adjustments with Application to Field Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176(1): 147–168. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01058.x.

Bienias, J.L., E.M. Sweet, and C.H. Alexander. 1990. A Model for Simulating Interviewer Travel Costs for Different Cluster Sizes. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20–27. Anaheim, CA. Available at: https://ww2.amstat.org/sections/srms/Proceedings/papers/1990_004.pdf (accessed October 14, 2017).

Blom, A.G. 2012. "Explaining Cross-Country Differences in Survey Contact Rates: Application of Decomposition Methods." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175(1): 217–242. Doi: http://dx.doi.org/10.1111/j.1467-985X.2011.01006.x.

Blom, A.G., E.D. de Leeuw, and J.J. Hox. 2011. "Interviewer Effects on Nonresponse in the European Social Survey." *Journal of Official Statistics* 27(2): 359–377.

Bruce, A. and J.G. Robinson. 2006. "Tract-Level Planning Database with Census 2000 Data." US Department of Commerce, US Census Bureau: Washington, DC, 226. Available at: https://www.census.gov/2010census/partners/pdf/TractLevelCensus-2000Apr_2_09.pdf (accessed October 14, 2017).

Campanelli, P., P. Sturgis, and S. Purdon. 1997. *Can You Hear Me Knocking?: Investigation into the Impact of Interviewers on Survey Response Rates*, National Centre for Social Research, London.

Chen, B.-C. 2012. "Simulating NHIS Field Operations." *Proceedings of the 2012 Research Conference*, Federal Committee on Statistical Methodology. Office of Management and Budget. Washington, DC. Available at: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/Chen_2012FCSM_II-A.pdf (accessed October 14, 2017).

Cochran, W.G. 1977. *Sampling Techniques* (Third ed.). New York: John Wiley and Sons.

Durrant, G.B. and F. Steele. 2009. "Multilevel Modelling of Refusal and Non-Contact in Household Surveys: Evidence from Six UK Government Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172(2): 361–381. Doi: http://dx.doi.org/10.1111/j.1467-985X.2008.00565.x.

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70(5): 646–675. Doi: http://dx.doi.org/10.1093/poq/nfl0.

Groves, R.M. and E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72(2): 167–189.

Hansen, M.H., W.N. Hurwitz, and W.G. Madow. 1953. *Sample Survey Methods and Theory*. New York: Wiley.

Health and Retirement Study. 2008. *Sample Evolution: 1992–1998*. Available at: http://hrsonline.isr.umich.edu/sitedocs/surveydesign.pdf (accessed February 6, 2014).

Heeringa, S.G. and J.H. Connor. 1995. *Technical Description of the Health and Retirement Survey Sample Design*. Available at: http://hrsonline.isr.umich.edu/sitedocs/userg/HRSSAMP.pdf (accessed February 6, 2014).

Judkins, D., J. Waksberg, and D. Northrup. 1990. Cost Functions for NHIS and Implications for Survey Design. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 34–43. Anaheim, CA.

Kalsbeek, W.D., O.M. Mendoza, and D.V. Budescu. 1983. "Cost Models for Optimum Allocation in Multi-Stage Sampling." *Survey Methodology* 9(2): 154–177.

Kalsbeek, W.D., S.L. Botman, J.T. Massey, and P.-W. Liu. 1994. "Cost-Efficiency and the Number of Allowable Call Attempts in the National Health Interview Survey." *Journal of Official Statistics* 10(2): 133–152.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley & Sons, Inc.

Lepkowski, J.M., W.D. Mosher, K.E. Davis, R.M. Groves, and J. Van Hoewyk. 2010. *The 2006–2010 National Survey of Family Growth: Sample design and analysis of a continuous survey*. Washington, DC: National Center for Health Statistics. Available at: https://www.cdc.gov/nchs/data/series/sr_02/sr02_150.pdf (accessed October 14, 2017).

Mayer, C.S. 1968. "A Computer System for Controlling Interviewer Costs." *Journal of Marketing Research* 5: 312–318.

Montaquila, J.M., J.M. Brick, M.C. Hagedorn, C. Kennedy, and S. Keeter. 2008. "Aspects of Nonrepsonse Bias in RDD Telephone Surveys." In *Advances in Telephone Survey*

Methodology, edited by J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, and R.L. Sangster, 561–586. Hoboken, NJ: John Wiley & Sons.

Morton-Williams, J. 1993. *Interviewer Approaches*. Hants, England: Aldershot.

Olson, K. and A. Peytchev. 2007. "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes." *Public Opinion Quarterly* 71(2): 273–286.

Olson, K. and J. Wagner. 2015. "A Feasibility Test of Using Smartphones to Collect GPS Information in Face-to-Face Surveys." *Survey Research Methods* 9(1): 1–13. Doi: http://dx.doi.org/10.18148/srm/2015.v9i1.6036.

O'Muircheartaigh, C. and P. Campanelli. 1999. "A Multilevel Exploration of the Role of Interviewers in Survey Non-Response." *Journal of the Royal Statistical Society, Series A* 162: 437–446. Doi: http://dx.doi.org/10.1111/1467-985X.00147.

Peachman, J. 1992. *Design and Methodology of 1991/92 Household Interview Survey*. Paper presented at the Australasian Transport Research Forum 17th Conference, 149–161. Canberra, Australia. Available at: http://atrf.info/papers/1992/1992_Peachman.pdf (accessed October 14, 2017).

Pickery, J. and G. Loosveldt. 2002. "A Multilevel Multinomial Analysis of Interviewer Effects on Various Components of Unit Nonresponse." *Quality and Quantity* 36: 427–437. Doi: https://doi.org/10.1023/A:1020905911108.

Purdon, S., P. Campanelli, and P. Sturgis. 1999. "Interviewers' Calling Strategies on Face-to-Face Interview Surveys." *Journal of Official Statistics* 15(2): 199–219.

Raudenbush, S.W. and A.S. Bryk. 2002. *Hierarchical linear models : applications and data analysis methods*. Thousand Oaks, Sage Publications.

Schouten, B., F. Cobben, P. Lundquist, and J. Wagner. 2016. "Does More Balanced Survey Response Imply Less Non-Response Bias?" *Journal of the Royal Statistical Society: Series A* (Statistics in Society) 179(3): 727–748. Doi: http://dx.doi.org/10.1111/rssa.12152.

Stoop, I., J. Billiet, A. Koch, and R. Fitzgerald. 2010. *Improving Survey Response: Lessons Learned from the European Social Survey*. West Sussex, UK: Wiley.

Stokes, S.L. and P.A. Jones. 1989. Evaluation of the Interviewer Quality Control Procedure for the Post-Enumeration Survey. *1989 Proceedings of the Survey Research Methods Section*. American Statistical Association. Alexandria, VA, 696–698.

Stroup, W. 2011. Living with Generalized Linear Models. Paper 349-2011, SAS Global Forum 2011. Available at: http://support.sas.com/resources/papers/proceedings11/349-2011.pdf (accessed October 14, 2017).

Sudman, S. 1965-66. "Time Allocation in Survey Interviewing and in Other Field Occupations." *Public Opinion Quarterly* 29: 638–648.

Sudman, S. 1967. *Reducing the Cost of Surveys*. Chicago: Aldine.

Sudman, S. 1978. "Optimum Cluster Designs Within a Primary Unit Using Combined Telephone Screening and Face-to-Face Interviewing." *Journal of the American Statistical Association* 73: 300–304. Doi: http://dx.doi.org/10.2307/2286656.

United States Census Bureau. 2013. "Metropolitan and Micropolitan Statistical Areas Main." Available at: http://www.census.gov/population/metro/. Last Revised: May 6, 2013; (accessed January 15, 2015).

Wagner, J., K. Olson, and M. Edgar. 2013. *Using GPS and Other Data to Assess Errors in Level-of-Effort Data in Field Surveys*. Paper presented at the Joint Statistical Meetings, August 2013. Montreal, Canada.

Wang, K. and P. Biemer. 2010. *The Accuracy of Interview Paradata: Results from a Field Investigation*. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Chicago, IL. Available at: https://www.rti.org/sites/default/files/resources/aapor10_biemer_pres.pdf (accessed October 14, 2017).

Weeks, M.F., R.A. Kulka, J.T. Lessler, and R.W. Whitmore. 1983. "Personal versus Telephone Surveys for Collecting Household Health Data at the Local Level." *American Journal of Public Health* 73: 1389–1394.

West, B.T. and K. Olson. 2010. "How Much of Interviewer Variance is Really Nonresponse Error Variance?" *Public Opinion Quarterly* 74(5): 1004–1026. Doi: http://dx.doi.org/10.1093/poq/nfq061.

West, B.T., K.B. Welch, and A.T. Galecki. 2015. *Linear Mixed Models: A Practical Guide Using Statistical Software* (Second ed.). Boca Raton, FL: CRC Press.

Received February 2014
Revised July 2017
Accepted November 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 239–263, http://dx.doi.org/10.1515/JOS-2018-0011

An Overview of Population Size Estimation where Linking Registers Results in Incomplete Covariates, with an Application to Mode of Transport of Serious Road Casualties

Peter G.M. van der Heijden[1], *Paul A. Smith*[2], *Maarten Cruyff*[3], *and Bart Bakker*[4]

We consider the linkage of two or more registers in the situation where the registers do not cover the whole target population, and relevant categorical auxiliary variables (unique to one of the registers; although different variables could be present on each register) are available in addition to the usual matching variable(s). The linked registers therefore do not contain full information on either the observations (often individuals) or the variables. By treating this as a missing data problem it is possible to construct a linked data set, adjusted to estimate the part of the population missed by both registers, and containing completed covariate information for all the registers. This is achieved using an Expectation-Maximization (EM)-algorithm. We elucidate the properties of this approach where the model is appropriate and in situations corresponding with real applications in official statistics, and also where the model conditions are violated. The approach is applied to data on road accidents in the Netherlands, where the cause of the accident is denoted by the police and by the hospital. Here the cause of the accident denoted by the police is considered as missing information for the statistical units only registered by the hospital, and the other way around. The method needs to be widely applied to give a better impression of the range of problems where it can be beneficial.

Key words: Dual system estimation; linkage; missing data; register; coverage.

1. Introduction

In recent years there has been continuing pressure on National Statistical Offices (NSOs) and other organisations producing official statistics to produce more, better quality and more detailed statistics, generally with decreasing resources. One of the important ways NSOs have responded has been to increase the use of administrative data sets, which provide relatively large amounts of information, generally at a small marginal cost. Often the desired range of statistical units or variables is not available on a single administrative data set, and therefore linking of administrative data sources (we call them registers) is also becoming more and more popular as a means to provide more comprehensive statistics.

There are several methodological problems that NSOs encounter when they are using registers for the production of official statistics. One is that registers, even when linked,

[1] Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands and University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: P.G.M.vanderHeijden@uu.nl
[2] University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: p.a.smith@soton.ac.uk
[3] Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands. Email: m.cruyff@uu.nl
[4] Statistics Netherlands, P.O.Box 24500, 2490 HA Den Haag, The Netherlands. Email: bfm.bakker@cbs.nl

under-cover the population of interest. A second problem is that missing values will be generated if two or more registers are linked. The values of variables that are only available in a subset of the registers are not known for the records that are not present in this subset. In this article, we present a framework for solving both undercoverage and these missing values in one procedure. We do not consider other methodological problems such as overcoverage and item missingness in single registers, although we return to them in the discussion. Figure 1 is a graphical representation of the linkage of two registers. The representation shows the linked data with observations (often individuals) in the rows and variables in the columns. The data for variables available only in register *A* are on the left and denoted by *a*, the data for variables only in register *B* are on the right and denoted by *b*, and in the middle are the data for variables that register *A* and *B* have in common, denoted by *ab*. Typically the variables in *ab* include the variables used for linking the registers.

As Figure 1 illustrates, each register has some unique variables. In *a* we find data for the covariates in register *A* that are *not* in register *B*. It follows that for the individuals in register *B* that are not in register *A* these covariates are missing. This is represented by the grey bitmap block at the bottom left in the representation in Figure 1. Similarly, individuals that are in register *A* but *not* in register *B* have missing values on the variables that are unique for register *B*, and this is represented by the grey bitmap block top right in Figure 1. In this article we consider the presence of the two grey bitmap blocks as a missing data problem that we solve by estimating the missing data. (It is evident that estimating missing covariate values only makes sense for covariates that pertain to all registers involved. An example where estimation of missing covariate values does not make sense: consider a population register coupled with a hospital register, then the

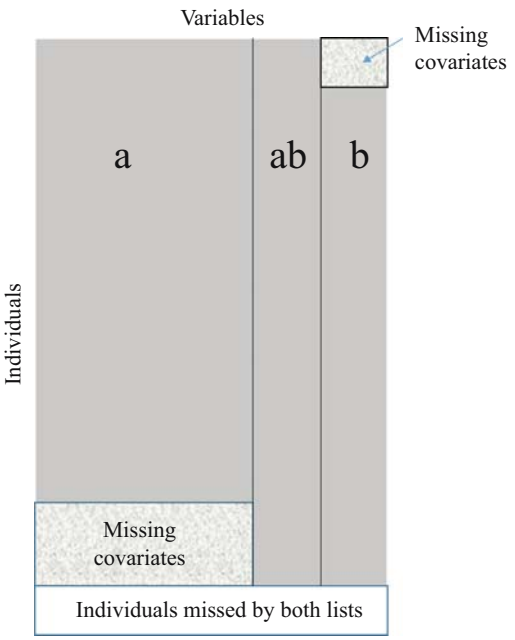


Fig. 1. Graphical representation of two linked registers, see text for details.

hospital register covariate "type of medical problem" should not be estimated for all individuals in the population register who do not appear in the hospital register, as it is likely that they do not have a medical problem at all.)

In addition, the linked registers may not cover the population perfectly, and two-source estimation may be applied to estimate that part of the population missed by both registers. This is depicted by the white area at the bottom of Figure 1. Notice that the aim is here not only to estimate the number of missing individuals but also their covariate data. Also notice that, in common with the basic population size estimation problem using two registers, individuals can be observed in three ways: individuals only in register *A* (on top), individuals in both register *A* and *B* (in the middle), and individuals only in register *B* (at the bottom). We note that when there are two registers, two-source estimation assumes independence conditional on covariates. Heterogeneity of inclusion probabilities can lead to marginal dependence between the registers. When this heterogeneity is caused by observed covariates, using these covariates in the model can compensate for this dependence. However, there may be dependence that is not caused by observed covariates and one way to handle this true dependence is by inclusion of a third register (compare the International Working Group for Disease Monitoring and Forcasting 1995) or by using latent variable models (compare Darroch et al. 1993; Fienberg et al. 1999). We will show that the approach we adopt can also be elaborated for more than two registers.

Thus there is a missing data problem in the covariates and a population size estimation problem, and both problems are handled simultaneously. In earlier work, Zwane and Van der Heijden (2007) and Van der Heijden et al. (2012) studied the situation where the missing variables are categorical and Zwane and Van der Heijden (2008) where they are continuous. In this article we review the case of categorical variables only, where the problem will be solved by applying the Expectation-Maximization (EM)-algorithm to estimate the missing observations in the context of population size estimation. We will in particular investigate the properties of the chosen solution as well as applying it within simulation studies.

Secondly, we will discuss an application where a single concept is measured by one variable in *A* and another in *B*, but where the validity of the variable in *A* is considered to be better than the validity of the variable in *B*. Notice that the concept is not represented by a single variable in part *ab* in Figure 1, but by a variable in part *A* and a different variable in part *B* and in this case the variable is clearly relevant in both (all) registers. Now the focus is on the estimation of the missing data of the grey bitmap part at the bottom left of the representation in Figure 1.

A good overview of two-source and multiple-source estimation where registers are linked is Bishop et al. (1975, Ch. 6). Important work in official statistics includes Wolter (1986), Bell (1993) and Griffin (2014) for the US Census, and by Brown et al. (1999), Brown et al. (2006) and Brown et al. (2011) for the UK. In epidemiology important reviews are by the International Working group for Disease Monitoring and Forecasting (1995) and Chao et al. (2001). For a Bayesian perspective, see Madigan and York (1997).

In this article the covariates in a population size estimation model play an important role. Earlier work in this area is from Bishop et al. (1975, Ch. 6), Alho (1990), Huggins (1989), Baker (1990), Tilling and Sterne (1999), and Zwane and Van der Heijden (2005); for a review see Pollock (2002). Bishop et al. (1975) discuss the use of categorical

covariates, and Alho (1990) and Zwane and Van der Heijden (2005) discuss how inclusion probabilities may be functions of continuous auxiliary information, where Alho (1990) predicts the inclusion probabilities using logistic regressions for two registers and Zwane and Van der Heijden (2005) generalize this to more than two registers. These papers do not discuss the problem of partly missing covariates.

The problem of partly missing covariates which we discuss here arises because the registers available describe different parts of populations, for example, the registers cover different but overlapping regions in a country, or cover different but overlapping periods in time. Zwane et al. (2004) and Sutherland et al. (2007) also approach this problem as a missing data problem, where, for the region example, the regional parts of a register that are missed by design are estimated using the EM-algorithm. Here the dependence structure between the registers in those regions that are observed by more than one register is projected onto those regions where one or more registers are missing. Zwane et al. (2004) and Sutherland et al. (2007) illustrate this for an example of six registers on spina bifida that are operative in different but overlapping time periods, where they fit log-linear models in the M-Step, and Pelle et al. (2016) fit multidimensional Rasch models to these data.

In the following we will first present the theory and properties of our approach, including the extension to more than two registers. This is followed by simulation studies showing the circumstances under which our approach is better than ignoring the additional variables. We end with an application to the estimation of the number of serious casualties from traffic accidents in the Netherlands measured by the police and by hospitals.

2. Population Size Estimation in the Presence of Missing Covariates: Theory

The basic idea of the methodology that we review can easily be explained by an example taken from Van der Heijden et al. (2012) and Gerritse et al. (2015b), involving the estimation of the population size of people with Afghan, Iranian, or Iraqi nationality (hereafter "AII") in the Netherlands, see Panel 1 in Table 1. Register A is a population register in the Netherlands and register B is a police register. From the population register A the variable Marital status is used, and denoted by X_1, with $X_1 = 1$ referring to married or living together and $X_1 = 0$ referring to unmarried, divorced, or widowed. From the police register B the variable "Police region where apprehended" is used, and denoted by X_2, with $X_2 = 1$ referring to one of the five biggest cities of the Netherlands and $X_2 = 0$ referring to the rest of the country. Notice that Marital status is not available in register B and "Police region where apprehended" is not available in register A. Clearly Marital status is a relevant variable for people in the police register; "Police region where apprehended" is not so obviously relevant for the population register, since most people will not have been apprehended. However, we can consider it as an approximation to usual residence, and therefore it is a relevant variable (though imperfectly measured in this source). If we compare the variables in Table 1 to Figure 1, we see that X_1 in Table 1 is a variable in region A in Figure 1, and X_2 in Table 1 is a variable in region B in Figure 1. In Table 1 A and B are variables denoting presence in registers A and B respectively, with categories 0 = no and 1 = yes; the variables A and B in Table 1 are dichotomous variables in Figure 1 in the areas A and B.

Table 1. Covariate X_1 (Marital status) is only observed in population register A and X_2 (Police region where apprehended) is only observed in police register B.

Panel 1: Observed counts of AII individuals

		$B=1$		$B=0$
		$X_2=0$	$X_2=1$	X_2 missing
$A=1$	$X_1=0$	259	539	13,898
	$X_1=1$	110	177	12,356
$A=0$	X_1 missing	91	164	–

Panel 2: Fitted values under $[AX_2][X_1X_2][BX_1]$

		$B=1$		$B=0$	
		$X_2=0$	$X_2=1$	$X_2=0$	$X_2=1$
$A=1$	$X_1=0$	259.0	539.0	4,510.8	9,387.2
	$X_1=1$	110.0	177.0	4,735.8	7,620.3
$A=0$	$X_1=0$	63.9	123.5	1,112.4	2,150.2
	$X_1=1$	27.1	40.5	1,167.9	1,745.4

The eight counts in Panel 1 of Table 1 correspond to Figure 1 as follows: four counts for AII individuals that are in both register A and register B, which cross–classify individuals using the variables X_1 and X_2; two counts for AII individuals that are only in register A, which categorize them using variable X_1 only, as variable X_2 is missing for the individuals in register A; and two counts for AII individuals that are only in register B, which categorize them using variable X_2 only, as variable X_1 is missing for the individuals in register B.

In Panel 2 of Table 1 the counts 13,898 and 12,356, and the counts 91 and 164, are distributed over the levels of the missing variables. For example, 13,898 is distributed over the levels of X_2 into 4,510.8 and 9,387.2, and the ratio of these two counts is equal to the ratio of the observed counts 259 and 539. Similarly, 91 is split up into 63.9 and 27.1, and the ratio of these two counts is equal to the ratio of the observed counts 259 and 110. As a result, in Panel 2 the odds ratio for the counts 259, 539, 110, and 177 is projected to the four cells on the right and the four cells at the bottom. The theoretical motivation of this projection is given by a Missing At Random (MAR) assumption, and the estimates are found using the EM-algorithm (Zwane and Van der Heijden 2007). The EM-algorithm is an iterative procedure where each iteration has an expectation (E) and a maximization (M) step. In the E-step the expectations of the missing values are found given the observed values and the fitted values under a model, here some log-linear model. The E-step yields completed data. Then, in the M-step, the log-linear model is fitted to the completed data and this updates the fitted values that are used in the next E-step. This proceeds until convergence. The algorithm has linear convergence, which may make the algorithm very slow. Yet the likelihood increases in each step and therefore convergence is guaranteed. We illustrate the EM-algorithm for the maximal model in the next section, but first we elaborate some theoretical properties of this approach.

In the lower right corner of Panel 2 of Table 1 the missing part of the population is estimated (compare the white area in Figure 1). This estimate is a by–product of the

estimation using the EM-algorithm. For example, the missed count for $X_1 = 0$ and $X_2 = 0$ is 1,112.4, and this value is found by assuming independence between A and B given X_1 and X_2, so that $4,510.8 \times 63.9/259 = 1,112.4$. This last step is made under the usual assumptions in population size estimation using two registers taking into account the covariates, that is, (i) perfect linkage, (ii) independence between A and B conditional on X_1 and X_2, (iii) for each of the four subpopulations the population is closed, and (iv) homogeneity of inclusion probabilities for A or B, conditional on X_1 and X_2. The use of the word "or" in assumption (iv) may come as a surprise as in many papers homogeneity is formulated as an assumption that should hold for both A and B. However, if it holds for only one of the registers, this is sufficient, see Chao et al. (2001) and Van der Heijden et al. (2012).

2.1. Maximal Models

The most complicated model that can be fitted is

$$log\,\pi_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2}, \tag{1}$$

with identifying restrictions that the parameters λ, λ_1^A, λ_1^B, $\lambda_1^{X_1}$, $\lambda_1^{X_2}$, $\lambda_{11}^{AX_2}$, $\lambda_{11}^{BX_1}$, $\lambda_{11}^{X_1X_2}$ and $\lambda_{11}^{AX_2}$ are free, and the other parameters are restricted to be zero. The two-factor interactions are closely related to odds ratios; for example, $exp\left(\lambda_{11}^{X_1X_2}\right)$ is the conditional odds ratio between X_1 and X_2. Another way to denote log-linear models is to use the highest fitted interactions to codify the model, the highest interactions implying the inclusion in the model of all lower order effects; for this model this corresponds to the notation $[AX_2][X_1X_2][BX_1]$. Model $[AX_2][X_1X_2][BX_1]$ has eight free parameters, namely an intercept, four main effects, and three interactions. This number of parameters corresponds to the number of counts in Panel 1 of Table 1, that is also eight. Notice that the term AX_1 is not included in the model, because when $A = 0$, X_1 is missing and unknown. Therefore only three counts are available for the term AX_1. On the other hand, for the term AX_2 four counts are available, namely (259 + 110), (539 + 177), 91 and 164, so this term is included in the model (and similarly for BX_1). A maximal model is also a saturated model in the sense that the fitted counts for a maximal model are equal to the observed counts. (Notice that violations of this model, such as dependence between A and B conditional on the covariates, cannot be tested. We come back to this issue in Section 3, where we investigate sensitivity to such model violations.)

The fitted values for this model are obtained with the EM-algorithm. The algorithm starts with the initial estimates $\hat{n}_{10(lk)}^{(0)}$ and $\hat{n}_{01(lk)}^{(0)}$, that are found by evenly distributing the observed frequencies $n_{10(l+)}$ and $n_{01(+k)}$ over the corresponding cells. In the first M-step, the log-linear model $[AX_2][X_1X_2][BX_1]$ is fitted to the completed data, with the cells corresponding to $(i,j) = (0,0)$ specified as structural zeros. This yields the estimates $\hat{\pi}_{ij(lk)}^{(1)}$, which are then used in the first E-step

$$\hat{n}_{10(lk)}^{(1)} = \frac{\hat{\pi}_{10(lk)}^{(1)}}{\hat{\pi}_{10(l+)}^{(1)}} n_{10(l+)}, \quad \hat{n}_{01(lk)}^{(1)} = \frac{\hat{\pi}_{01(lk)}^{(1)}}{\hat{\pi}_{01(+k)}^{(1)}} n_{01(+k)}, \tag{2}$$

to compute the updates $\hat{n}_{10(lk)}^{(1)}$ and $\hat{n}_{01(lk)}^{(1)}$. These estimates are then used in the second M-step to find the updates $\hat{\pi}_{ij(lk)}^{(2)}$, and so on until convergence is reached at iteration t.

Equation (1) also allows for an alternative way to estimate the four cells in the lower right part of Table 1. For example, the upper left element $1,112.4 = \exp(\hat{\lambda})$, as for the cell with indices $(i,j,k,l) = (0,0,0,0)$ the parameter values are zero except for the intercept. Similarly, $2,150.2 = \exp(\hat{\lambda} + \hat{\lambda}_1^{X_1})$. In other words, the parameters of the model are estimated and projected to cells that refer to the part of the population missed by both registers.

Model 1 can easily be extended when there are additional covariates. For example, consider the situation that in addition to X_1 being observed in A and X_2 in B a variable X_3 is observed in A and B; then the maximal and saturated model is $[AX_2X_3][X_1X_2X_3][BX_1X_3]$. And, as a second example, consider the situation that in addition to X_1 being observed in A and X_2 in B a variable X_4 is only observed in A, then the maximal model is $[AX_2][X_1X_2X_4][BX_1X_4]$.

For each of the three models discussed it is possible to investigate whether more restrictive models also fit the data. For example, for the model $[AX_2]$ $[X_1X_2][BX_1]$ it is useful to investigate whether one of the interactions can be eliminated without the fit deteriorating. For example, if the covariate X_1 is statistically independent from the covariate X_2, then the model becomes $[AX_2][BX_1]$, and under this model A and B are statistically independent, and not independent conditional on X_1 and X_2.

Example. For model 1 the likelihood ratio chi-square is zero with zero degrees of freedom. We may want to investigate whether imposing the additional restriction $\lambda_{kl}^{X_1X_2} = 0$ is allowed, so that the model becomes $[AX_2][BX_1]$. The difference between the likelihood ratio chi-squares for these two models is 3.2 (df is 1), which is not significant at the five percent level. The estimated population size under model 1 is 33,769.9, whereas for model 1 with $\lambda_{kl}^{X_1X_2} = 0$ it is 33,764.2, only marginally different. This corresponds to the odds ratio estimated from the four elements where X_1 and X_2 are both observed, 259, 539, 110, and 177, which yields a value of 0.7732 with a 95 percent confidence interval of (0.5842, 1.0234). The z-statistic to test whether the odds-ratio is significantly different from 1 is 1.798, which is not significant in a two-sided test but is significant in a one-sided test.

2.2. *Collapsibility, Active and Passive Variables*

The maximal models just discussed have interesting properties in terms of collapsibility over variables X_1 and X_2 (Van der Heijden et al. 2012). We use the following terminology. We use the word *marginalize* to refer to the contingency table formed by considering a subset of the original variables. We use the word *collapsibility* to refer to the situation that when a table is marginalized the population size estimate remains invariant. Using these terms the properties of maximal models can be easily explained using interaction graphs of the log-linear models involved. See Figure 2. Log-linear model $[AX_2X_3][X_1X_2X_3][BX_1X_3]$ has graph M_1. This is a maximal model. The log-linear model where X_1 and X_2 are conditionally independent given the variables A, B, and X_3 is $[AX_2X_3][BX_1X_3]$ and this model has graph M_2. What follows are three models where one of X_1, X_2, or X_3 is not available. In model M_3 the variable X_1 is not available, and the log-linear model is $[AX_2X_3][BX_3]$. In model M_4 the variable X_2 is not available, and the log-linear model is $[AX_3][BX_1X_3]$. Finally, in model M_5 the variable X_3 is not available, and the log-linear

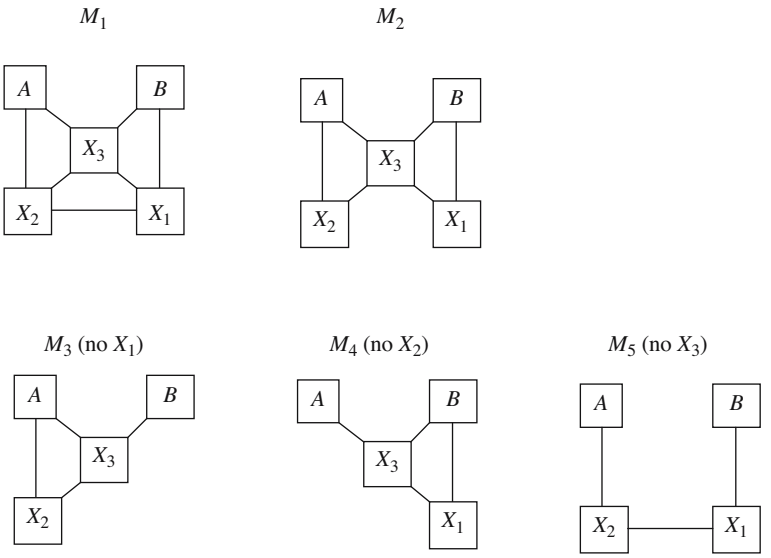


Fig. 2. Graph representations of some log-linear models.

model is $[AX_2][X_1X_2][BX_1]$. This last model is the model that is the focal point in the exposition in this article.

Interaction graphs are useful tools for assessing collapsibility. We make use of the concept of a *short path* (Whittaker 1990): two registers A and B are connected by a path if there is a sequence of adjacent edges connecting the variables A and B in the graph. A short path from A to B is a path that does not contain a sub-path from A to B. The rule is that when a covariate is on a so-called short path from A to B, the contingency table cannot be marginalized over this variable, and vice versa, that is when a covariate is not on a short path, the contingency table can be marginalized over this variable. We now discuss this for models M_1 to M_5. In M_1 the data cannot be marginalized over any of the variables X_1 to X_3. The reason is that there are two short paths, namely $A - X_2 - X_1 - B$ and $A - X_3 - B$, and each of the variables X_1 to X_3 is on one of the two short paths. In M_2 the data are collapsible over X_1 and over X_2, the reason being that the only short path is $A - X_3 - B$. In M_3 the data are collapsible over X_2, the reason being that the only short path is $A - X_3 - B$. In M_4 the data are collapsible over X_1, the reason being that the only short path is $A - X_3 - B$. In M_5 the data cannot be marginalized over X_1 and X_2, because both variables are on the short path $A - X_2 - X_1 - B$.

When a model (or graph) is collapsible over a variable, this means that in both the original model and collapsed model the same estimate of the population size is obtained. For example, models M_2, M_3, and M_4 yield the same population size estimate, and this estimate is identical to the population size estimate of model $[AX_3][BX_3]$. However, it may still be interesting to fit a model M_3, for example, because then this total population size estimate is spread out over the levels of variables X_1 and X_2. In Van der Heijden et al. (2012) the variables X_1 and X_2 in model M_2 are referred to as passive, in the sense that they do not have an impact on the estimate of the total population size. In contrast, variables X_1

and X_2 in model M_1 are referred to as active, because these variables do influence the total population size estimate.

Example. In the former section we saw that in model $[AX_2][X_1X_2][BX_1]$ the additional restriction $\lambda_{kl}^{X_1X_2} = 0$ does not deteriorate the fit, so that a more parsimonious model is $[AX_2][BX_1]$. As there is no short path any more between A and B, this means that we can marginalize over X_1 and X_2, showing that the population size estimate for model $[AX_2] \times [BX_1]$ is identical to the population size estimate for model $[A][B]$. We do not state that the original table should necessarily be marginalized over X_1 and X_2, because the original table can give insight into how the total population size is spread out over the levels of X_1 and X_2. Van der Heijden et al. (2009) and Van der Heijden et al. (2012) consider other examples with a larger number of covariates, namely five. They show that, by estimating the missing covariates and the number of individuals missed completely, the coverage of the population register can be evaluated in terms of the five covariates.

2.3. Precision and Sensitivity

Figure 1 illustrated that there are two estimation problems: estimating the missing covariates (the grey bitmap parts) and estimating the number of individuals (and their covariate values) missed by both A and B (the white parts in Figure 1). For both estimation problems we are interested in the precision when the model assumptions are true, and the sensitivity of the outcomes to deviations from the model assumptions.

We first discuss precision and start with the precision of the estimates for the missing covariates. Here precision is to be understood as an overall term referring to the variance of the estimates. Under the EM approach the model fitted is $[AX_2][X_1X_2][BX_1]$. As can be seen in Table 1, the odds ratio $(259 \times 110)/(539 \times 177) = 0.7732$, is used to calculate the expectations for the part of the table where X_1 is missing and the part where X_2 is missing. Under the model, the more precise this odds ratio, the more precise these expectations. This precision is directly related to the size of the population that is in both A and B: the larger this size, the smaller the standard error of the odds ratio and the standard errors of the estimates and the larger the precision.

The precision of the data for the individuals missed by A and B is the outcome of two sources: first, the precision of the estimates of the missing covariates that we just discussed, and, second, implied coverage. Precision of the estimates of the missing covariates has a direct impact on the precision of the data for the individuals missed. Consider again Table 1. Because, in Panel 2 of Table 1, the estimate $1,112.4 = 4,510.8 \times 63.9/259$, when the estimates 63.9 and 4,510.8 are imprecise, the estimate 1,112.4 will be imprecise as well.

The second source of imprecision is related to implied coverage. We explain this for $(X_1, X_2) = (0, 0)$. For the population register A the coverage of A implied by B is $259/(259 + 63.9) = 0.802$. However, for the police register B the coverage of B implied by A is only $259/(259 + 4,510.8) = 0.057$. The equation $1,112.4 = 4,510.8 \times 63.9/259$ shows that if either or both of these implied coverages is low, the estimated number of missed individuals is large relative to the number of individuals seen, and hence imprecise.

Estimates of the precision can be obtained using the parametric bootstrap (compare Buckland and Garthwire 1991). The parametric bootstrap provides a simple way to find the

confidence intervals when the contingency table is not fully observed. To compute the bootstrapped confidence intervals for a specific log-linear model, we need to first compute the population size under this model and the probabilities on the completed data under this model, that is, by including the cells that cannot be observed by design. A first multinomial sample is drawn given these parameters, and the sample is then reformatted to be identical to the observed data (for example, the sample in the format of Panel 2 of Table 1 is recoded into the format of Panel 1). The specific log-linear model used is then fitted to the resulting data, resulting in an estimate of the population size. Then this is repeated K times. By ordering the K bootstrap population size estimates, a percentile confidence interval can be constructed. We use this approach later on.

Up to this point we have discussed precision when the model assumptions are correct. We now discuss the sensitivity of the estimates to violations of the assumptions of the model. It is possible to investigate whether maximal models can be reduced by setting some parameters equal to zero. For example, in model M_5 (i.e., Equation 1) it is possible to test whether the parameter $\lambda_{kl}^{X_1X_2}$ is needed to give an adequate description of the data. However, it is not possible to test whether parameters that are *not* included in the maximal model, should be included. In other words, we cannot reject the MAR assumption using the data.

However, as was shown in this context by Gerritse et al. (2015b), it is possible to investigate for a particular data set how sensitive the outcome of the maximal model is to the assumption that certain parameters are zero. Take model M_5. The maximal model assumes that three two-factor interactions are zero, that is, $\lambda_{ik}^{AX_1} = \lambda_{jl}^{BX_2} = \lambda_{ij}^{AB} = 0$, and all three- and four-factor interactions are zero. Consider $\lambda_{ik}^{AX_1} = 0$. The maximal model is extended with a fixed parameter value for $\lambda_{ik}^{AX_1}$. We denote such a fixed parameter with the tilde $\tilde{\lambda}$, and the model to be fitted becomes

$$log\,\pi_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2} + \tilde{\lambda}_{ik}^{AX_1}. \tag{3}$$

Such a model can be fitted for a range of values of $\lambda_{11}^{AX_1}$. Appropriate values can be chosen by making use of the fact that log-linear parameters are closely related to odds ratios. Technically, the model may be fitted as a log-linear Poisson regression with offset $\exp\left(\tilde{\lambda}_{ik}^{AX_1}\right)$ (see Gerritse et al. 2015b, for details).

Gerritse (2016) argues that the sensitivity of outcomes of the analyses to violation of the independence assumption and to violation of perfect linkage is larger when the implied coverage is lower. In the absence of covariates this can be explained as follows. Let m_{ij} be the expected count for cell (i,j) $(i,j = 0,1)$, where m_{00} is the missing count to be estimated. Under independence the odds ratio is 1, that is, $m_{00}m_{11}/m_{01}m_{10} = 1$, so that $m_{00} = m_{01}m_{10}/m_{11}$. Under dependence with odds ratio θ, $m_{00}\theta = m_{01}m_{10}/m_{11}$. Thus, the smaller the overlap in cell (1,1), and hence the smaller the coverage, the larger the estimated value for cell (0,0), and this holds both for independence and dependence. In the same way, when links are missed, this increases the expected values m_{01} and m_{10} and decreases m_{11}, with the result that m_{00} is larger, and this effect is larger the smaller the overlap m_{11}.

Example. We carried out sensitivity analyses for the omission of parameters $\tilde{\lambda}_{ij}^{AB}$, $\tilde{\lambda}_{ik}^{AX_1}$, and $\tilde{\lambda}_{jl}^{BX_2}$. The results are shown in Table 2. Conditional odds ratios of 0.67 and 1.5 are used. For our example the model without the fixed parameters has an estimated missed

Table 2. Sensitivity analyses. The maximal model is $[AX_1][X_1X_2][BX_2]$*, where* $\hat{m}_{00} = 6{,}176$ *and* $\hat{N} = 33{,}770$*. Fixed conditional odds ratios are plugged in for* $\tilde{\lambda}_{ij}^{AB}$, $\tilde{\lambda}_{ik}^{AX_1}$ *and* $\tilde{\lambda}_{jl}^{BX_2}$.

Term	Size(OR)	$\hat{m}_{00}$	N	Size(OR)	$\hat{m}_{00}$	N
$\tilde{\lambda}_{ij}^{AB}$	0.67	4,117	31,711	1.5	9,264	36,858
$\tilde{\lambda}_{ik}^{AX_1}$		6,736	34,330		5,711	33,305
$\tilde{\lambda}_{jl}^{BX_2}$		6,136	33,730		6,220	33,814

population size of $\hat{m}_{00} = 6,176$ and an estimated population size of $\hat{N} = 33,770$. Violation of the model because there is direct dependence between A and B in the form of conditional odds ratios 0.67 or 1.5 has a large effect, because this leads to estimated missed population sizes of $\hat{m}_{00} = 4,117$ and $\hat{m}_{00} = 9,264$ respectively. Plugging in a missed odds ratio $\tilde{\lambda}_{jl}^{BX_2}$ has a minor effect on the estimated missed population size: for conditional odds ratios of 0.67 and 1.5 it leads to estimated values 6,136 and 6,220, both values being close to 6,176. However, plugging in a missed odds ratio $\tilde{\lambda}_{ik}^{AX_1}$ has a larger effect on the estimated missed population size: for conditional odds ratios of 0.67 and 1.5 it leads to estimated values 6,736 and 5,711.

2.4. *Extension to More than Two Registers*

The advantage of being able to use more than two registers is that the restrictive (conditional) independence assumption between variables A and B can be replaced by less restrictive assumptions. For example, in the situation of three registers without covariates, the saturated model is the model with all two-factor interactions. Now it is possible to search for more restrictive models that still describe the data well. One can consult the references provided in the introduction for details, see, for example, Bishop et al. (1975).

For three registers the problem of incomplete covariates has been studied by Zwane and Van der Heijden (2007), who show that for this problem the EM-algorithm can easily be adapted. Van der Heijden et al. (2012) discuss graph representations of the models and collapsibility, but do not touch incomplete covariates.

An interesting official statistics application is found in Gerritse et al. (2015a). The problem is to estimate the number of usual residents for the Dutch census 2011. Here usual residence is defined as, roughly, living in the Netherlands for a continuous period of twelve months before the reference time. Three registers are available, namely the population register, the employment register and a crime suspects register. Given this definition we are interested in a dichotomized version of duration, namely longer than or shorter than a year. From both the population register and the employment register residence duration can be derived (for details see Gerritse et al. 2015a), so when people are only in the population register or only in the job register a measurement for a persons' duration is available. For persons who are both in the population register and the employment register the overlapping durations are reconciled and dichotomized. The crime suspects register has no variable for duration. This is not problematic for persons who are also in the population register or the employment register, because then the residence variable of the latter can be used, but it is problematic for individuals who are only in the crime suspects

Table 3. Polish individuals by the population register, the employment register and the crime suspects register, by usual residence. The counts for the two cells labeled "missing" add up to 1,043. Data from Gerritse et al. (2015a).

Usual residence	Population	Employment	Crime suspects	
			Yes	No
No	Yes	Yes	32	3,523
		No	34	3,225
	No	Yes	149	60,190
		No	missing	0
Yes	Yes	Yes	183	21,309
		No	195	14,052
	No	Yes	81	20,216
		No	missing	0

register. See Table 3, taken from Gerritse et al. (2015a), where counts for people born in Poland and registered in one or more of the three registers are displayed. We find a $2 \times 2 \times 2$ table for residence duration longer than a year and a $2 \times 2 \times 2$ table for residence duration shorter than a year, where two cells are indicated with the label 'missing'. As these two cells refer to persons only in the police register, the sum of the counts for the two cells is known, namely 1,043. The EM-algorithm is used to distribute these 1,043 persons over the two cells under some log-linear model, and the parameters of the final model are projected on the two (0,0,0) cells to find the number of persons missed by all three registers. We refer to the Supplementary materials, Section 1, for further details (Available online at: www.dx.doi.org/10.1515/jos-2018-0011).

A similar example can be found in Héraud-Bousquet et al. (2012), where there are three registers, and in two of the registers place of birth is available, but in a third register it is not. For those individuals only in the third register the missing values are imputed using multiple imputation. Multiple imputation has wider application when covariates are continuous instead of categorical, when the EM-algorithm loses its simplicity. Zwane and Van der Heijden (2008) apply multiple imputation using predictive mean matching in this situation.

3. Simulations

Earlier simulation results can also be found in Zwane and Van der Heijden (2007) for two registers and two partially observed covariates. These results are not completely transparent as the covariates used in the simulation are correlated continuous variables that are dichotomized. Thus the true model structure from which samples are drawn cannot easily be understood from the perspective of a log-linear model. In the simulations that we present here the true model is a log-linear model in which marginal probabilities and conditional odds ratios are specified to describe the dependence between the variables. We refer to the Supplementary materials, Section 2, for details on how true models are generated (available online at: www.dx.doi.org/10.1515/jos-2018-0011).

We carried out simulations to compare the behaviour of the classical model (denoted by LL), where incomplete covariates are ignored, with the model where incomplete

covariates are completed with the EM-algorithm (denoted by EM). For each choice of conditional odds ratios this yields population probabilities from which we sample. In each instance of the simulation study 25,000 samples are taken. For LL, for each sample the classical model $[A][B]$ is estimated on the marginal table formed from A and B, where the sampled count in cell $(A, B) = (0, 0)$ is made missing, and subsequently estimated assuming independence between A and B. Similarly, for EM for the same samples the model $[AX_2][X_1X_2][BX_1]$ is estimated, where the four cells where $(A, B) = (0, 0)$ are made missing.

In the first simulation study the population model is $[AX_2][X_1X_2][BX_1]$, so that the model estimated by EM is identical to the true model. The prespecified marginal probabilities are $P(A = 1) = 0.3$, $P(B = 1) = 0.3$, $P(X_1 = 0) = 0.5$ and $P(X_2 = 0) = 0.5$. Conditional odds ratios different from 1 are specified between A and X_2, between X_1 and X_2 and between B and X_1, so that the true model is $[AX_1][X_1X_2][BX_2]$. We denote the conditional odds ratio between A and X_2 by $OR(A, X_2)$. Note that the theoretical results in earlier sections show that, when one of the three conditional odds ratios $OR(A, X_2)$, $OR(B, X_1)$ or $OR(X_1, X_2)$ is 1, the model is collapsible over the covariates so that identical results are found for LL and EM. Therefore conditional odds ratios equal to 1 are not used. Also note that, for example, $OR(A, X_2) = OR(B, X_1) = 0.5$ leads to the same population probabilities as $OR(A, X_2) = OR(B, X_1) = 2$, as this is equivalent to the recoding of levels 0 and 1 in X_1 and X_2. Therefore we only use odds ratios of 2.

In Table 4 results are reported. In the upper part the true population size is 1,000. We first plug in conditional odds ratios of moderate size. In the first two lines the three odds ratios plugged in are $OR(A, X_2) = OR(B, X_2) = OR(X_1, X_2) = 2$. The average observed n, over 25,000 samples is 511, which is approximately $1,000 \times (1 - 0.7 \times 0.7)$, where 0.7 is the probability of not being selected in A or B. Note that the implied coverage, derived by collapsing over the covariates, is low, namely 0.3, that is, given population A, when linking to population B 70 percent of the observations in B were not seen before (in A). Under LL, the average estimated mean is 1,014.9 (with SE = 76.3 calculated over the 25,000 samples), the average estimated median is 1,009.9 (with SE = 76.4) and the RMSE is 77.7. Under EM, the average mean is 1,004.9 (SE = 75.4), the average median is 1,000.1 (SE = 75.6) and RMSE is 75.6. For $N = 1,000$ two other triples of conditional odds ratios are investigated. As expected, under EM the average mean and (in particular) the average median under the log-linear model are very close to the population value, where under LL there is some bias. Notice that the median has less bias than the mean, due to the non–normality of the distribution of estimates. With the population size of 1,000, the RMSE's of LL and EM are close. In the following four instances the population size is 10,000. The bias of the means and medians become a bit smaller, and as the standard errors become smaller (due to the increased population size) the RMSE's of EM become smaller than those of LL. The same holds for $N = 50,000$. It seems that the bias found for LL is approximately equally large but opposite for conditional odds ratios $OR(X_1, X_2) = 0.5$ and $OR(X_1, X_2) = 2$, and this is in contrast to the results in Zwane and Van der Heijden (2007).

In Table 5 the coverage is higher, with $P(A = 1) = P(B = 1) = 0.6$. When the coverage is higher, the part of the population missed is smaller, and violation of assumptions will have a smaller effect. This is also apparent by comparing Table 5 with Table 4, which

Table 4. Simulations under the model, with lower coverage. $P(A = 1) = 0.3$, $P(B = 1) = 0.3$, $P(X_1 = 0) = 0.5$ *and* $P(X_2 = 0) = 0.5$. *The conditional odds ratios refer to* $OR(A, X_2)$, $OR(B, X_1)$ *and* $OR(X_1, X_2)$.

	N	Odds ratios	Mean (n)	Mean	Median	SE mean	SE med.	RMSE
LL	1,000	2,2,0.5	511	1,014.9	1,009.9	76.3	76.4	77.7
EM				1,004.9	1,000.1	75.4	75.6	75.6
LL	1,000	2,2,2	509	995.2	990.8	73.8	73.9	74.0
EM				1,005.3	1,000.6	75.4	76.0	76.1
LL	1,000	2,2,5	508	984.1	979.1	72.4	72.6	74.1
EM				1,007.1	1,000.6	77.3	77.5	77.6
LL	10,000	2,2,0.5	5109	10,104.0	10,098.2	237.0	237.1	258.8
EM				10,005.1	10,000.3	234.2	234.3	234.3
LL	10,000	2,2,2	5092	9,910.7	9,906.2	228.7	228.8	245.6
EM				10,008.3	10,003.3	234.4	234.5	234.6
LL	10,000	2,2,5	5081	9,792.0	9,789.0	226.0	226.0	307.1
EM				10,007.2	10,003.3	239.5	239.5	239.6
LL	50,000	2,2,0.5	25,546	50,502.5	50,497.4	531.8	531.8	731.6
EM				50,007.4	50,004.5	524.2	524.2	524.3
LL	50,000	2,2,2	25,456	49,522.5	49,524.1	514.5	514.5	702.0
EM				50,008.2	50,010.5	527.2	527.3	527.3
LL	50,000	2,2,5	25,402	48,935.5	48,932.5	499.4	499.4	1,175.8
EM				50,004.7	49,998.3	529.0	529.1	529.1

Table 5. Simulations. $P(A = 1) = 0.6$, $P(B = 1) = 0.6$, $P(X_1 = 0) = 0.5$ *and* $P(X_2 = 0) = 0.5$. *The conditional odds ratios refer to* $OR(A, X_2)$, $OR(B, X_1)$ *and* $OR(X_1, X_2)$.

	N	Odds ratios	Mean (n)	Mean	Median	SE mean	SE med.	RMSE
LL	10,000	2,2,0.5	9906	10,001.7	10,001.9	10.8	10.8	10.9
EM				9,999.9	10,000.0	10.8	10.8	10.8
LL	10,000	2,2,2	9903	9,998.1	9,998.2	11.0	11.0	11.2
EM				9,999.9	10,000.0	11.0	11.0	11.0
LL	10,000	2,2,5	9901	9,995.9	9,996.0	11.0	11.0	11.8
EM				10,000.0	10,000.1	11.1	11.1	11.1

shows that the bias for LL in Table 5 is smaller than the bias in Table 4. The bias in EM is negligible, in particular when N increases.

Simulations suggested by the Census Coverage Survey for England and Wales are reported in the Supplementary materials, section 3 (available online at: www.dx.doi.org/10.1515/jos-2018-0011). We also did simulations where the model $[AX_2][X_1X_2][BX_1]$, assumed in the EM approach, is violated. These results can be found in the Supplementary materials, section 4 (available online at: www.dx.doi.org/10.1515/jos-2018-0011). Overall the simulations show that, when the MAR assumptions are fulfilled, the EM approach does better, though sometimes only slightly better, than the traditional approach. When the MAR assumptions are not fulfilled, the bias can be substantial, in particular when the inclusion probabilities are low.

4. Novel Application: The Same Variable Measured in Both Registers

We present a novel application of the above methodology. It concerns two registers that both measure the same variable, and the measure in one register is generally considered to be more trustworthy, or valid, than the measure of the same variable in the other register. This is closely related to the classical two-phase sampling problem, where there is an inexpensive but low quality measurement which can be obtained from a large sample, and a more expensive and more accurate approach which is used on a subsample. Two-phase sampling concentrates on combining the small sampling variance of the large sample measure with the measurement accuracy of the small sample measure. In our case we will apply the EM-algorithm to complete the missing information on the highest quality measure, and additionally to provide this information for statistical units which are missed in both the registers (a situation which cannot generally be handled by two-phase sampling). The example we deal with is the number of serious road injuries in the Netherlands. The first author was consulted by the Ministry of Transport with the question whether the current methodology applied for estimating this number was sufficiently appropriate. In the Netherlands the number of serious road injuries is important because it is used for assessing the road safety target.

In the Netherlands there are two parties that can deliver information on serious road injuries, namely the police and hospitals. Both parties are usually present after the occurrence of such an accident. The police are supposed to record the accident and its cause in the police crash record database, but this regularly does not happen for some

Table 6. Road accidents in the Netherlands in 2000, from Reurings and Stipdonk (2011). Motorized vehicle involved X_1 is only observed in the Police register (A) and Motorized vehicle involved X_2 is only observed in hospital register (B). Levels of X_1 and X_2 are 1 = yes, 2 = no.

Panel 1: Observed counts

		$B=1$		$B=0$	
		$X_2=1$	$X_2=2$	X_2 missing	Total
$A=1$	$X_1=1$	5,970	287	1,351	7,608
	$X_1=2$	28	256	70	354
$A=0$	X_1 missing	2,947	4,120	–	7,067
Total		8,945	4,663	1,421	15,029

Panel 2: Fitted values under $[AX_1][X_1Y]$

		$B=1$		$B=0$	
		$X_2=1$	$X_2=2$	X_2 missing	Total
$A=1$	$X_1=1$	5,970.0	287.0	1,351.0	7,608.0
	$X_1=2$	28.0	256.0	70.0	354,0
$A=0$	$X_1=1$	2,509.6	120.6	567.9	3,198.1
	$X_1=2$	437.4	3,999.4	1,093.6	5,530.4
Total		8,945.0	4,663.0	3,082.5	16,690.5

Panel 3: Fitted values under $[AX_2][X_1X_2][BX_1]$

		$B=1$		$B=0$		
		$X_2=1$	$X_2=2$	$X_2=1$	$X_2=2$	Total
$A=1$	$X_1=1$	5,970.0	287.0	1,289.0	62.0	7,608.0
	$X_1=2$	28.0	256.0	6.9	63.1	354,0
$A=0$	$X_1=1$	2,933.2	2,177.6	633.3	470.2	6.214,3
	$X_1=2$	13.8	1,942.4	3.4	478.8	2.438,4
Total		8,945.0	4,663.0	1.932,6	1,074.1	16,614.7

reason, such as that it is not clear which police officer has to file the accident report, or that the injury is not considered very serious. The hospital that treats the seriously injured, can report the cause of the injury in the hospital inpatient registry but this is sometimes forgotten and then such a patient's connection to a traffic accident is lost. Thus there are two register sources that both have coverage problems. Many details of the registration by the police and the hospitals can be found in Reurings and Stipdonk (2011), who report research conducted at the SWOV Institute for Road Safety Research. They state that the police database in particular suffers from serious underreporting, and is inaccurate in indicating injury severity, whereas the hospital database is inaccurate in indicating that a patient was involved in a road crash but in principle contains all serious road injuries.

For the year 2000 Reurings and Stipdonk (2011) present the data in the upper panel of Table 6. (We refer to their paper for a detailed discussion regarding the linking of the two

registers.) The police register has a larger undercoverage than the hospital register. Yet it is reasonable to assume that, where the police registers do record the mode of transport of injured persons, they do this more accurately than the hospital. The reason is that assessing the cause of accidents is a more important function for the police, because liability plays a role, than of the hospital, which is more concerned about the type of serious casualty and who will be focused more on health related issues than on the cause and details of the accident. Notice that in the 2×2 subtable that is fully observed, there are 287 joint classifications not in agreement where the police recorded the involvement of a motorized vehicle but the hospital recorded that no motorized vehicle was involved, and 29 vice versa.

As it turns out, two approaches can be taken for solving the missing data problem and subsequently estimating the number of accidents missed by both registers for the $2 \times 2 \times 2 \times 2$ table. We discuss these options and then generalize to a situation where the number of levels of the variables X_1 and X_2, Cause of the accident, is increased from two to seven.

4.1. The $2 \times 2 \times 2 \times 2$ Table

As a first approach, Reurings and Stipdonk (2011) set up a system of linear equations to estimate the number of seriously injured. They report 10,804 seriously injured in motorized accidents and 5,891 seriously injured in non-motorized accidents. Using a log-linear modelling framework that includes missing data we can obtain their results as follows. We define a new variable Y with three levels, namely $(X_2 = 1, B = 1), (X_2 = 2, B = 1)$ and $(X_2 =$ missing). We then fit model $[AX_1][X_1Y]$ with X_1–values missing for $A = 0$. The estimates using our procedure should in principle be identical to Reurings and Stipdonk's estimates but they are slightly different (probably due to rounding), see Panel 2 of Table 6, in the two last lines, and these lead to estimates of (7,608 + 3,198.1 =) 10,806.1 for motorized and 5,884.4 for non-motorized accidents. In this approach the relative frequencies for 5,970, 287, and 1,351 are identical to those for 2,509.6, 120.6, and 567.9, and similarly for 28.0, 256.0, and 70.0 to 437.4, 3.999.4, and 1,093.6, while at the same time the counts 2,947 and 4,120 are split up over the missing levels of X_1. Notice that we estimate that only (567.9 + 1,093.6 =) 1,661.5 accidents with serious road injuries are missed by both registers, which is approximately ten percent of the total estimated population size. 95 percent confidence intervals of the estimates 10,806.1 and 5,884.4 are obtained using the parametric bootstrap by the percentile method with 10,000 bootstrap samples, and this yields 10,532 – 11,054 and 5,512 – 6,305.

As the second approach, we apply the methodology to this table that we applied before in Table 1. That is, we assume that the hospital Cause of accident is missing for those accidents only registered by the police whereas we assume that the police Cause of accident is missing for those accidents only registered by the hospital, and fit model $[AX_2][X_1X_2][BX_1]$. See Panel 3 in Table 6. This leads to very different estimates for motorized and non-motorized accidents, namely 13,823 (95 percent CI 13,568 – 14,072) and 2,791 (2,551 – 3,037). In this approach the four odds ratios for all combinations of register A and B are assumed to be equal, and the counts 2,947 and 4,120 are now split up in a way different from the first approach.

We make a few remarks. First, when we compare both approaches we have a preference for our own approach using model $[AX_2][X_1X_2][BX_1]$ over the approach by Reurings and

Stipdonk using model $[AX_1][X_1Y]$. This preference is not based on model fit as both models are saturated and have a perfect fit. Instead we make a judgement based on a professional opinion. We find it is reasonable to assume that, for example, the count 2,947 for which X_1 is missing, should be split over motorized and non-motorized in the same way as when X_1 is not missing. Our approach is plausible, simple and transparent, as in the saturated model we present here the estimates can be found by hand. The plausibility of the approach by Reurings and Stipdonk can be argued, but it is less simple and transparent, as it needs an iterative procedure, and in the next section we will see that it can have numerical problems. We obtain additional support from the model-based bootstrap applied to $[AX_2][X_1X_2][BX_1]$ which gives smaller confidence intervals (14,072 – 13,568 = 504 and 486) than the estimates under model $[AX_1][X_1Y]$ (11,054 – 10,532 = 522 and 793). This strategy is in line with Elliott and Little (2000)'s principles for choosing between saturated models where, after a series of principles fail to distinguish models, then principle 5 suggests using the model that gives estimates with reduced variance.

Second, when we compare our log-linear modelling procedure with the approach by Reurings and Stipdonk of solving a system of linear equations, a number of differences are apparent. Our approach is flexible because extra variables can be incorporated easily. This will in principle also be the case in Reurings and Stipdonk's approach. However, when estimates become unstable due to low observed counts, our approach allows for constraints on the log-linear parameters that can stabilize the model. The modelling approach has the advantage that it always produces maximum likelihood estimates, whereas solving a system of linear equations only leads to maximum likelihood estimates when the estimates are non-negative. Also, we think that the flexibility of our approach is important, because Reurings and Stipdonk (2011) report that they applied the method three times separately, namely for the covariates transport mode (reported here), region and injury severity. This has the drawback that three different estimates of the population size will result. In our methodology it is easy to include all three covariates simultaneously, and this will yield a single total population size that is consistent over the three covariates. It also allows investigation of the relationships between the three covariates.

As a third remark, in situations like this a practical approach is often taken (Reurings and Stipdonk 2011 are a noteworthy exception) when a measure of some variable in register A is considered more trustworthy than a different measure of the same variable in register B, so after linking registers A and B a new, composite variable is created that makes the best of the information. In this new measure we fill in the values of the variable from register A when it is available, we fill in the values of the variable from register B for the observations that were missed by register A, and some ad hoc solution is found for the observations that were missed by both registers. In the approaches presented here, however, for those observations that were missed by register A we translate the values in register B into what would have been found in register A using the subtable of $A = 1$ and $B = 1$ to give the structure for those observations only found in register B, 2,947 and 4,120 at the bottom of Panel 1 of Table 1.

Last, notice that the odds ratio in this observed subtable is typically very large (in the upper part of 6 it is almost 200), and in both approaches the odds ratio for the subtable of $A = 1$ and $B = 1$ is used to find the estimates in the subtables of $A = 0$ and $B = 1$.

4.2. The $2 \times 2 \times 7 \times 7$ *Table*

The reason that the Ministry asked Van der Heijden for a consultation had to do with a generalization of the method applied by the SWOV Institute for Road Safety Research. See Table 7 taken from Reurings and Bos (2012, 25) where we find for 2010 a much more detailed coding of motorized mode of transport: where in Table 6 this only had one coding, it now has six codings, namely "Sitting in car", "Driving motorbike", "Driving moped", "Bicycles in motorized accident", "Pedestrians in motorized accident", and "Other in motorized accident". Of course, this finer coding into seven levels can be useful for assessing the cause of a rise or decline in accidents. Notice the occasional low off-diagonal counts, that are attractive because they make the data plausible (we do not want "non-motorized" to be mixed up a lot with "sitting in car". A second difference between the data for the years 2000 and 2010 is that the police registered many fewer accidents: in 2000 the number in the police register was around 7,000 compared with 8,000 missed by the police but found in the hospital registration, but in 2010 these numbers are approximately 3,500 and 14,000. In the same period, the quality of the hospital register went up: in 2000 1,400 accidents were observed by the police but not by the hospital, but in 2010 this was only approximately 400.

The SWOV Institute for Road Safety Research generalized their approach of using a system of linear equations and found unstable estimates for some cells, including estimated counts that were negative. Using log-linear model $[AX_1][X_1Y]$, where Y has eight categories, the EM-algorithm also produces unstable results in the sense that convergence is not reached with 10^6 iterations, where in that last iteration two lines of estimates where $A = 0$ consisted of 0's only. Therefore we will only present results for the approach using model $[AX_2][X_1X_2][BX_1]$.

Table 7. Road accidents in the Netherlands in 2010. Data from Reurings and Bos (2012, 25). Motorized vehicle involved X_1 is only observed in Police register (A) and Motorized vehicle involved X_2 is only observed in hospital register (B). m.a. = motorized accident.

Observed counts

			$B = 1$				$B = 0$		
			X_2				X_2		
X_1	1	2	3	4	5	6	7	missing	Total
$A = 1$									
1. Sitting in car	856	7	12	26	61	62	18	130	1,172
2. Driving motorbike	3	261	33	0	7	5	2	20	331
3. Driving moped	7	83	504	19	8	60	21	47	749
4. Bicycles in m.a.	55	2	10	523	38	29	139	96	892
5. Pedestrians in m.a.	9	0	2	11	208	33	3	35	301
6. Other in m.a.	20	1	18	4	7	17	2	22	91
7. Non-motorized	2	0	0	9	1	7	82	12	113
$A = 0$									
missing	1,100	860	1,530	844	482	540	8,578	–	13,934
Total	2,052	1,214	2,109	1,436	812	753	8,845	362	17,583

Table 8. Motorized vehicle involved X_1 is only observed in Police register (A) and Motorized vehicle involved X_2 is only observed in hospital register (B). Fitted values (rounded) under model $[AX_2][X_1X_2][BX_1]$.

	$B=1$								$B=0$							
	X_2								X_2							
X_1	1	2	3	4	5	6	7	Total	1	2	3	4	5	6	7	Total
$A=1$																
1. Sitting in car	856.0	7.0	12.0	26.0	61.0	62.0	18.0	1,042.0	106.8	0.9	1.5	3.2	7.6	7.7	2.2	129.9
2. Driving motorbike	3.0	261.0	33.0	0.0	7.0	5.0	2.0	311.0	0.2	16.8	2.1	0.0	0.5	0.3	0.1	20.0
3. Driving moped	7.0	83.0	504.0	19.0	8.0	60.0	21.0	702.0	0.5	5.6	33.7	1.3	0.5	4.0	1.4	47.0
4. Bicycles in m.a.	55.0	2.0	10.0	523.0	38.0	29.0	139.0	796.0	6.6	0.2	1.2	63.1	4.6	3.5	16.8	96.0
5. Pedestrians in m.a.	9.0	0.0	2.0	11.0	208.0	33.0	3.0	266.0	1.2	0.0	0.3	1.4	27.4	4.3	0.4	35.0
6. Other in m.a.	20.0	1.0	18.0	4.0	7.0	17.0	2.0	69.0	6.4	0.3	5.7	1.3	2.2	5.4	0.6	21.9
7. Non-motorized	2.0	0.0	0.0	9.0	1.0	7.0	82.0	101.0	0.2	0.0	0.0	1.1	0.1	0.8	9.7	11.9
$A=0$																
1. Sitting in car	989.1	17.0	31.7	37.1	89.1	157.2	578.3	1,899.5	123.4	2.1	4.0	4.6	11.1	19.6	72.1	236.9
2. Driving motorbike	3.5	634.1	87.2	0.0	10.2	12.7	64.3	812.0	0.2	40.8	5.6	0.0	0.7	0.8	4.1	52.2
3. Driving moped	8.1	201.6	1,331.8	27.1	11.7	152.1	674.7	2,407.1	0.5	13.5	89.2	1.8	0.8	10.2	45.2	161.2
4. Bicycles in m.a.	63.6	4.9	26.4	745.6	55.5	73.5	4,465.7	5,435.2	7.7	0.6	3.2	89.9	6.7	8.9	538.6	655.6
5. Pedestrians in m.a.	10.4	0.0	5.3	15.7	303.8	83.7	96.4	515.3	1.4	0.0	0.7	2.1	40.0	11.0	12.7	67.9
6. Other in m.a.	23.1	2.4	47.6	5.7	10.2	43.1	64.3	196.4	7.4	0.8	15.2	1.8	3.3	13.7	20.5	62.7
7. Non-motorized	2.3	0.0	0.0	12.8	1.5	17.7	2,634.4	2,668.7	0.3	0.0	0.0	1.5	0.2	2.1	313.0	317.1

Table 9. Parametric bootstrap point estimates of causes according to the police, with 95 percent confidence interval (percentile method) and median, under model $[AX_2][X_1X_2][BX_1]$.

	Mean	2.5 percent	Median	97.5 percent
1. Sitting in car	3,307.1	2,997.9	3,300.6	3,644.9
2. Driving motorbike	1,195.0	1,071.2	1,190.8	1,336.8
3. Driving moped	3,317.2	2,996.6	3,312.6	3,657.6
4. Bicycles in m.a.	6,981.8	6,382.4	6,980.5	7,583.1
5. Pedestrians in m.a.	884.7	749.4	881.0	1,046.1
6. Other in m.a.	350.8	238.4	344.3	498.9
7. Non-motorized	3,099.3	2,565.1	3,094.4	3,646.8

The estimates under model $[AX_2][X_1X_2][BX_1]$ can be found in Table 8. In order to investigate the stability of the estimates we used a parametric bootstrap. The results are reported in Table 9. The seven estimated total numbers of severely injured are rather stable.

We conclude that, where classification by mode of transport in 2000 was stable, the refined classification of "motorized" into six categories in 2010 is usable for policy purpose when model $[AX_2][X_1X_2][BX_1]$ is applied.

5. Discussion

In this article we have presented a methodological framework that may be useful for the production of official statistics based on linked registers where additional categorical auxiliary variables are available. The methodology has potential for simultaneously solving the problems of undercoverage and of missing covariate values for those persons who are missed in some or all of the registers. This corresponds to solving the missing data problem for the grey bitmap and white parts in Figure 1.

5.1. Extensions

The EM-algorithm can also be used to solve the problem of missing data in covariates that are incompletely measured. There are many reasons why such data may be missing, including administrative errors or lags in recording data. If there is only a single register this is a simple missing data problem, but in the case of more than one register the extra information can help to complete these variables. The software we employ, the CAT-procedure in R, is able to handle this problem (Meng and Rubin 1991; Schafer 1997a,b).

Multiple imputation provides an alternative method for dealing with missing values in covariates. It was used, next to EM, by Gerritse et al. (2015a) and they argue that in their example, multiple imputation is more flexible. Their point is that in Table 3 the persons in the two cells labelled missing are most similar to persons not in the population register, and imputing from this subpopulation is easily accomplished using multiple imputation. But this approach is separate from the estimation of the unobserved part of the population, and does not benefit from the integrated way of dealing with these two issues.

Multiple imputation is however more natural in the case of continuous covariates, as used in Zwane and Van der Heijden (2008). Further research into the benefits of improving estimation using continuous covariates is also desirable. A more general strategy for

official statistics from linked registers which includes options for using categorical and continuous auxiliary variables in the estimation could then emerge, and an important element of that would be to have more examples of the usefulness of the approaches presented in this article.

If the framework is used to produce register based official statistics in a complex system with many registers, then it is more challenging to devise a procedure which ensures consistency between different outputs. Unless all the registers are linked, using the EM approach for different groups of registers using the same covariates, as would be likely in the case for age and gender, would lead to inconsistent outcomes. It is an open research question of how to build in this consistency.

The approaches presented here deal only with the problem of undercoverage. However, many registers also contain overcoverage, and this can have an effect on the undercoverage estimation by increasing the number of records to be linked. This will generally inflate population size estimates by inflating the number of records appearing in only one register, though it could have the opposite effect if the overcovered records appear in both registers. Zhang (2015) provides a framework for models to deal with overcoverage error, but it is important to have at least one source that does not suffer from overcoverage in order to make a suitable adjustment. More work is needed on how the estimation of undercoverage and overcoverage can be integrated into a set of procedures which can be applied in a wide range of situations including the production of official statistics.

5.2. Conclusion

The simulation studies show that, in comparison with the classical method where those partially observed covariates are ignored, the EM approach performs slightly better when the underlying MAR assumption and the conditional independence assumption for inclusion in the registers is met. When these assumptions are violated, both models can be severely biased.

In the last example in this article we showed how this missing data approach can be applied to the situation where a covariate of interest is measured in both registers.

Theoretically, the methodology can also be used when the number of covariates is large, where stability can be improved by making some of the covariates passive (compare Van der Heijden et al. 2012). In this instance there is little practical experience and we hope that this methodology will be used more so that the practical benefits become clearer.

6. References

Alho, J.M. 1990. "Logistic Regression in Capture-Recapture Models." *Biometrics* 46(3): 623–635. Doi: http://dx.doi.org/10.2307/2532083.

Baker, S.G. 1990. "A Simple EM Algorithm for Capture-Recapture Data with Categorical Covariates (with discussion)." *Biometrics* 46: 1193–1197. Doi: http://dx.doi.org/10.2307/2532461.

Bell, W.R. 1993. "Using Information from Demographic Analysis in Post-Enumeration Survey Estimation." *Journal of the American Statistical Association* 88(423): 1106–1118. Doi: http://dx.doi.org/10.2307/2290805.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. *Discrete Multivariate Analysis, Theory and Practice*. New York: McGraw-Hill. Doi: http://dx.doi.org/10.1007/978-0-387-72806-3.

Brown, J., O. Abbott, and I. Diamond. 2006. "Dependence in the 2001 One-Number Census Project." *Journal of the Royal Statistical Society: Series A* (Statistics in Society) 169: 883–902. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00431.x.

Brown, J., O. Abbott, and P.A. Smith. 2011. "Design of the 2001 and 2011 Census Coverage Surveys for England and Wales." *Journal of the Royal Statistical Society: Series A* (Statistics in Society) 174(4): 881–906. Doi: http://dx.doi.org/10.1111/j.1467-985X.2011.00697.x.

Brown, J.J., I.D. Diamond, R.L. Chambers, L.J. Buckner, and A.D. Teague. 1999. "A Methodological Strategy for a One-Number Census in the UK." *Journal of the Royal Statistical Society: Series A* (Statistics in Society) 162(2): 247–267. Doi: http://dx.doi.org/10.1111/1467-985X.00133.

Buckland, S. and P. Garthwire. 1991. "Quantifying Precision of Mark-Recapture Estimates Using the Bootstrap and Related Methods." *Biometrics* 47: 255–268. Doi: http://dx.doi.org/10.2307/2532510.

Chao, A., P. Tsay, S. Lin, W. Shau, and D. Chao. 2001. "The Applications of Capture-Recapture Models to Epidemiological Data." *Statistics in Medicine* 20: 3123–3157. Doi: http://dx.doi.org/10.1002/sim.996.

Darroch, J., S. Fienberg, G. Glonek, and B. Junker. 1993. "A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability." *Journal of the American Statistical Association* 88: 1137–1148. Doi: http://dx.doi.org/10.2307/2290811.

Elliott, M.R. and R.J.A. Little. 2000. "A Bayesian Approach to Combining Information from a Census, a Coverage Measurement Survey, and Demographic Analysis." *Journal of the American Statistical Association* 95: 351–362. Doi: http://dx.doi.org/10.1080/01621459.2000.10474205.

Fienberg, S., M. Johnson, and B. Junker. 1999. "Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Multiple Lists." *Journal of the Royal Statistical Society: Series A* 162: 383–406. Doi: http://dx.doi.org/10.1111/1467-985X.00143.

Gerritse, S.C. 2016. "An Application of Population Size Estimation to Official Statistics. Sensitivity of Model Assumptions and the Effect of Implied Coverage." Utrecht University (dissertation), Utrecht, 2016. Available at: https://dspace.library.uu.nl/handle/1874/337476 (accessed February 1, 2018).

Gerritse, S.C., B.F.M. Bakker, and P.G.M. van der Heijden. 2015a. "Different Methods to Complete Datasets Used for Capture-Recapure Estimation: Estimating the Number of Usual Residents in the Netherlands." *Statistical Journal of IAOS* 31: 613–627. Doi: http://dx.doi.org/10.3233/SJI-150938.

Gerritse, S.C., P.G.M. van der Heijden, and B.F.M. Bakker. 2015b. "Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Loglinear Models." *Journal of Official Statistics* 31: 357–379. Doi: http://dx.doi.org/10.1515/jos-2015-0022.

Griffin, R.A. 2014. "Potential Uses of Administrative Records for Triple System Modeling for Estimation of Census Coverage Error in 2020." *Journal of Official Statistics* 30: 177–189. Doi: http://dx.doi.org/10.2478/jos-2014-0012.

Héraud-Bousquet, V., F. Lot, M. Esvan, F. Cazein, C. Laurent, J. Warszawski, and A. Gallay. 2012. "A Three-Source Capture-Recapture Estimate of the Number of New HIV Diagnoses in Children in France from 2003–2006 with Multiple Imputation of a Variable of Heterogeneous Catchability." *BMC infectious diseases* 12(1) : 1. Doi: http://dx.doi.org/10.1186/1471-2334-12-251.

Huggins, R.M. 1989. "On the Statistical Analysis of Capture Experiments." *Biometrika* 76(1): 133–140. Doi: http://dx.doi.org/10.1093/biomet/76.1.133.

International Working Group for Disease Monitoring and Forecasting. 1995. "Capture-Recapture and Multiple Record Systems Estimation. Part I. History and Theoretical Development." *American Journal of Epidemiology* 142: 1047–1058. Doi: http://dx.doi.org/10.1093/oxfordjournals.aje.a117558.

Madigan, D. and J.C. York. 1997. "Bayesian Methods for Estimation of the Size of a Closed Population." *Biometrika* 84: 19–31. Doi: http://dx.doi.org/10.1093/biomet/84.1.19.

Meng, X.L. and D.B. Rubin. 1991. "IPF for Contingency Tables with Missing Data via the ECM Algorithm." In Proceedings of the Statistical Computing Section of the American Statistical Association, 244–247. Washington D.C.: American Statistical Association.

Pelle, E., D.J. Hessen, and P.G.M. van der Heijden. 2016. "A Log-Linear Multidimensional Rasch Model for Capture-Recapture." *Statistics in Medicine* 35: 622–634. Doi: http://dx.doi.org/10.1002/sim.6741.

Pollock, K.H. 2002. "The Use of Auxiliary Variables in Capture-Recapture Modelling: an Overview." *Journal of Applied Statistics* 29: 85–102. Doi: http://dx.doi.org/10.1080/02664760120108430.

Reurings, M.C.B. and N.M. Bos. 2012. "Ernstig Verkeersgewonden in de Periode 2009 en 2010. Update van de Cijfers." Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV, ref.; R-2012-7, Leidschendam, 2012. Available at: https://www.narcis.nl/publication/RecordID/oai:library.swov.nl:129380.

Reurings, M.C. and H.L. Stipdonk. 2011. "Estimating the Number of Serious Road Injuries in the Netherlands." *Annals of Epidemiology* 21(9): 648–653. Doi: http://dx.doi.org/10.1016/j.annepidem.2011.05.007.

Schafer, J. 1997a. *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall/CRC.

Schafer, J. 1997b. *Imputation of Missing Covariates under a General Linear Mixed Model*. PennState University, Department of Statistics.

Sutherland, J.M., C.J. Schwarz, and L.-P. Rivest. 2007. "Multilist Population Estimation with Incomplete and Partial stratification." *Biometrics* 63: 910–916. Doi: http://dx.doi.org/10.1111/j.1541-0420.2007.00767.x.

Tilling, K. and J.A.C. Sterne. 1999. "Capture-Recapture Models Including Covariate Effects." *American Journal of Epidemiology* 149(4): 392–400. Doi: http://dx.doi.org/10.1093/oxfordjournals.aje.a009825.

Van der Heijden, P.G.M., E. Zwane, and D. Hessen. 2009. "Structurally Missing Data Problems in Multiple List Capture-Recapture Data." *Advances in Statistical Analysis* 93: 5–21. Doi: http://dx.doi.org/10.1007/s10182-008-0098-6.

Van der Heijden, P.G.M., J. Whittaker, M. Cruyff, B. Bakker, and R. van der Vliet. 2012. "People Born in the Middle East but Residing in the Netherlands: Invariant Population Size Estimates and the Role of Active and Passive Covariates." *The Annals of Applied Statistics* 6(3): 831–852. Doi: http://dx.doi.org/10.1214/12-AOAS536.

Whittaker, J. 1990. *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Wolter, K.M. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81(394): 337–346. Doi: http://dx.doi.org/10.1080/01621459.1986.10478277.

Zhang, L.-C. 2015. "On Modelling Register Coverage Errors." *Journal of Official Statistics* 31: 381–396. Doi: http://dx.doi.org/10.1515/jos-2015-0023.

Zwane, E. and P.G.M. van der Heijden. 2007. "Analysing Capture-Recapture Data when some Variables of Heterogeneous Catchability are not Collected or Asked in All Registrations." *Statistics in Medicine* 26: 1069–1089. Doi: http://dx.doi.org/10.1002/sim.2577.

Zwane, E. and P.G.M. van der Heijden. 2008. "Capture-Recapture Studies with Incomplete Mixed Categorical and Continuous Covariates." *Journal of Data Science* 6: 557–572.

Zwane, E., K. van der Pal-de Bruin, and P.G.M. van der Heijden. 2004. "The Multiple-Record Systems Estimator when Registrations Refer to Different but Overlapping Populations." *Statistics in Medicine* 23: 2267–2281. Doi: http://dx.doi.org/10.1002/sim.1818.

Zwane, E. and P. van der Heijden. 2005. "Population Estimation Using the Multiple System Estimator in the Presence of Continuous Covariates." *Statistical Modelling* 5(1): 39–52. Doi: http://dx.doi.org/10.1191/1471082X05st086oa.

Received April 2016
Revised September 2017
Accepted September 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 265–301, http://dx.doi.org/10.1515/JOS-2018-0012

Factor Structural Time Series Models for Official Statistics with an Application to Hours Worked in Germany

Roland Weigand[1], *Susanne Wanger*[2], *and Ines Zapf*[3]

We introduce a high-dimensional structural time series model, where co-movement between the components is due to common factors. A two-step estimation strategy is presented, which is based on principal components in differences in a first step and state space methods in a second step. The methods add to the toolbox of official statisticians, constructing timely regular statistics from different data sources. In this context, we discuss typical measurement features such as survey errors, statistical breaks, different sampling frequencies and irregular observation patterns, and describe their statistical treatment. The methods are applied to the estimation of paid and unpaid overtime work as well as flows on working-time accounts in Germany, which enter the statistics on hours worked in the national accounts.

Key words: Unobserved components model; state space model; national accounts; overtime work; working-time accounts.

1. Introduction

In a very important and publicly visible field of official statistics, early releases of economic production or labor market indicators are constructed on a regular monthly or quarterly basis. Several surveys and other data sources are typically used to update these time series based on the information available so far. The importance of timely and precise measures of the economy is emphasized in a large literature on real-time data analysis, which shows that data revisions pose a severe challenge to forecasters and policymakers; see, for example, Croushore (2011). Hence, on the side of statistical agencies, most prominently for quarterly national accounts, efforts are made to produce accurate statistics by bringing together a large amount of primary data sources, typically surveys; see Bureau of Economic Analysis (2017), Wood and Elliott (2007), and Federal Statistical Office (2008) for GDP calculation in the US, in the UK, and in Germany.

The current article is a methodological contribution to this field of activity. For the estimation of a target series θ_t such as real GDP or hours worked in the past quarter, we make use not only of currently available surveys z_t that aim to measure θ_t, but notably also of the history of such surveys, $z_{t-1}, z_{t-2}, \ldots$, and of a possibly very large set of additional indicators, $x_t, x_{t-1}, \ldots$, which are in some way related to θ_t. Hence, in the terminology of survey or small area statistics, we discuss a new model and an

[1,2,3] Institute for Employment Research (IAB) Regensburger Strasse 104, 90478 Nuremberg, Germany. Emails: roland.weigand@posteo.de, susanne.wanger@iab.de, and ines.zapf@iab.de
Acknowledgments: The authors are grateful for helpful comments by Enzo Weber, by participants of the IAB-Bundesbank seminar 2013 in Frankfurt am Main and of the German Statistical Week 2014 in Hannover.

estimator that borrows strength both over time and from related variables in a data-rich environment.

Our proposed approach is based on factor models for high-dimensional time series which have become an indispensable tool for macroeconomic fore- and nowcasting as well as structural modeling; see Bai and Ng (2008) as well as Stock and Watson (2011) for recent surveys. In this field, seasonally adjusted variables typically enter the model in first or second differences, while the factors are modeled as a stationary VAR process. Methods for handling nonstationary variables are also available (Bai and Ng 2004), and unit-root versions of the factor-augmented VAR as well as error correction models are an area of active research; see, among others, Banerjee et al. (2014).

We propose a concurrent parametrization for large factor models of nonstationary variables which we formulate in the structural time series framework of Harvey (1991). Factor structures on the trend, seasonal, cyclical and irregular components allow to model the co-movements of a large number of time series in a parsimonious, componentwise manner. The popular common trends or common cycle models emerge as special cases, but a common features assumption, restricting the idiosyncratic part to be stationary or even serially uncorrelated, is not necessarily imposed in our framework. Rather, the idiosyncratic part may be characterized by trend, cycle and seasonal components as well.

For a straightforward and computationally feasible implementation of the approach, a principal component analysis is combined with state space methods in the spirit of Bräuning and Koopman (2014). We extract the principal components of suitably differenced data to account for nonstationarity of the idiosyncratic part. Re-cumulated factors are modeled jointly with the series of primary interest using likelihood-based techniques within a state space framework. In Monte Carlo simulations, we find that this method performs well, irrespectively of whether a common features assumption holds.

From the perspective of data construction, we discuss several advantages and possible modifications of our model in state space form. Primary sources in official statistics are typically subject to survey errors and statistical breaks. They may be collected at different sampling frequencies, while changing survey designs lead to irregular measurement patterns. Since the key part of our model is formulated in state space form, it is well-suited to handle these patterns. It produces efficient estimates of the target series when different surveys measure the same underlying series. Information from the past of the series is processed, and additional strength is borrowed from a large number of related series with correlated components. Seasonally adjusted time series, using all available data for the adjustment, are obtained as a by-product of the procedure.

The potential of the state space approach for official statistics has already been pointed out by other researchers. Uses in several areas of official statistics have been highlighted by Durbin (2000). There are examples where state space methods are applied for seasonal adjustment, while Pfeffermann (1991) and Tiller (1992) discuss signal extraction from repeated survey data. In small area statistics, state space models help obtain disaggregate figures from surveys by borrowing strength both over time and space, see Pfeffermann and Tiller (2006) and Krieg and van den Brakel (2012). In that context, Bollineni-Balabay et al. (2015) pursue the estimation of aggregates along with the small-area domains in the presence of survey redesigns and variance breaks. Durbin and Quenneville (1997) and Quenneville and Gagné (2013) introduce benchmark constraints drawn from precise but

low-frequency census data to correct the preliminary survey estimates, while Harvey and Chung (2000) discuss modeling data from different sources, and Moauro and Savio (2005) is concerned with temporal disaggregation as required by national statistical agencies.

We apply our methodology to the statistics of hours worked in Germany. High-quality data on hours worked are a key for understanding aggregate labor market dynamics, for example, to track business cycles, to assess reactions to shocks such as the 2008/09 financial and economic crisis (Burda and Hunt 2011), and to confront macroeconomic theory with time series evidence (Ohanian and Raffo 2012). Timely figures on hourly labour productivity are considered as being important, for example, for well-guided wage negotiations and monetary policy.

In Germany, working time statistics are constructed within the working time measurement concept of the Institute for Employment Research (IAB). The componentwise accounts provide a comprehensive figure of hours worked and contributes results to the German national accounts; see Wanger et al. (2016). In the measurement of overtime hours and flows on working-time accounts (WTA), we use household and business surveys, while additionally drawing on several labor market and business cycle indicators. Lacking continuously available survey data on working-time account net flows, the latter is based on the unobserved trend and cycle components for transitory overtime hours as well as regular and actual hours worked.

The article is structured as follows: Section 2 describes the model and its statistical treatment, Section 3 illustrates alternative measurement schemes faced in official statistics and Section 4 presents finite sample properties of the procedure. Section 5 applies the methods to the German statistics of hours worked, while the last section concludes.

2. A High-Dimensional Structural Time Series Model

2.1. The Factor Model

This article presents a new model and its implementation for official statistics. It extends the scope of multivariate structural time series models (STSM) discussed by Harvey and Koopman (1997) to high-dimensional applications. As the point of departure, an N-dimensional vector time series y_t is decomposed into trend μ_t, seasonal γ_t, cycle c_t, and irregular components u_t, according to

$$y_t = \mu_t + \gamma_t + c_t + u_t, \tag{1}$$

where the terms on the right are unobserved stochastic processes. Additional components such as calendar effects or outliers can be straightforwardly incorporated through the use of dummy regressors given this additive formulation but are not considered in this article. After describing the dynamic specification of the components we introduce a factor structure that handles cross-series linkages within the groups of components, and the statistical treatment of the model.

We use a standard specification for the dynamics of each component and characterize the slow movements by local linear trends

$$\mu_{t+1} = \mu_t + \nu_t + \xi_t, \quad \nu_{t+1} = \nu_t + \zeta_t,$$

where $\xi_t \sim \text{iid } N(0, \Sigma_\xi)$ and $\zeta_t \sim \text{iid } N(0, \Sigma_\zeta)$ are independent Gaussian white noise sequences. For a model frequency of s observations per year, the seasonal components are

$$\gamma_{t+1} = -\sum_{j=0}^{s-2} \gamma_{t-j} + \omega_t, \quad \omega_t \sim \text{iid } N(0, \Sigma_\omega).$$

Alternatively, a trigonometric specification for the seasonal components can be used (Durbin and Koopman 2012, Sec. 3.2.1). An individual cycle component $\tilde{c}_{it}$, $i = 1, \ldots, N$ evolves jointly with the auxiliary process $\tilde{c}^*_{it}$ as

$$\begin{pmatrix} \tilde{c}_{i,t+1} \\ \tilde{c}^*_{i,t+1} \end{pmatrix} = \rho_i \begin{pmatrix} \cos\lambda_i & \sin\lambda_i \\ -\sin\lambda_i & \cos\lambda_i \end{pmatrix} \begin{pmatrix} \tilde{c}_{it} \\ \tilde{c}^*_{it} \end{pmatrix} + \begin{pmatrix} \kappa_{it} \\ \kappa^*_{it} \end{pmatrix}, \quad \begin{pmatrix} \kappa_{it} \\ \kappa^*_{it} \end{pmatrix} \sim \text{iid } N(0, \Sigma_{\kappa,ii} I),$$

where λ_i is the dominant frequency and $0 < \rho_i < 1$ denotes the dampening factor. As for the trends and seasonal components, linkages between the individual cycles are introduced through covariances between the disturbances, collected in Σ_κ. To gain flexibility on the temporal timing of the co-movement, we introduce phase shifts $\delta_2, \ldots, \delta_N$ between the cycles by setting $c_{it} = \tilde{c}_{it}\cos\lambda_i\delta_i + \tilde{c}^*_{it}\sin\lambda_i\delta_i$, $i = 1, \ldots, N$, where $\delta_1 = 0$ as a normalization and δ_i measures the lead time of cycle j against the cycle of the first variable; see Rünstler (2004) and Valle e Azevedo et al. (2006). Finally, the irregular noise term is given by $u_t \sim \text{iid } N(0, \Sigma_u)$. For simplicity we assume that all groups of shocks ξ_t, ζ_t, ω_t κ_t and u_t are mutually independent. Correlated components in the spirit of Morley et al. (2003) could be straightforwardly adapted as long as suitable identification conditions are met.

Our focus is on cases where N, the number of series in y_t is large, and hence a curse of dimensionality occurs in the unrestricted model (1). For full covariance matrices Σ_ξ, Σ_ζ, Σ_ω, Σ_κ and Σ_u, there are $O(N(N+1))$ variance parameters to be estimated, which makes the application practically infeasible even for moderate values of N. In such situations, factor models have been found useful for different purposes in economics and finance. They allow a parsimonious representation of the cross-section dependencies between panel units or time series variables. Within our STSM setup, we consider common factors for each group of components. Denoting the common components by a C superscript and the idiosyncratic terms by I, our model is given by

$$y_t = \Lambda_\mu \mu^C_t + \Lambda_\gamma \gamma^C_t + \Lambda_c c^C_t + \Lambda_u u^C_t + \mu^I_t + \gamma^I_t + c^I_t + u^I_t. \tag{2}$$

The common components are of dimensions r_μ, r_γ, r_c and r_u, respectively, which are typically substantially smaller than N, while Λ_k, $k \in \{\mu, \gamma, c, u\}$ are $N \times r_k$ loading matrices of full column rank. All components follow the same dynamics as those described below (1), and are driven by shocks with covariance matrices Σ^C_l and Σ^I_l for $l \in \{\xi, \zeta, \omega, \kappa, u\}$. The idiosyncratic components are assumed mutually uncorrelated and hence Σ^I_l are diagonal, so that the number of parameters is reduced to an order $O(N)$ for fixed factor dimensions.

Our decision of using a factor model to circumvent the curse of dimensionality is popular in the econometrics field, since this "reduced rank sparsity" can easily handle high correlations between the series due to business cycle linkages. Especially cyclical

movements are candidates for such a rank reduction, but also long-term trends may be linked by identical underlying driving forces. The alternative way to avoid the dimensionality problem, the so-called "zero sparsity" where correlations are set to zero, is often less plausible in such setups. In the empirical application, we practice a mix of reduced rank and zero sparsity: Component variances and correlations are set to zero when this is statistically appropriate after the model dimension has been drastically reduced by the factor approach.

Identification of the latent components in (2) is achieved if two conditions hold: (a) The processes μ_t, γ_t, c_t and u_t from (1) are separately identified through their difference in dynamics, and (b) for each of these dynamic components, for example for $\Lambda_\mu \mu_t^C + \mu_t^I$, the common $\Lambda_\mu \mu_t^C$ and idiosyncratic μ_t^I are distinguished as in classical factor models by the assumption that the idiosyncratic series are mutually uncorrelated in the cross-section dimension. Condition (a) is standard both in univariate and multivariate structural time series models and unproblematic in the uncorrelated components case considered here. It is discussed among others by Harvey (1991, sc. 4.4). Condition (b) is not related to the dynamic properties of the series, but only draws on the correlations between the series which are due to a low-dimensional factor process. The autocorrelation and even nonstationarity, for example of $\Lambda_\mu \mu_t^C + \mu_t^I$, does not interfere with the identification problem since the setup can be easily transformed to the classical "white noise" factor model of Anderson (1984) by a univariate time series filter; the reversed filter applied to the identified components would in principle recover the original autocorrelation structure. Clearly, as in classical factor models, the loadings and factors are identified only up to rotation, so that the additional normalizations that the upper $r_k \times r_k$ block of suitable loading matrices are identity will be used in Subsection 2.2. The chosen identification, however, does not matter for the purpose of this article which is estimation of a target series rather than structural inference on the factors.

The factor STSM can be represented in the notation of a standard multivariate STSM (1) if a similar cycle assumption holds, that is, if all ρ_i and λ_i are identical for both the common and idiosyncratic components. However, the factor structure imposes restrictions on the disturbance covariance matrices, which are given by

$$\Sigma_\xi = \Lambda_\mu \Sigma_\xi^C \Lambda_\mu' + \Sigma_\xi^I, \quad \Sigma_\zeta = \Lambda_\mu \Sigma_\zeta^C \Lambda_\mu' + \Sigma_\zeta^I, \quad \Sigma_\omega = \Lambda_\gamma \Sigma_\omega^C \Lambda_\gamma' + \Sigma_\omega^I,$$
$$\Sigma_\kappa = \Lambda_c \Sigma_\kappa^C \Lambda_c' + \Sigma_\kappa^I, \quad \Sigma_u = \Lambda_u \Sigma_u^C \Lambda_u' + \Sigma_u^I.$$

If one or more of the columns of Λ_i are linearly dependent with those of Λ_j, $i \neq j$, the stacked loadings (Λ_μ, Λ_γ, Λ_c, Λ_u) have a reduced column rank denoted by $r < r_\mu + r_\gamma + r_c + r_u$. Then, the cross-section correlations between variables in y_t can be traced back to a smaller number of common sources than there are common structural time series components. This possibly smaller dimensional latent process is given by the r-dimensional compound factors denoted by f_t with a corresponding full-rank $N \times r$ loading matrix Λ, such that $y_t = \Lambda f_t + \mu_t^I + \gamma_t^I + c_t^I + u_t^I$. The compound factors are related to the common components by

$$f_t = \Gamma_\mu \mu_t^C + \Gamma_\gamma \gamma_t^C + \Gamma_c c_t^C + \Gamma_u u_t^C,$$

where $\Gamma_k = (\Lambda'\Lambda)^{-1}\Lambda'\Lambda_k$ are $r \times r_k$ matrices of full column rank. Again, factors f_t, loadings Λ and Γ_k are only identified up to linear combinations, but for a chosen rotation the common components μ_t^C, γ_t^C, c_t^C and u_t^C are identified (up to rotation) through their different dynamics, and hence can be estimated from f_t by the state space approach described by Harvey (1991). As an example with the richest dynamic structure possible for a given r, consider the case with $r = r_\mu = r_\gamma = r_c = r_u$ and $\Lambda = \Lambda_\mu = \Lambda_\gamma = \Lambda_c = \Lambda_u$. The factors $f_t = \mu_t^C + \gamma_t^C + c_t^C + u_t^C$ then follow a structural time series process and consist of trend, irregular, seasonal and cyclical components themselves.

Factor structures in the multivariate STSM framework have been studied before in the econometrics literature, albeit with a different scope. Models for a moderate number of series have been used to investigate common trends (and thus cointegration) or common cycles in their dynamics; see, for example, Harvey (1991, Sec. 8.5), Valle e Azevedo et al. (2006) or the software implementation of Koopman et al. (2009). In the standard setup, the idiosyncratic part is a white noise process, or at least has different dynamic properties from the factors'. Identification of the factor (e.g., μ_t^C in the common level model $y_t = \Lambda_\mu \mu_t^C + \varepsilon_t^I$) is therefore achieved in both of the ways (a) and (b) discussed above at the same time: In the common levels model μ_t^C is the only source of autocorrelation and also the only source of correlation between the series. This restricts the model in a very relevant way and makes it less applicable for high-dimensional settings, since especially in high dimensionsid iosyncratic errors with restricted dynamic properties (or even white noise) are unrealistic and a high factor dimension would be needed to provide a reasonable approximation to the data. Eickmeier (2009), among others, finds unit roots in the idiosyncratic part of many macroeconomic time series, so that a common trends assumption fails for a reasonable factor dimension.

Our Model (2), in contrast, allows factors and idiosyncratic part to have the same types of components as the common part, and hence to consist of trend, seasonal, cycle and noise. In this way, we may obtain a more parsimonious structure with less factors when a larger panel of data is considered. Our model is rather general in that it allows for co-movements in each of the components, while a common features restriction is possible by setting the respective idiosyncratic components, say trends or cycles, to zero. As we describe in the next section, our model allows a computationally feasible treatment even in the high-dimensional case, since it naturally allows a combination of PCA and state space methods. In contrast, in common cycles or common trends models the components are typically filtered out from a full state space approach which becomes cumbersome for a larger number of series and factors.

In the high-dimensional factor framework, by far the most popular approach for dynamic modeling is by estimating factors by principal components, and using VAR models for observed series and estimated factors, resulting in VAR-based dynamic factor or so-called factor-augmented VAR (FAVAR) models; see, for example Stock and Watson (2005) or Bernanke et al. (2005). Model (2) has several benefits also relative to such VAR-based approaches. Firstly, the structural approach offers insights into the nature of co-movements between the series, which can be assigned to specific components: Is it because of business cycles or rather correlated trends that macroeconomic time series co-move? Are there joint sources of changing seasonal patterns in several branches of the economy? Can common irregular components like weather effects be identified that

transitorily hit several output measures? Secondly, in the context of filtering a signal from sparsely available data, the structural time series setup imposes a parsimonious parametrization which stabilizes the estimates. In the application to official statistics, all components help estimate the different features of the target series while taking into account all relevant information from related series. Thirdly, using information from many series may also lead to important improvements of seasonal adjustment procedures over univariate approaches.

2.2. *Estimation by Collapsing the Factor Space*

We suggest an estimation procedure of the model given by a combination of principal component and state space techniques along the lines of Bräuning and Koopman (2014). Assume that we are primarily interested in a low-dimensional subprocess z_t holding N_z series of the available data, while the complete set of time series is separated according to $y_t = (x_t', z_t')'$. In forecasting applications, z_t will hold at least the series to be predicted, while the estimation of official statistical figures typically requires the series z_t to consist of the major surveys measuring the target series. Unlike Bräuning and Koopman (2014), we assume that all variables in y_t are generated by the same model, (2) in our case, and hence variables in x_t and z_t are treated symmetrically in terms of the model but not in terms of the estimation procedure.

To estimate the space of compound factors f_t in a first step, we apply a suitable principal components analysis to x_t. By using the data x_t in differences, we avoid possible inconsistencies due to nonstationary idiosyncratic components, and thus adapt ideas of Bai and Ng (2004) to our setting. More concretely, denoting by L the lag operator, by $\Delta := (1 - L)$ the standard difference and by $\Delta_s := (1 - L_s)$ the seasonal difference operator, we obtain factor loadings $\bar{\Lambda}$ as $\sqrt{N_x}$ times the orthonormal eigenvectors corresponding to the r largest eigenvalues of $\sum_{t=1}^{T} (\Delta\Delta_s x_t)(\Delta\Delta_s x_t)'$. Estimated factors are obtained by re-cumulating the principal components in differences, or from the level data as $\bar{f}_t = \bar{\Lambda}' x_t$, which differs from the re-cumulation approach through the effects of initial values. Under an additional assumption on the factor loadings, the results of Bai and Ng (2002) are applicable to the variables in differences; see Appendix A. Among other things, this assures consistency (up to rotation and net of the effects of starting values) of $\bar{f}_t$ for f_t at a fixed t as N and T tend to infinity. In the setup (2), the differenced series are usually autocorrelated as are the residuals from the principal components approach. However, as long as differences of sufficient orders are taken, the autocorrelation is weak in the sense of Bai and Ng (2002, Assumption C) and consistency of the factors in this *approximate* factor framework is ensured.

The principal components approach is typically not optimal and comes with an efficiency loss, for example, in the situation of outliers due to nongaussianity, of heteroskedasticity of the idiosyncratic components, or of autocorrelation. We propagate its use as asimple, well-understood and popular first-step estimator, but of course improved versions are available and can also be applied in our setup (see, e.g., Breitung and Tenhofen 2011).

To gain information on the common components and their relation to the variables in z_t, we consider the joint model of f_t and z_t within the state space setup. Replacing the

compound factors f_t by their estimates, the model is given by

$$\begin{pmatrix} \bar{f}_t \\ z_t \end{pmatrix} = \begin{pmatrix} \Gamma_\mu \\ \Lambda_\mu \end{pmatrix} \mu_t^C + \begin{pmatrix} \Gamma_\gamma \\ \Lambda_\gamma \end{pmatrix} \gamma_t^C + \begin{pmatrix} \Gamma_c \\ \Lambda_c \end{pmatrix} c_t^C + \begin{pmatrix} \Gamma_u \\ \Lambda_u \end{pmatrix} u_t^C + \begin{pmatrix} e_t \\ \mu_t^I + \gamma_t^I + c_t^I + u_t^I \end{pmatrix}, \quad (3)$$

where e_t is the error of $\bar{f}_t$ estimating f_t. As a slight abuse of notation, the idiosyncratic components and loadings are those corresponding to the elements in z_t only. While the compound factor estimates $\bar{f}_t$ are identified by the standard normalization of principal components, the common structural time series components are made unique by setting

$$\Gamma_\mu = \begin{pmatrix} I_{r_\mu} \\ \Gamma_\mu^{(2)} \end{pmatrix}, \quad \Gamma_\gamma = \begin{pmatrix} I_{r_\gamma} \\ \Gamma_\gamma^{(2)} \end{pmatrix}, \quad \Gamma_c = \begin{pmatrix} I_{r_c} \\ \Gamma_c^{(2)} \end{pmatrix}, \quad \Gamma_u = \begin{pmatrix} I_{r_u} \\ \Gamma_u^{(2)} \end{pmatrix},$$

and the common components may have unrestricted disturbance covariance matrices. Setting the upper block of the loading matrices to identity is only one of many ways to prohibit observationally equivalent rotations of factors and loadings (see e.g., Bai and Ng 2013), but especially fore- and nowcasts of the series do not depend on such normalizations.

Under the given restrictions, the model can be operationalized by ignoring the error from principal components estimation, and hence setting $e_t = 0$, which is justified as an approximation especially for large N. The unknown hyperparameters of (3) are estimated by maximum likelihood using the state space approach.

Alternatively, a multivariate STSM without the restrictions of (3) can be fitted to the joint process of principal components and variables of interest. This second strategy allows for correlation between the idiosyncratic terms of z_t, while the model nests the factor STSM specification (3). Given typical factor dimensions of less than five and a univariate or low-dimensional z_t, especially the latter estimation approach can easily be conducted in one of several available software packages such as STAMP (Koopman et al. 2009) or those described by Commandeur et al. (2011) and articles in the same special issue. For the computations in this article, the KFAS package for R is used (Helske 2016).

Empirically, the compound factor dimension can be inferred from the data y_t in suitable differences, for example by the methods proposed by Bai and Ng (2002). Alternatively, different (small) values of r can be considered and robustness with respect to this choice can be assessed in practice. Subsequently, for a given r, beginning from $r_\mu = r_\gamma = r_c = r_u = r$, the dimension of each common component may be determined in a general-to-specific sequential testing procedure based on (3).

3. Observation Schemes

The factor STSM introduced in this article has advantages in filtering latent series from incomplete measures which is a key issue in official statistics. For this purpose, we assume that a latent N_θ dimensional process θ_t of target series instead of observed z_t is modeled to follow the factor STSM (2), and that the observations collected in z_t are related to θ_t through a dynamic measurement relationship

$$z_t = d_t + M_t(L)\theta_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H_t), \quad (4)$$

where d_t holds possible survey bias terms and statistical breaks, $M_t(L) = M_{t0} + M_{t1}L + \ldots + M_{tl}L^l$ are $N_z \times N_\theta$ matrix lag polynomials holding the dynamic measurement coefficients, while ε_t is a vector of survey errors with possibly time-varying covariance matrices H_t. The latter need not necessarily follow a white noise process, but can, for example, contain autocorrelation due to survey overlap, which may be treated by the methods of Pfeffermann and Tiller (2006). We review some of the cases that the general mechanism (4) captures, and propose its implementation in the state space form which is given in Appendix B.

The measurement scheme (4) is sufficiently flexible to allow for several surveys estimating the same underlying concept, for missing data and for time-varying observation patterns. Consider an example where θ_{1t}, for example, paid overtime hours per week and employee, is measured by two surveys z_{1t} and z_{2t}, for example, the German Socio-Economic Panel (GSOEP), and the German Microcensus, as it is the case in the application to German hours worked data below. The measurement mechanism is then given by

$$z_{1t} = \theta_{1t} + \varepsilon_{1t}, \quad z_{2t} = d_2 + \theta_{1t} + \varepsilon_{2t}. \tag{5}$$

In this simple example, with $M_t(L) = (1, 1)'$, the scheme brings together contradicting surveys, where differences are explained by the survey errors ε_{1t} and ε_{2t}. The variances of these errors depend on the design and size of the survey and are likely to change over time. By including an unknown constant d_2 in the second measurement equations, it is possible to correct for a bias in one of the sources. Similarly, if statistical breaks, like changes in the survey questionnaire, occur in one or more of the data sources, these may be explicitly accounted for by level shifts in d_t, and hence leave the measured θ_t unaffected. In case of changing seasonal patterns or covariance structures of the components, however, a time-varying transition rather than measurement equation has to take this into account, a topic that we do not consider in this article.

Different sampling frequencies of regular surveys, or data missing for other reasons, are also covered by the measurement scheme (4), which is an important topic in the existing nowcasting literature (Giannone et al. 2008). Considering a quarterly stock variable which is measured only at the end of the quarter, we observe the monthly value $z_{1t} = \theta_{1t}$ only when t is the last month of a quarter, while values of z_t are missing two thirds of the time. Returning to the bivariate example, if in period t no survey z_{1t} is conducted, we obtain a trivial equation

$$0 = 0 \cdot \theta_{1t} + 0, \quad z_{2t} = d_2 + \theta_{1t} + \varepsilon_{2t} \tag{6}$$

by specifying $M_t(L) = (0, 1)'$ and $H_{11,t} := \mathrm{Var}(\varepsilon_{1t}) = 0$. Hence, no information is gained by the first survey in that period. Therefore, information about θ_{1t} stem firstly from other surveys z_{2t}, secondly from past and future values of z_{1t} through the dynamics of the system, or thirdly from additional indicators correlated with θ_{1t} through the common components. If one survey z_{1t} is used as a benchmark and hence the resulting estimate of θ_{1t} should exactly match that survey, this is reached by setting $\varepsilon_{1t} = 0$. Further relevant methods for benchmarking are discussed in Durbin and Quenneville (1997) and Quenneville and Gagné (2013).

In contrast to the previous example of a stock variable where survey interviews reflect observations on one period t, in reality reference intervals may span more than one period in terms of the model frequency. This is typically the case for flow variables like GDP where we observe the sum of the monthly flows $z_{1t} = \theta_t + \theta_{t-1} + \theta_{t-2}$ at the end of the quarter. With a lag polynomial $M(L) = 1 + L + L^2$, we can also write $z_{1t} = M(L)\theta_t$. As another example, since 2005 the German Microcensus has a continuous interview policy and allows an evaluation of quarterly averages of quantities such as overtime hours worked per week. If the model is formulated at monthly frequency, an observation z_{1t} of a flow variable, corresponding to the second quarter 2006, refers to the mean of the underlying θ_{1t}, $\theta_{1,t-1}$ and $\theta_{1,t-2}$ of April, May and June. The measurement equation reflects this by assigning the quarterly value z_{1t} to the last month of the quarter and selecting

$$z_{1t} = \frac{1}{3}\theta_{1t} + \frac{1}{3}\theta_{1,t-1} + \frac{1}{3}\theta_{1,t-2} = M_t(L)\theta_{1t} \tag{7}$$

where z_{1t} contains values only at the end of the quarter of each year, and where $M_t(L) = \frac{1}{3} + \frac{1}{3}L + \frac{1}{3}L^2$ is the lag polynomial that reflects that measurement scheme. The change from a fixed reference week to continuous interviews is reflected by a change in the time-varying observation polynomial $M_t(L)$, so that $M_t(L) = 1$ for periods t before 2005.

For other surveys, the observation scheme is still more general. For example, for household panel studies such as the GSOEP or the U.S. panel study of income dynamics (PSID), the field period spans several months and changes from year to year. Assigning the resulting yearly figure to the December of each survey year, an observation equation

$$z_{1t} = M_{t,dec}\theta_{1t} + \ldots + M_{t,jan}\theta_{1,t-11} \tag{8}$$

reflects the time-varying shares M_{tj} of observations in each month j, relative to all observations in that year. Figure 1 shows the distribution of the GSOEP interviews for

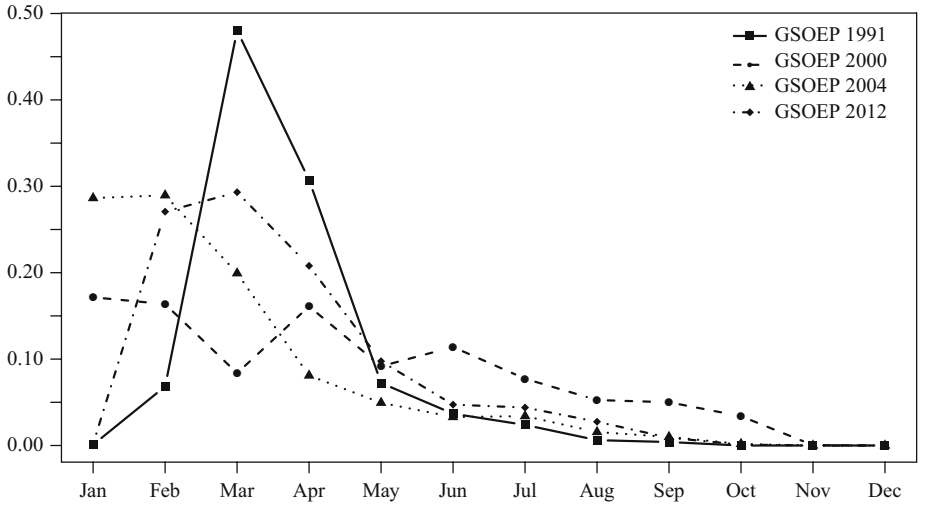


Fig. 1. Distribution of GSOEP interviews over certain years. The fraction of interviews for each month is shown for 1991, 2000, 2004, and 2012.

selected years, namely for 1991 (solid line), for 2000 (dashed), for 2004 (dotted), and for 2012 (dash-dotted).

Note that the measurement scheme might interfere with the identification of a given model. If an underlying series θ_{1t} is measured by a low-frequency, say quarterly, series alone, then seasonal components of some frequency may be unobservable; see Harvey (1991, Sec. 6.3). One obvious way to circumvent this is to use quarterly time-varying dummy variables that do not aim to estimate intra-quarter seasonality, while one could alternatively use a trigonometric seasonal specification (Durbin and Koopman 2012, Sec. 3.2.2) and skip frequencies higher than the observation frequency.

The Model (3) with the measurement scheme (4) can be stated in state space form (Appendix B). After estimating the model hyperparameters by the methods described in Subsection 2.2, estimated θ_t, $t = 1, \ldots, T$ using all available data are obtained by a state smoothing algorithm (Durbin and Koopman 2012, Sec. 4.3). The application of a smoother rather than a Kalman filter means that also past data are revised as soon as new information comes in. The smoother is constructed in a way that the revision optimally reflects the new information, given the model structure and its parameters. Since it is current practice in national accounts to revise also recent quarters, the use of a smoother automatically implements this revision together with the computation of a new quarter. Hence, no additional updating mechanism or model using more data is needed.

4. A Monte Carlo Study

A Monte Carlo study is conducted to shed light on the practical performance of the proposed methods in finite samples. Different aspects of the procedure are analysed. Firstly, the difference-based principal components approach is studied in the case of factor STSM processes for different data generating mechanisms and sample sizes, and compared to principal components in levels. Secondly, the estimation of the latent target process is evaluated and compared to standard benchmarks such as univariate models or standard principal-components based factor models.

4.1. Data Generating Processes

Four data generating processes are chosen to mimic different situations of practical relevance. We consider (1) cases with linearly independent loadings for the distinct common components and (2) cases with identical loadings, where principal components estimate a compound factor process. Furthermore, while typically (A) the idiosyncratic components have a structural time series structure with trend, seasonal and possibly cycle components, we additionally consider a common features assumption with (B) serially uncorrelated idiosyncratic components. We introduce the data generating processes as combinations of these characteristics in turn.

(1A) To define the first data generation mechanism as the case with linearly independent loadings and without common features, we consider the process (2) with $s = 4$ and where the cyclical components have frequency $\lambda = 0.2$ and dampening factor $\rho = 0.97$. The parameters in the loading matrices are randomly chosen for each draw. Denoting the uniform distribution between a and b by $U(a, b)$, they are given for $i = 1, \ldots, N$ and

$j = 1, \ldots, r$ by

$$\Lambda_{\mu,ij} \sim U(0, \chi), \quad \Lambda_{\gamma,ij} \sim U(0, \chi), \quad \Lambda_{c,ij} \sim U(0, \chi\varsigma), \quad \Lambda_{u,ij} \sim U(0, \chi\varsigma),$$

where the parameter $\mathcal{X}$ captures the overall importance of the common components relative to the idiosyncratic ones, while ς determines the relative size of stationary versus nonstationary components. The components are generated using innovation covariance matrices with

$$\Sigma^I_{\xi,ii} \sim U(0,1)^2, \quad \Sigma^I_{\zeta,ii} \sim U(0,1/10)^2, \quad \Sigma^I_{\omega,ii} \sim U(0,1)^2,$$

$$\Sigma^I_{\kappa,ii} \sim U(0,\varsigma)^2, \quad \Sigma^I_{u,ii} \sim U(0,\varsigma)^2$$

for idiosyncratic components and $\Sigma^C_\xi = 10\Sigma^C_\zeta = \Sigma^C_\omega = \Sigma^C_\kappa = \Sigma^C_u = I$ for common components, respectively.

(1B) The second data generating process is characterized by the same parameters for the common components as in (1A), but a common features assumption is imposed and hence the idiosyncratic components are subject to

$$\Sigma^I_\xi = \Sigma^I_\zeta = \Sigma^I_\omega = \Sigma^I_\kappa = 0, \quad \Sigma^I_{u,ii} \sim U(0,5)^2.$$

(2A) To introduce cases with linearly dependent common components loadings, the third data generating process sets the compound loading matrix Λ according to

$$\Lambda_{\mu,ij} = \Lambda_{\gamma,ij} = \frac{1}{\varsigma}\Lambda_{u,ij} \sim U(0, \chi),$$

and drops the cyclical components from the processes. The remaining variances Σ^I_ξ, Σ^I_ζ, Σ^I_ω, Σ^I_u, Σ^C_ξ, Σ^C_ζ, Σ^C_ω, and Σ^C_u correspond to those in (1A).

(2B) The last data generating process drops the trend and seasonal from the idiosyncratic components of the previous one, so that the only difference to (2A) is in the covariance matrices

$$\Sigma^I_\xi = \Sigma^I_\zeta = \Sigma^I_\omega = 0, \quad \Sigma^I_{u,ii} \sim U(0,5)^2.$$

4.2. Estimation of the Compound Factor Space

We first assess the performance of the principal components procedure based on differenced data $\Delta\Delta_4 y_t$ which we have proposed as a first step in estimating the factor STSM. For all data generating processes and different values for the time and cross-section dimensions T and N, we simulate 1,000 trajectories and repeatedly estimate the compound factor process f_t by $\bar{f}_t$ as explained in Subsection 2.2. We compare the results to the principal component method using the data in levels. At this step, no maximum likelihood estimation of the structural model is performed and hence the dynamic properties are not taken into account. Therefore, only the space of compound factors can be estimated, which is identified only up to rotation. The measure of estimation error has to take this lack of identification into account, and hence we rotate each factor estimate to achieve the best predictive power for the true factors by least squares. Since the overall level of mean

squared error is very different across estimated components and has no natural interpretation, we standardize the mean squared error by reporting the adjusted R^2 from regressing each of the true compound factors on the estimated factors $\bar{f}_t$. This measure is strictly decreasing in the (root) mean squared error and emphasizes the size of the error relative to the overall factor variation. To enforce stationarity for these evaluation equations, we apply the regressions in differences ($\Delta_4\Delta$) of the true factors and their estimates. These R^2 are averaged over the iterations.

Table 1 gives results for the data generating process (1A) with linearly independent component loadings and without a common components structure. There, the true factor process consists of the $r = 4$ common structural times series components, $f_t = (\mu_t^C, \gamma_t^C, c_t^C, u_t^C)'$, which allows for an evaluation of each component separately. Overall, the principal components in differences outperform the estimates based on levels data. The difference between the methods is most pronounced for larger N and T. The estimates in differences clearly improve with N, but also slightly with T, with R^2 becoming close to one for each component in large samples. The level estimate, however, especially the stationary components in the baseline case with $\chi = 1$ and $\varsigma = 1$, does not show a clear improvement with larger N. The precision typically even worsens with larger T, which is the result of inconsistency when the idiosyncratic components are nonstationary; see Bai and Ng (2004) for the I(1) framework. The results are robust to changing the scale of the stationary components to $\varsigma = 2$ and of the common factors to $\chi = 2$. These changes lead to the expected results that the stationary common components are estimated more precisely in the former case, while the overall precision increases in the latter case.

In Table 2, we show results for the process (1B) which entails the common features assumption that the idiosyncratic components are white noise. Compared to (1A), the overall picture changes. Now, the estimates in levels are better than their difference-based counterparts, most strikingly for larger N. The difference-based estimates still improve both with N and with T. The precisions of the two estimators for the stationary components are closer to each other for $\varsigma = 2$ and for $\chi = 2$, but still the level-based estimates dominate the difference-based ones almost uniformly.

Results for the data generating processes with identical loadings for all common components are depicted in Table 3. We evaluate the precision of r principal components estimating the r compound factors $f_t = \mu_t^C + \gamma_t^C + u_t^C$ for $r \in \{1, 2\}$ by means of the adjusted R^2 as before. The mean of the adjusted R^2 over both evaluation regressions is computed in the case $r = 2$.

For $r = 1$, the adjusted R^2 are very close to one for all chosen specifications. Thus, when compared to Tables 1 and 2, the performance is seemingly enhanced if the components can be estimated in aggregated form, which reduces the compound factor dimension relative to the first two data generating processes. However, the higher uncertainty of the first two cases likely recurs in the second step when distinct structural time series components are estimated from the compound factors in a state space framework. The outcomes for $r = 2$ reveal a loss of precision and visible differences between the specifications and estimators. The patterns described for the first two data generating processes are confirmed here. Most notably, without the common feature assumption the difference estimator outperforms the level estimator again, while the latter is slightly better in case of common features.

Table 1. Precision of common component estimation by principal components in levels and differences ($\Delta_4\Delta$) for process (1A) without common features. The mean of the adjusted R^2 from regressions of true common components on estimated factors is given.

				pca in levels				pca in differences			
χ	s	T	N	μ_t	γ_t	c_t	u_t	μ_t	γ_t	c_t	u_t
1	1	250	10	0.260	0.506	0.270	0.338	0.229	0.551	0.310	0.430
1	1	250	50	0.427	0.633	0.272	0.338	0.507	0.836	0.637	0.756
1	1	250	100	0.499	0.688	0.270	0.333	0.736	0.917	0.812	0.873
1	1	250	500	0.634	0.780	0.278	0.338	0.941	0.983	0.959	0.973
1	1	500	10	0.260	0.496	0.253	0.333	0.232	0.548	0.306	0.418
1	1	500	50	0.369	0.588	0.260	0.336	0.564	0.846	0.676	0.773
1	1	500	100	0.427	0.638	0.260	0.346	0.770	0.922	0.830	0.884
1	1	500	500	0.550	0.720	0.257	0.338	0.951	0.984	0.964	0.976
1	1	1000	10	0.254	0.469	0.257	0.339	0.230	0.548	0.317	0.422
1	1	1000	50	0.309	0.504	0.262	0.348	0.588	0.853	0.698	0.785
1	1	1000	100	0.316	0.503	0.260	0.352	0.783	0.924	0.839	0.888
1	1	1000	500	0.347	0.520	0.265	0.355	0.953	0.985	0.966	0.977
1	2	250	10	0.136	0.287	0.445	0.499	0.096	0.291	0.483	0.603
1	2	250	50	0.254	0.366	0.550	0.516	0.147	0.627	0.799	0.861
1	2	250	100	0.336	0.414	0.584	0.522	0.260	0.783	0.889	0.924
1	2	250	500	0.517	0.516	0.655	0.535	0.806	0.951	0.976	0.984
1	2	500	10	0.127	0.277	0.420	0.499	0.094	0.288	0.490	0.603
1	2	500	50	0.209	0.343	0.439	0.523	0.157	0.655	0.808	0.867
1	2	500	100	0.259	0.371	0.440	0.525	0.369	0.807	0.898	0.928
1	2	500	500	0.425	0.459	0.448	0.532	0.864	0.959	0.979	0.985
1	2	1000	10	0.125	0.263	0.403	0.502	0.091	0.289	0.483	0.600
1	2	1000	50	0.151	0.275	0.407	0.528	0.167	0.668	0.813	0.869
1	2	1000	100	0.160	0.274	0.420	0.528	0.453	0.819	0.901	0.931
1	2	1000	500	0.180	0.294	0.418	0.535	0.883	0.962	0.980	0.986
2	1	250	10	0.492	0.716	0.487	0.473	0.497	0.783	0.605	0.695
2	1	250	50	0.717	0.841	0.514	0.453	0.878	0.959	0.911	0.939
2	1	250	100	0.767	0.869	0.537	0.447	0.941	0.980	0.956	0.970
2	1	250	500	0.825	0.901	0.628	0.422	0.988	0.996	0.991	0.994
2	1	500	10	0.482	0.717	0.417	0.484	0.500	0.788	0.603	0.695
2	1	500	50	0.668	0.820	0.387	0.465	0.884	0.959	0.914	0.940
2	1	500	100	0.714	0.845	0.373	0.463	0.942	0.980	0.957	0.970
2	1	500	500	0.764	0.872	0.378	0.460	0.988	0.996	0.992	0.994
2	1	1000	10	0.453	0.682	0.413	0.493	0.514	0.791	0.609	0.700
2	1	1000	50	0.546	0.726	0.387	0.480	0.885	0.960	0.916	0.940
2	1	1000	100	0.566	0.740	0.396	0.480	0.943	0.980	0.958	0.971
2	1	1000	500	0.620	0.773	0.378	0.466	0.989	0.996	0.992	0.994

These outcomes suggest that the estimator choice should be based on whether the idiosyncratic components are white noise or not, and that unnecessary differencing should be avoided. The difference-based estimator appears as a robust choice since it is consistent in both settings while the level-based estimator does not necessarily improve with sample size in the general framework of this article.

Table 2. Precision of common component estimation by principal components in levels and differences ($\Delta_4\Delta$) for process (1B) with common features. The mean of the adjusted R^2 from regressions of true common components on estimated factors is given.

				pca in levels				pca in differences			
χ	s	T	N	μ_t	γ_t	c_t	u_t	μ_t	γ_t	c_t	u_t
1	1	250	10	0.150	0.331	0.160	0.185	0.079	0.230	0.101	0.163
1	1	250	50	0.300	0.556	0.343	0.237	0.120	0.381	0.157	0.263
1	1	250	100	0.414	0.673	0.468	0.275	0.136	0.440	0.177	0.305
1	1	250	500	0.756	0.900	0.802	0.742	0.176	0.631	0.234	0.519
1	1	500	10	0.145	0.331	0.154	0.180	0.075	0.221	0.098	0.156
1	1	500	50	0.297	0.559	0.343	0.234	0.117	0.383	0.156	0.252
1	1	500	100	0.410	0.677	0.471	0.276	0.132	0.444	0.179	0.298
1	1	500	500	0.763	0.905	0.810	0.794	0.199	0.741	0.302	0.618
1	1	1000	10	0.143	0.338	0.158	0.175	0.073	0.220	0.100	0.148
1	1	1000	50	0.296	0.564	0.342	0.231	0.115	0.382	0.159	0.246
1	1	1000	100	0.414	0.678	0.475	0.287	0.133	0.448	0.183	0.300
1	1	1000	500	0.767	0.907	0.816	0.818	0.247	0.821	0.460	0.710
1	2	250	10	0.110	0.249	0.386	0.402	0.057	0.163	0.281	0.411
1	2	250	50	0.257	0.494	0.651	0.606	0.069	0.213	0.412	0.566
1	2	250	100	0.403	0.659	0.785	0.785	0.074	0.272	0.535	0.688
1	2	250	500	0.767	0.905	0.947	0.957	0.101	0.653	0.843	0.918
1	2	500	10	0.109	0.252	0.391	0.399	0.054	0.158	0.285	0.403
1	2	500	50	0.261	0.506	0.654	0.630	0.063	0.211	0.426	0.577
1	2	500	100	0.402	0.666	0.790	0.801	0.069	0.288	0.595	0.732
1	2	500	500	0.768	0.906	0.947	0.959	0.112	0.786	0.901	0.940
1	2	1000	10	0.109	0.250	0.388	0.394	0.051	0.148	0.279	0.396
1	2	1000	50	0.260	0.509	0.658	0.645	0.060	0.208	0.440	0.589
1	2	1000	100	0.404	0.668	0.791	0.808	0.068	0.330	0.639	0.755
1	2	1000	500	0.768	0.908	0.948	0.960	0.194	0.835	0.922	0.949
2	1	250	10	0.286	0.541	0.331	0.299	0.162	0.442	0.216	0.315
2	1	250	50	0.565	0.792	0.633	0.534	0.213	0.636	0.303	0.499
2	1	250	100	0.716	0.884	0.774	0.764	0.288	0.784	0.437	0.663
2	1	250	500	0.929	0.974	0.946	0.955	0.759	0.953	0.850	0.926
2	1	500	10	0.281	0.540	0.337	0.296	0.156	0.433	0.216	0.306
2	1	500	50	0.571	0.793	0.632	0.564	0.217	0.656	0.322	0.512
2	1	500	100	0.721	0.885	0.776	0.783	0.329	0.818	0.519	0.710
2	1	500	500	0.929	0.975	0.946	0.958	0.868	0.966	0.911	0.947
2	1	1000	10	0.282	0.543	0.332	0.291	0.151	0.426	0.213	0.304
2	1	1000	50	0.571	0.799	0.635	0.582	0.225	0.677	0.347	0.527
2	1	1000	100	0.725	0.887	0.779	0.791	0.407	0.838	0.593	0.743
2	1	1000	500	0.930	0.975	0.947	0.959	0.900	0.970	0.931	0.954

4.3. Estimation of Latent Processes

In a second part of this Monte Carlo study, we assess the two-step procedure with respect to its ability to estimate a latent process θ_{1t} from incomplete data z_{1t} and additional high-dimensional data x_t. We delete $N/4$ of the observations in z_{1t} which is generated together

Table 3. Precision of compound factor estimation by principal components in levels and differences ($\Delta_4\Delta$) for processes (2A) and (2B) (without and with common features). The mean of the adjusted R^2 from regressions of true common components on estimated factors is given.

				$r = 1$				$r = 2$			
				(2A)		(2B)		(2A)		(2B)	
χ	ς	T	N	level	diff	level	diff	level	diff	level	diff
1	1	250	10	0.970	0.994	0.998	0.985	0.644	0.760	0.463	0.283
1	1	250	50	0.973	0.999	1.000	0.999	0.798	0.951	0.752	0.494
1	1	250	100	0.969	0.999	1.000	1.000	0.843	0.976	0.850	0.614
1	1	250	500	0.978	1.000	1.000	1.000	0.932	0.995	0.964	0.919
1	1	500	10	0.977	0.995	0.998	0.986	0.628	0.764	0.469	0.279
1	1	500	50	0.983	0.999	1.000	0.999	0.772	0.952	0.753	0.507
1	1	500	100	0.987	0.999	1.000	1.000	0.825	0.976	0.850	0.684
1	1	500	500	0.990	1.000	1.000	1.000	0.915	0.995	0.964	0.945
1	1	1000	10	0.976	0.994	0.998	0.985	0.616	0.767	0.465	0.280
1	1	1000	50	0.990	0.999	1.000	1.000	0.761	0.953	0.754	0.527
1	1	1000	100	0.992	0.999	1.000	1.000	0.809	0.977	0.850	0.733
1	1	1000	500	0.996	1.000	1.000	1.000	0.907	0.995	0.964	0.954
1	2	250	10	0.965	0.994	0.998	0.992	0.642	0.746	0.586	0.414
1	2	250	50	0.972	0.999	1.000	1.000	0.801	0.951	0.850	0.694
1	2	250	100	0.978	0.999	1.000	1.000	0.843	0.975	0.916	0.857
1	2	250	500	0.974	1.000	1.000	1.000	0.928	0.995	0.981	0.972
1	2	500	10	0.970	0.993	0.998	0.995	0.627	0.750	0.593	0.414
1	2	500	50	0.985	0.999	1.000	1.000	0.772	0.952	0.847	0.731
1	2	500	100	0.987	0.999	1.000	1.000	0.829	0.976	0.916	0.878
1	2	500	500	0.991	1.000	1.000	1.000	0.913	0.995	0.981	0.977
1	2	1000	10	0.976	0.993	0.998	0.995	0.615	0.756	0.593	0.413
1	2	1000	50	0.986	0.999	1.000	1.000	0.759	0.952	0.849	0.756
1	2	1000	100	0.991	0.999	1.000	1.000	0.807	0.976	0.916	0.888
1	2	1000	500	0.996	1.000	1.000	1.000	0.905	0.995	0.981	0.979
2	1	250	10	0.985	0.999	0.999	0.999	0.867	0.930	0.710	0.538
2	1	250	50	0.990	1.000	1.000	1.000	0.948	0.988	0.914	0.872
2	1	250	100	0.988	1.000	1.000	1.000	0.965	0.994	0.955	0.941
2	1	250	500	0.988	1.000	1.000	1.000	0.985	0.999	0.991	0.988
2	1	500	10	0.991	0.998	1.000	0.999	0.844	0.931	0.708	0.540
2	1	500	50	0.994	1.000	1.000	1.000	0.933	0.988	0.915	0.885
2	1	500	100	0.997	1.000	1.000	1.000	0.955	0.994	0.955	0.945
2	1	500	500	0.998	1.000	1.000	1.000	0.982	0.999	0.991	0.990
2	1	1000	10	0.993	0.998	1.000	0.999	0.837	0.929	0.711	0.544
2	1	1000	50	0.996	1.000	1.000	1.000	0.928	0.988	0.915	0.890
2	1	1000	100	0.998	1.000	1.000	1.000	0.949	0.994	0.955	0.948
2	1	1000	500	0.998	1.000	1.000	1.000	0.980	0.999	0.991	0.990

with x_t as a factor STSM for $y_t' = (z_{1t}, x_t')$. Different alternative approaches are considered to estimate θ_{1t} for each observation where z_{1t} is missing. As an infeasible benchmark, (i) the factor STSM with known factor process f_t is considered in state space form, where parameters are determined by maximum likelihood and missing values are estimated by

the state smoother. As the feasible counterpart, (ii) the two-step estimate proposed in this article is used, where f_t is estimated by the principal components based on data in differences $\Delta\Delta_4 x_t$. As one further straightforward benchmark we use (iii) a univariate STSM which neglects information from x_t. Comparison between (ii) and (iii) straightforwardly illustrates the effect of taking into account the common part θ_t^C versus ignoring it.

As a simple competitor that also uses time series information on z_{1t} only, we interpolate the series using (iv) a local mean of available $\Delta_4 z_{1t}$ in the range of ± 20 observations near the period to be estimated. Cross-section information, but not the dynamics of the system are utilized by static regression-based predictions of θ_{1t} using the difference-based principal components of x_t as predictors. The regression is run (v) in levels, (vi) applying a yearly difference operator Δ_4 to z_{1t} and the principal component, or (vii) applying the difference operator $\Delta\Delta_4$ which is sufficient to make the variables stationary. A comparison to the full state space model, possibly using the common features restriction, would allow a measure of the undergone efficiency loss by our method but is beyond the scope of this article: The high dimension makes the treatment computationally intractable both here and in similar empirical problems, so that we omit it from this comparison.

Table 4 shows the corresponding root mean squared errors (RMSE) from estimating θ_{1t} according to the data generating process (1B) with $r = 1$. Not surprisingly, the infeasible estimator (i) outperforms the others, while the feasible two-step strategy (ii) of utilizing the factor STSM comes a close second. The loss from having to estimate f_t is rather small in this specification, and amounts to less than five percent of the overall RMSE in most cases. Clearly, this result may strongly depend on the data generating process and the corresponding precision of the principal components method. The differences vanish with larger N.

The univariate STSM approach (iii) comes in third place, but missing information on the factors leads to an efficiency loss which is more pronounced if either $\varsigma = 2$ which increases the noise which is unpredictable by univariate methods, or if $\chi = 2$ where the information content of x_t is higher. However, taking the dynamics into account appropriately pays off, which turns out from a comparison to the naïve local averaging method that performs clearly worse than all STSM approaches. The static regression estimation with principal components as predictors (v) leads to very spurious results in levels, while it still does not lead to a relevant improvement even over the local averaging method when it is applied in differences (vi, vii).

5. Application to German Hours Worked Statistics

We apply the proposed techniques to the measurement of several components of hours worked in Germany. Official statistics on hours worked per person and the overall volume of work are determined by the IAB which contributes the corresponding time series to the German national accounts. The working time measurement concept is a componentwise system where collective, calendar, cyclical, personal and other components are determined separately on a quarterly basis since 1991, and results are disaggregated according to industries, regions, and employment status; see Wanger (2013) and Wanger et al. (2016) for recent overviews.

Table 4. RMSE of estimating N/4 randomly chosen missing values for z_{1t} in process (2A) with r = 1. (i) Factor STSM with known factors f_t, (ii) factor STSM with differenced pca factor estimates, (iii) univariate STSM, (iv) mean of yearly differences of z_{1t} within ±20 observations, (v)–(vii) static OLS of z_{1t} on pca in levels and differences.

χ	s	T	N	STSM f_t	STSM pca	STSM univar.	Mean Δ_4	OLS level	OLS Δ_4	OLS $\Delta\Delta_4$
1	1	250	10	1.126	1.210	1.627	2.518	4.743	2.217	2.490
1	1	250	50	1.148	1.166	1.653	2.550	4.772	2.189	2.411
1	1	250	100	1.151	1.159	1.634	2.512	4.584	2.133	2.416
1	1	250	500	1.160	1.161	1.663	2.558	4.575	2.136	2.397
1	1	500	10	1.187	1.261	1.644	2.548	10.621	2.642	2.555
1	1	500	50	1.181	1.193	1.612	2.502	9.964	2.498	2.420
1	1	500	100	1.180	1.187	1.644	2.567	9.733	2.498	2.393
1	1	500	500	1.165	1.166	1.598	2.501	9.803	2.454	2.359
1	1	1000	10	1.275	1.328	1.636	2.567	26.310	3.195	2.562
1	1	1000	50	1.255	1.265	1.614	2.526	24.397	2.955	2.388
1	1	1000	100	1.268	1.274	1.630	2.554	24.386	2.986	2.386
1	1	1000	500	1.267	1.270	1.624	2.549	24.330	2.970	2.368
1	2	250	10	1.522	1.650	2.245	3.182	4.899	2.649	3.310
1	2	250	50	1.556	1.582	2.283	3.224	4.923	2.597	3.203
1	2	250	100	1.565	1.576	2.254	3.187	4.728	2.561	3.209
1	2	250	500	1.568	1.570	2.290	3.238	4.727	2.555	3.194
1	2	500	10	1.569	1.692	2.274	3.238	10.705	3.064	3.428
1	2	500	50	1.546	1.568	2.220	3.158	10.062	2.881	3.198
1	2	500	100	1.545	1.557	2.272	3.243	9.826	2.878	3.166
1	2	500	500	1.543	1.545	2.212	3.167	9.889	2.836	3.139
1	2	1000	10	1.595	1.700	2.259	3.246	26.281	3.561	3.400
1	2	1000	50	1.555	1.575	2.220	3.183	24.436	3.292	3.149
1	2	1000	100	1.573	1.583	2.241	3.218	24.424	3.328	3.151
1	2	1000	500	1.587	1.589	2.252	3.232	24.356	3.323	3.154
2	1	250	10	1.200	1.277	2.514	3.837	4.756	2.225	2.496
2	1	250	50	1.232	1.248	2.543	3.886	4.768	2.189	2.412
2	1	250	100	1.221	1.229	2.497	3.800	4.577	2.131	2.414
2	1	250	500	1.244	1.247	2.563	3.909	4.575	2.136	2.397
2	1	500	10	1.404	1.466	2.534	3.884	10.652	2.654	2.563
2	1	500	50	1.384	1.394	2.456	3.781	9.972	2.502	2.422
2	1	500	100	1.395	1.402	2.542	3.915	9.732	2.498	2.393
2	1	500	500	1.357	1.361	2.440	3.771	9.811	2.455	2.358
2	1	1000	10	1.683	1.719	2.512	3.893	26.520	3.209	2.571
2	1	1000	50	1.664	1.673	2.488	3.849	24.408	2.955	2.388
2	1	1000	100	1.677	1.683	2.518	3.901	24.416	2.985	2.385
2	1	1000	500	1.687	1.692	2.505	3.888	24.327	2.969	2.368

During a major revision in 2014 which also affected the methodology of the working time measurement, state space techniques were introduced to enhance the estimation precision for components with incomplete data sources, or where more than one source is used in the measurement. In this section, we describe the computation of the cyclical

components paid and unpaid overtime work as well as flows on working-time accounts. These are of primary importance when assessing the business cycle fluctuations of hours worked in real time.

5.1. Paid and Unpaid Overtime Hours

The computations of overtime hours in the working-time measurement concept are primarily based on two yearly surveys. In the GSOEP, employed persons are asked for the number of performed overtime hours in the recent month and the way overtime work is typically compensated. From the responses, yearly time series on paid and unpaid overtime hours since the 1980s can be constructed, but as has been mentioned in Section 3, a changing distribution of interviews over the year has to be taken into account when considering a target series of higher frequency. As a second primary data source, the Microcensus offers information on paid and unpaid overtime hours since 2010 on the basis of quarterly averages.

The main problem of constructing a quarterly time series in real time is the substantial publication lag of each of the sources, since results from the GSOEP are available approximately twelve months after the end of a reference year, while the Microcensus results typically come in July of the following year. Hence, information regarding the first quarter of each year is available only after about 21 months (GSOEP) and 16 months (Microcensus), respectively. Additionally, the determination of intra-year fluctuations before 2010 is challenging, since until then only yearly GSOEP data are available. In response, we gather additional indicators to tackle these problems and to achieve the highest possible precision for the given available data.

As an additional data source, we consider the Ifo Business Survey from the Ifo Institute (Leibniz Institute for Economic Research at the University of Munich). In this survey, establishments are asked in the last month of each quarter whether their employees currently perform overtime work. Along with the log of the GSOEP and of the Microcensus measures of overtime hours per week (z_{1t} and z_{2t}, respectively), the logarithmic fraction of establishments with overtime work enters the model as a third series of interest, z_{3t}.

Further economic and labor market indicators (x_t) are used to compute principal components which enter the factor STSM. Here, we use real gross domestic product, the production index, new orders for all manufacturing industries, the number of employed persons, real compensation per employee (all from the Federal Statistical Office), registered unemployment (from the Federal Employment Agency), business expectations, business assessment and the employment barometer (from the Ifo Institute) as well as the willingness to buy index (from GfK Nuremberg). These variables are considered informative when assessing the current business cycle and labor market development, and hence for the amount of overtime work. We refrain from using a data set of higher dimension, since the additional data are likely to introduce irrelevant information and require a higher number of factors. At the same time, we keep the updating process simple by this choice.

Principal component estimates are computed after applying the natural logarithm to all variables except business expectations, business assessment, the employment barometer

and the willingness to buy index. Seasonally adjusted data are used in x_t, so that yearly differences are not needed to remove seasonal nonstationarity. Additionally, there is no evidence for a changing slope in the processes: The p-values for tests of $\Sigma_\zeta = 0$ in univariate STSM is 0.97 and 0.07 for the first and second principal component (based on second differences), respectively. Hence, we base the subsequent analysis on re-cumulated principal components from first differences of the raw data. Data gaps and mixed frequency issues in x_t are resolved by the algorithm described by Stock and Watson (2002).

The resulting models for paid and unpaid overtime hours, respectively, are formulated in terms of a monthly model frequency to precisely capture the timing of the measurement process. Along with the r estimated factors $\bar{f}_t$, which capture the compound common components θ_t^C on a monthly basis, the measurement model is given by

$$\begin{pmatrix} \bar{f}_t \\ \log(ot\_gsoep_t) \\ \log(ot\_mc_t) \\ \log(ot\_ifo_t) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ d_2 \\ 0 \end{pmatrix} + \begin{pmatrix} I & 0 & 0 \\ 0 & M_{11,t}(L) & 0 \\ 0 & M_{21,t}(L) & 0 \\ 0 & 0 & M_{33,t}(L) \end{pmatrix} \begin{pmatrix} \theta_t^C \\ \theta_{1t} \\ \theta_{2t} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \varepsilon_{2t} \\ 0 \end{pmatrix}.$$

The model comprises a monthly variable $\bar{f}_t$, a yearly ot_gsoep_t which is brought into the model at the last month of the year only, the quarterly ot_mc_t, which is brought in at the last month of the quarter, and the likewise quarterly ot_ifo_t, which refers to a single month of each quarter where it comes into the model. The GSOEP measurement scheme $M_{11,t}(L)$ is determined by the changing proportion of GSOEP-interviews in each month as in (8) so that $M_{11,t}(L) = M_{11,t,dec} + M_{11,t,nov}L + \ldots + M_{11,t,jan}L^{11}$. As an example for the year 2013, the obsevation of ot_gsoep_t is given for $t = 2013M12$, while all other months are missing. The proportion of interviews accross months in 2013 leads to $M_{11,2013M12,jan} = 0.0010$ since only 0.10% of interviews took place in January, $M_{11,2013M12,feb} = 0.2656$ since 26.56% of the sample were interviewed in February, $M_{11,2013M12,mar} = 0.2824$ for a proportion of 28.24% of interviews in March, and so on. In contrast, $M_{11,2013M01,m}, \ldots, M_{11,2013M11,m} = 0$ for all m, since the yearly value is brought into the model in the last month of the year, and hence all other months are missing. The Microcensus measures the same underlying θ_{1t} as the GSOEP, but by quarterly averages according to (7), so that $M_{21,t}(L) = \frac{1}{3} + \frac{1}{3}L + \frac{1}{3}L^2$ for each t reflecting the last month of the quarter (March, June, September, or December), and $M_{21,t}(L) = 0$ for other values of t. The Ifo measure of overtime refers to a single month, and hence $M_{33,t}(L) = 1$ if data are available in month t and $M_{33,t}(L) = 0$, otherwise.

Since for a long sample of data prior to 2010 the GSOEP is the only available statistic directly measuring θ_{1t}, we implement this source as a benchmark, and force a weighted average of θ_{1t} to fit the yearly GSOEP figure exactly. Recent Microcensus figures, in contrast, enter the model with an adjustment term d_2, and the survey error ε_{2t} is modeled with a fixed variance which is estimated within the state space model. Since quantitative information on the survey autocorrelation due to overlap is not available, we model the survey error as serially uncorrelated.

Since the Ifo series does not directly measure the overtime hours, but rather serve as a correlated indicator, the inclusion of a measurement error is not important here. The measurement error would be anyway exchangeable with the noise term in the dynamic model, because the Ifo series estimates a single underlying series, and because the measurement scheme does not involve a filter of lagged values so that there is no autocorrelation induced by the measurement scheme.

We choose an unrestricted multivariate STSM formulation of θ_t^C and θ_t as the dynamic model. In contrast to the Formulation (3), this approach allows for correlation between θ_{1t} and θ_{2t} (fraction of establishments with overtime) beyond their dependence on the common components, while nesting the strict factor specification. The need for this additional flexibility is reasonable since the Ifo survey measures a concept relatively close to the target series, and may provide specific information beyond the overall business cycle.

A large gain in parsimony is associated by using only one factor instead of all ten indicators in the model and hence drastically reducing the model by means of reduced rank sparsity. Additionally, however, we conduct model selection and reduce the model parameters by dropping different individual components from the model. The decision to drop components is drawn from sequential tests based on each series individually. Augmented Dickey Fuller tests, with lag lengths determined by AIC, fail to reject unit roots for each of the series considered in the models (the exception being $\bar{f}_{1t}$, with a p-value of 0.0029). We hence include unit root components for all series and let $\xi_{it} \neq 0$ in general. We test the presence of slope changes ζ_t, white noise terms u_t, cyclical components c_t, and changes to the seasonal pattern ω_t in univariate STSMs, and present the p-value of the corresponding hypotheses in Table 5. The time series of Microcensus data is not sufficiently long for univariate analyses so that we base the specification for paid and unpaid overtime on the yearly GSOEP series.

On a five percent significance level, there is no evidence for a changing slope in either of the series. This again supports the computation of principal components based on first differences rather than second differences of the data. We hence set Σ_ζ to zero in both the model for paid and for unpaid overtime hours. According to additional test results, we include a white noise term for the principal components, but not for the series in z_t in what follows. There is relatively strong evidence on the presence of cyclical components which seems to be needed in each of the observed series. Finally, the Ifo survey is the only series with a seasonal component that is reasonably modeled with a fixed seasonal pattern. The

Table 5. P-values from testing different null hypotheses on the presence of several components in univariate structural time series models. The tests refer to the full model in the alternative. Models are formulated at the original data frequency (monthly for $\hat{f}_t$ yearly for GSOEP, quarterly for Ifo Business Survey).

	$\hat{f}_{1t}$	$\hat{f}_{2t}$	Paid Ot. (GSOEP)	Unpaid Ot. (GSOEP)	Overtime (Ifo)
$H_0 : \varsigma_t = 0$	0.4213	0.0898	0.6815	0.9735	1.0000
$H_0 : u_t = 0$	0.0008	0.0433	1.0000	0.0676	1.0000
$H_0 : c_t = 0$	0.0000	0.0010	0.0053	0.0133	0.0000
$H_0 : \omega_t = 0$	—	—	—	—	0.4396

Table 6. Estimated parameters for cyclical components in models for paid and unpaid overtime hours (first two columns) and flows on working time accounts (last two columns). A similar cycles assumption is imposed in each of the models.

	Paid Ot.	Unpaid Ot.	Inflow WTA	Outflow WTA
Dampening factor ρ	0.9832	0.9880	0.9835	0.9835
Angle frequency λ	0.1155	0.1128	0.1198	0.1198
Period $\frac{2\pi}{\lambda}$	54.42	55.70	52.45	52.45
Cycle standard deviation	8.45	3.89	15.04	9.34
Cycle shock correlation with κ_t^c	0.69	0.45	0.60	−0.42
Phase shift δ	−2.60	−3.15	0.89	−7.79

seasonal figure in overtime hours comes in only trough the short Microcensus time series and has therefore also to be set fixed.

Considering the joint dynamic process introduced above, a decomposition

$$\begin{pmatrix} \theta_t^C \\ \theta_{1t} \\ \theta_{2t} \end{pmatrix} = \begin{pmatrix} \mu_t^C \\ \mu_{1t} \\ \mu_{2t} \end{pmatrix} + \begin{pmatrix} 0 \\ \gamma_{1t} \\ \gamma_{2t} \end{pmatrix} + \begin{pmatrix} c_t^C \\ c_{1t} \\ c_{2t} \end{pmatrix} + \begin{pmatrix} u_t^C \\ 0 \\ 0 \end{pmatrix}$$

applies, with dynamic components driven by the processes introduced below Equation (1). There, Σ_κ and Σ_ξ are full symmetric $(r+2)\times(r+2)$ parameter matrices, while Σ_u is a scalar and $\Sigma_\zeta = \Sigma_\omega = 0$.

Both for paid and unpaid overtime hours, models with $r = 1$ are estimated as the baseline specifications, which appears reasonable due to the relatively small number of indicators in x_t and avoids parameter abundance. Setting $r = 2$ while using the same modeling strategy does not change the estimated time series in a relevant way. We assess whether the data are consistent with a similar cycles assumption ($\rho_i = \rho$, $\lambda_i = \lambda$) and whether the overtime measures θ_{1t} and θ_{2t} have the same cycle shift with respect to the business cycle factor ($\delta_2 = \delta_3$). These restrictions are rejected neither for paid, nor for unpaid overtime hours on a 5% significance level, so that they are maintained. The estimated cyclical parameters are shown in the left two columns of Table 6.

For both models, we find that the cycles are relatively persistent, with a dampening factor close to one, and that a typical cycle lasts about four and a half years. The cycles are shifted by approximately three months to the right relative to the business cycle of the principal component, so that a peak in overtime hours typically lags behind that of the factor. Paid overtime hours appear to be more pro-cyclical, since the standard deviation of the factor (log-scale × 100) is more than twice as large as that for unpaid overtime hours. At the same time, paid overtime hours exert a stronger correlation with the business cycle.

Further results on the volatility and correlations of cycle and trend shocks are given for paid overtime in Table 7 and for unpaid overtime in Table 8. We observe strong positive cycle correlations also for the Ifo overtime hours, which justifies the inclusion of this series. From the standard deviations of ξ_{1t} for both model, it can be seen that the trend is

Table 7. Estimated standard deviations (main diagonal) and correlations (below diagonal) of cycle shocks κ_t (left) and trend shocks ξ_t (right) for paid overtime model.

	κ^C	κ_1	κ_2		ξ^C	ξ_1	ξ_2
κ^C	1.10	—	—	ξ^C	0.60	—	—
κ_1	0.69	1.55	—	ξ_1	−0.70	1.72	—
κ_2	0.95	0.43	5.78	ξ_2	−0.94	0.41	4.12

more volatile for unpaid than for paid overtime, mirroring the larger impact of the cycle on the paid overtime hours.

Figure 2 shows the smoothed estimate for paid overtime hours. The observations of the GSOEP (round points, placed in March of each year) and Microcensus (crosses, net of the constant d_2) are shown along with the trend μ_{1t} (dotted), the seasonally adjusted estimate $\mu_{1t} + c_{1t}$ (dashed) and the overall smoothed series including the seasonal component (solid). The ordinate axis is depicted on a logarithmic scale to reflect the logarithmic model formulation. The nearly linear long-term downward trend is visibly superimposed by stochastic cycles which had a pronounced effect during the 2008/09 financial and economic crisis and reflects well-known patterns from cyclical output movements. The fixed seasonal component, which shows higher overtime usage in the second half of the year, stems mostly from the short sample of Microcensus observations, and should therefore be treated with care.

The unpaid overtime hours, shown in Figure 3, are driven by a rather volatile trend which closely follows the observations. There are several periods of longer upward or downward movements, and although unpaid hours rose in tendency over the whole sample, there is a decline since about 2006 until now. The cycle is rather small, which reflects the low business cycle sensitivity of this working-time component, while the seasonal component is positive in the first and fourth quarter.

We assess the stability of the models by studying estimated parameters in different subsamples. We estimate both the paid and the unpaid overtime models for subsamples of two thirds of the monthly observations (200 of the 300 months from 1991 to 2015). Table 9 shows results for paid overtime in the left and results for unpaid overtime in the right panel, where the subsample ranging from January 1991 to August 2007 is denoted by Smpl 1, the subsample from March 1995 to October 2011 is denoted by Smpl 2, and the subsample from May 1999 to December 2015 is denoted by Smpl 3.

We find notable differences in the cyclical properties: The first subsample has a more persistent cycle (higher ρ), smaller frequency (smaller λ) and a longer phase shift of the overtime variables (absolutely larger ξ) for both models. Also the standard errors of trend

Table 8. Estimated standard deviations (main diagonal) and correlations (below diagonal) of cycle shocks κ_t (left) and trend shocks ξ_t (right) for unpaid overtime model.

	κ^C	κ_1	κ_2		ξ^C	ξ_1	ξ_2
κ^C	1.14	—	—	ξ^C	0.41	—	—
κ_1	0.45	0.60	—	ξ_1	0.21	2.87	—
κ_2	0.94	0.74	5.51	ξ_2	−1.00	−0.28	4.54

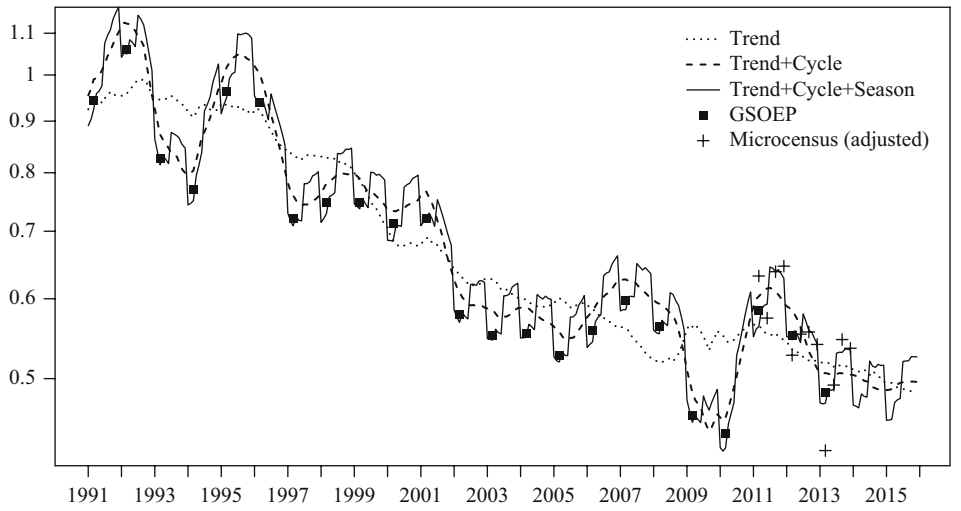


Fig. 2. Paid overtime hours per week. The trend, cycle and seasonal figures are obtained by the state smoother and shown along with the GSOEP and Microcensus observations. The latter is adjusted for the constant d_2.

shocks ($sd(\xi_{jt})$) are subject to change, most prominently for the first and second series of the paid overtime model, where the standard deviations change by a factor of two or more, and large trend variance in a given sample is associated with a smaller cycle variance (smaller $sd(\kappa_{jt})$). The correlations ($corr(\xi_{jt}, \xi_{it})$ and $corr(\kappa_{jt}, \kappa_{it})$) even change sign in some cases. The apparent structural instability is due to multiple maxima of the likelihood function, where different local maxima dominate in different subsamples. Improved stability, for example, by averaging over different local maxima or applying numerical

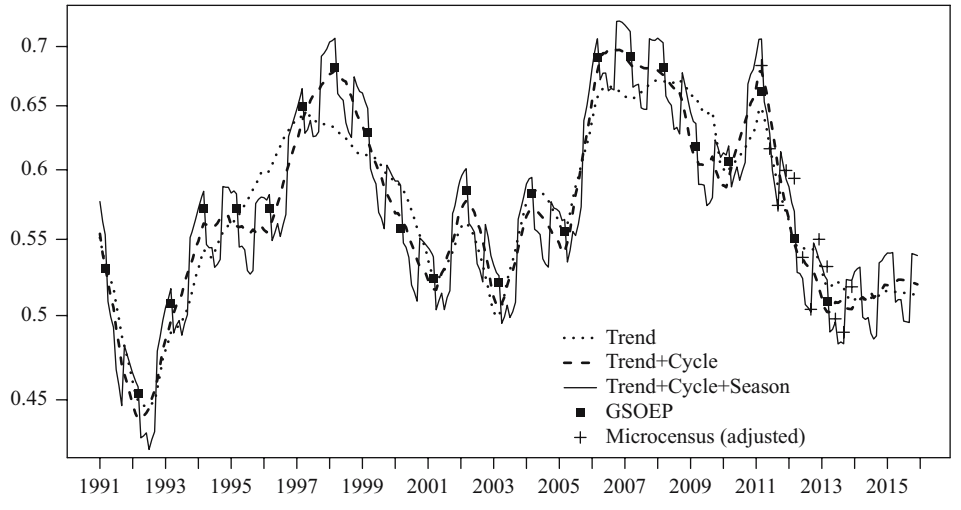


Fig. 3. Unpaid overtime hours per week. The trend, cycle and seasonal figures are obtained by the state smoother and shown along with the GSOEP and Microcensus observations. The latter is adjusted for the constant d_2.

Table 9. Estimated parameters for paid (left) and unpaid (right) overtime models and different subsamples each with two-thirds of the observations: First (Smpl 1), middle (Smpl 2), and last portion of the data (Smpl 3).

Par.	Smpl 1	Smpl 2	Smpl 3
ρ_1	0.9926	0.9803	0.9805
ρ_2	0.9926	0.9803	0.9805
ρ_3	0.9926	0.9803	0.9805
λ_1	0.0714	0.1064	0.0995
λ_2	0.0714	0.1064	0.0995
λ_3	0.0714	0.1064	0.0995
ξ_1	0.0000	0.0000	0.0000
ξ_2	−9.4151	−2.9368	−4.8008
ξ_3	−9.4151	−2.9368	−4.8008
$sd(\xi_{1t})$	0.3430	0.1706	0.5064
$sd(\xi_{2t})$	1.0588	2.5260	1.4812
$sd(\xi_{3t})$	5.6845	5.9511	5.7918
$corr(\xi_{1t}, \xi_{2t})$	−0.9936	−0.7646	−0.9998
$corr(\xi_{1t}, \xi_{3t})$	−0.9998	−0.3016	0.3731
$corr(\xi_{2t}, \xi_{3t})$	0.9956	−0.3837	−0.3893
$sd(\kappa_{1t})$	0.9774	1.3254	1.2326
$sd(\kappa_{2t})$	2.6402	0.8432	1.3161
$sd(\kappa_{3t})$	6.9054	5.9609	5.4484
$corr(\kappa_{1t}, \kappa_{2t})$	0.7252	0.9432	0.9397
$corr(\kappa_{1t}, \kappa_{3t})$	0.9372	0.8800	0.8430
$corr(\kappa_{2t}, \kappa_{3t})$	0.4404	0.6725	0.9761

Par.	Smpl 1	Smpl 2	Smpl 3
ρ_1	0.9944	0.9900	0.9809
ρ_2	0.9944	0.9900	0.9809
ρ_3	0.9944	0.9900	0.9809
λ_1	0.0718	0.1065	0.0999
λ_2	0.0718	0.1065	0.0999
λ_3	0.0718	0.1065	0.0999
ξ_1	0.0000	0.0000	0.0000
ξ_2	−6.9145	−5.3627	−2.9933
ξ_3	−6.9145	−5.3627	−2.9933
$sd(\xi_{1t})$	0.3935	0.6908	0.2875
$sd(\xi_{2t})$	3.6219	3.1814	3.0331
$sd(\xi_{3t})$	5.7547	6.1983	5.5740
$corr(\xi_{1t}, \xi_{2t})$	0.4744	0.9061	0.7796
$corr(\xi_{1t}, \xi_{3t})$	−0.9431	0.5471	0.1223
$corr(\xi_{2t}, \xi_{3t})$	−0.1546	0.1417	−0.5262
$sd(\kappa_{1t})$	0.9394	1.1645	1.2796
$sd(\kappa_{2t})$	0.1545	0.6507	0.7645
$sd(\kappa_{3t})$	6.8822	5.8553	5.3710
$corr(\kappa_{1t}, \kappa_{2t})$	−0.9954	−0.5995	0.4459
$corr(\kappa_{1t}, \kappa_{3t})$	0.9998	0.8345	0.8958
$corr(\kappa_{2t}, \kappa_{3t})$	−0.9963	−0.0594	0.7973

integration in a Bayesian approach are beyond the scope of this article, but are an important topic of future research.

To illustrate the effect of using a multivariate approach, we contrast the result to a univariate structural time series approach. To mimic the dynamic specification in the multivariate approach, we fit a model consisting of level and cycle to the yearly GSOEP series, but omit the seasonal which would not be identified. We use the same measurement scheme as $M_{11,t}(L)$ above in the multivariate model. In Figure 4 the seasonally adjusted multivariate estimate (dashed line) for paid overtime hours is shown along with the GSOEP observations (points, again placed in March of each year) and the smoothed estimate from the univariate approach (solid line). We see only slight differences between the lines in the time span before 2013: The GSOEP observations are used as a benchmark for the yearly weighted means of the estimates, and hence the latter do not move too far away from the former. As a slight deviation, we see that a decline in overtime hours before the global financial crisis is detected already in 2007 using all available indicators, while the univariate GSOEP model gives a smoother change to the "crisis regime". After 2010 when the Microcensus data come in and especially after the last GSOEP observation the multivariate estimates clearly uses more relevant information than just an extrapolation of the dynamics. Hence the additional indicators have the greatest impact where information is most valuable: at the current edge. The figure for unpaid overtime hours, which shows only minor differences between the univariate and the multivariate model, is available from the author upon request.

Finally, we investigate the advantage of our modeling approach with respect to the quality of early estimates in official statistics. A numerical assessment of the overall precision of our estimates of paid and unpaid overtime hours is not possible: The truth is not known and hence a straightforward benchmark as in the simulation study is not

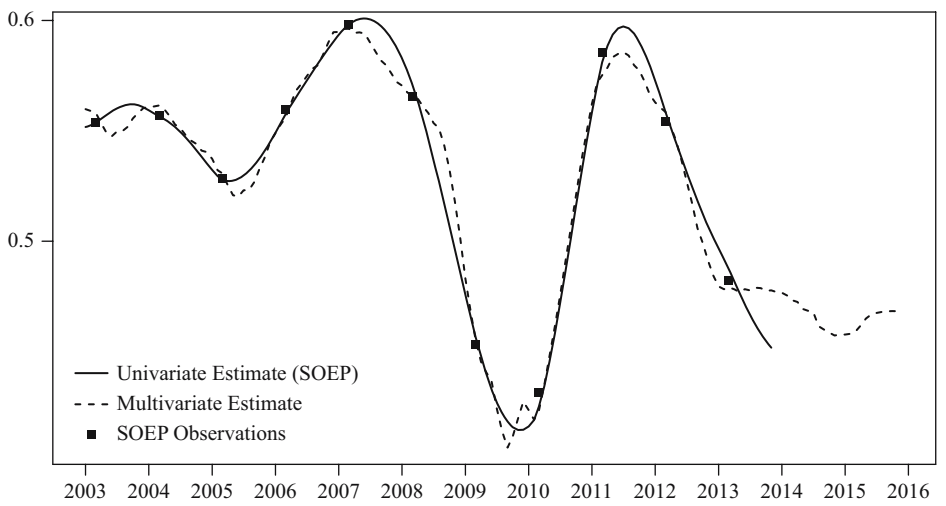


Fig. 4. Paid overtime hours per week. The seasonal adjusted multivariate estimate (dashed) is shown along with a univariate structural time series model (solid) using SOEP data alone. The univariate model implies the same dynamic properties as the multivariate model described in Subsection 5.1 and uses the same measurement scheme with respect to the SOEP observations.

available. We can however assess the timeliness of precise estimates by an ex-post comparison of early estimates in real time to the final estimates where all data are available. We figure out the value of additional information beyond the GSOEP data by comparing the multivariate model to the univariate benchmark. Since the seasonality in the multivariate model can be estimated only in recent years, the seasonal patterns in the Microcensus would cause problems in the pseudo real-time study. We hence omit the Microcensus data from the multivariate model and ignore seasonality both for the univariate and multivariate model in our study.

We conduct the real-time experiment as follows: For the construction of data for a given quarter, we take the indicators contained in x_t and the Ifo Business survey for that quarter as known. The GSOEP data are used with a time lag; the data from two years earlier are used so that for example from the beginning of 2015 the data for 2013 are available. This is a realistic timing since first estimates are typically constructed in the middle of the subsequent quarter. Having constructed pseudo-real-time estimates by a Kalman filter for each of the two models and each quarter from 2004Q1 to 2015Q4, we compare them to the ex-post estimate of each model using all available data. From this comparison, the biases and root mean squared errors can be computed which are shown in Table 10. There, along with the full evaluation sample (all), also results for successive subsamples are given which divide the evaluation sample in four intervals of three years.

Overall, in terms of the RMSE, we find that the multivariate model outperform the univariate approach for paid overtime, while the approaches perform similar for unpaid overtime where the univariate approach is slightly better. This is in line with the finding that paid overtime is more correlated to business cycle indicators and thus the latter help estimate paid overtime hours better than unpaid hours. The very parsimonious univariate model has a smaller bias for both target measures. Considering the subsamples, it is reassuring for the multivariate approach that the latter outperforms the univariate approach in the last subsample, where more data are used in the Kalman filter. The multivariate model is likely to dominate in larger samples which is reflected here. In sum, it is preferable to use a factor approach as the one considered in this article if the target series has a strong business cycle correlation, and if long enough time series are available.

Table 10. Bias and root mean squared error (RMSE) for paid (left) and unpaid (right) overtime model and univariate (univ.) and multivariate (mult.) model. An evaluation sample from 2003Q4 to 2015Q4 is used (all) and four successive subsamples thereof.

	Paid Overtime				Unpaid Overtime			
	BIAS		RMSE		BIAS		RMSE	
Smpl	univ.	mult.	univ.	mult.	univ.	mult.	univ.	mult.
all	−4.047	−5.680	10.846	9.582	0.217	2.373	11.064	11.402
1	4.212	−0.761	7.681	8.126	−2.888	−11.052	10.648	12.830
2	−11.157	−12.923	14.241	13.613	−4.426	2.935	11.991	9.268
3	−4.712	−9.152	13.034	10.665	−0.059	10.311	8.521	11.451
4	−4.530	0.116	6.232	1.474	8.243	7.300	12.643	11.764

5.2. Net Flows on Working Time Accounts

Not all additional hours worked by employees in a given period lead to a definitive increase in the amount of labor over a longer time span. Some of them, termed transitory overtime hours, are compensated by leisure time in a future period. The number of these additional hours worked hence raise the credits on WTA, which are formal arrangements to record such additional hours worked. When measuring hours actually worked per period, the statistician has to track such inflows on WTA which raise hours worked, but also the outflows from WTA which reduce the overall hours worked.

Only few data sources are available which allow to measure in- and outflows from WTA in Germany on a regular basis. Besides paid and unpaid overtime hours, the GSOEP questionnaire asks for overtime hours which are compensated with time-off, and which we hence treat as inflows on WTA. A question regarding the reduction of such hours has been included in the questionnaire only in 2014 and the results are not yet available. A similar objection is faced by a new question regarding balances on WTA in the IAB Job Vacancy Survey. It has been included in the establishment survey in 2013 and therefore still lacks a sufficient history to base long time series estimates thereupon.

The Microcensus holds additional information on WTA flows over a longer time span, which we exploit in our estimation strategy. Each employed household member is asked for the regular weekly hours worked and for hours worked last week. If both differ, the main reason for that difference is inquired, where possible answers include "compensation for more hours worked (e.g., flexible working hours)" if actual hours were lower and "hours for the accumulation of the time credit or for the reduction of time dept" if they were higher than usual. These or analogous questions are available for the whole estimation period.

Since only the *main* reason for a difference is asked for in the Microcensus, there are likely further WTA in- or outflows that are not revealed by the survey participants and hence the results are biased. Our strategy thus combines information on the level of gross inflows from the GSOEP with cyclical variations of the Microcensus figures on in- and outflows around their trends to arrive at a final estimate of net flows. The maintained assumption is that even if both WTA in- and outflows follow (possibly stochastic) trends, the latter should be identical so that there is no long-run discrepancy between the both, and the net flows average to zero in the long run. This allows us to estimate the trend by use of the GSOEP series, while relative deviations from it are determined from the Microcensus. Stated jointly with the estimated factors, the measurement model is

$$\begin{pmatrix} \bar{f}_t \\ \log(in\_mc_t) \\ \log(out\_mc_t) \\ \log(in\_gsoep_t) \end{pmatrix} = \begin{pmatrix} 0 \\ d_{1t} \\ d_{2t} \\ 0 \end{pmatrix} + \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & M_{11,t}(L) & 0 & 0 \\ 0 & 0 & M_{22,t}(L) & 0 \\ 0 & 0 & 0 & M_{33,t}(L) \end{pmatrix} \begin{pmatrix} \theta_t^C \\ \theta_{1t} \\ \theta_{2t} \\ \theta_{3t} \end{pmatrix}.$$

The measurement polynomials $M_{11,t}(L)$, $M_{22,t}(L)$ and $M_{33,t}(L)$ are again designed to fit the characteristics of the Microcensus and GSOEP surveys and in particular the distribution of interviews over the time-spans for which the surveys can be distinctly evaluated. In contrast to the overtime models in Subsection 5.1, for which Microcensus data are available only since 2010, the restructuring of the Microcensus has to be taken into

account in the current WTA model. In 2005, a fixed reference week each year (or less frequently before 1995) was replaced by a continuous interviewing policy. This leads to a change in the polynomials, which are given by $M_{11,t}(L) = M_{22,t}(L) = 1$ before 2005 and $M_{11,t}(L) = M_{22,t}(L) == \frac{1}{3} + \frac{1}{3}L + \frac{1}{3}L^2$ since then. Additionally, a level shift results from the changing survey practice at this time which we model by setting d_{1t} and d_{2t} to nonzero constants before 2005 and to zero afterwards. For the GSOEP measurement, given by $M_{33,t}(L)$, we use the same approach as for the paid and unpaid overtime models described in Subsection 5.1.

In contrast to the case where two surveys measure the same underlying process, in the case of one survey per series the survey error variance cannot be estimated from the data when the underlying process has an additional noise term. The survey error variance could be rather set fixed, based on further information from the survey design. As the pure sampling uncertainty is very small for the large sample of the Microcensus, we do not model survey errors explicitly in this case and set $\varepsilon_t = 0$, while allowing irregular components within the model for θ_t.

We again conduct model selection by studying the individual processes first, and include only components in the joint model which appear worthwhile from the univariate tests. We thus again include a unit root component for all series in order to reflect results from Augmented Dickey Fuller tests. A univariate analysis of the individual components similar to Table 5 reveals that the GSOEP inflow series has a significant slope change (p-value 0.04), while a noise term finds more support from the data than a cycle (for which the p-value is 0.16).

For the Microcensus series, we set $\zeta_t = 0$ and include a noise term along with the cycle and random walk trends, which is also supported by statistical tests in the multivariate model. As an a-priori modeling decision to gain parsimony, correlations between the GSOEP and other series are not considered and hence the former is used solely to extract its trend by univariate filtering and smoothing. The model is thus given by

$$\begin{pmatrix} \theta_t^C \\ \theta_{1t} \\ \theta_{2t} \\ \theta_{3t} \end{pmatrix} = \begin{pmatrix} \mu_t^C \\ \mu_{1t} \\ \mu_{2t} \\ \mu_{3t} \end{pmatrix} + \begin{pmatrix} 0 \\ \gamma_{1t} \\ \gamma_{2t} \\ 0 \end{pmatrix} + \begin{pmatrix} c_t^C \\ c_{1t} \\ c_{2t} \\ 0 \end{pmatrix} + \begin{pmatrix} u_t^C \\ u_{1t} \\ u_{2t} \\ u_{3t} \end{pmatrix},$$

where Σ_ζ has a single nonzero element associated with θ_{3t}, Σ_u is diagonal, and Σ_ξ as well as Σ_κ are block diagonal with a full upper left 3×3 submatrix.

The properties of the cyclical components of WTA in- and outflows are summarized in the right two columns of Table 6. Again, we cannot reject the similar cycles restriction, and the common period and the dampening factor are similar to the case of overtime hours. Both components have a relatively strong cyclical pattern, and WTA inflows have the highest cycle standard deviation among the variables under consideration. Not surprisingly, shocks to inflows are positively, while outflow shocks are negatively related to business cycle shocks. The phase shifts mean that typically seven months after employees have built up most credit on the accounts, the outflows peak and reduce the savings on WTA.

The cyclical patterns of in- and outflows are shown in Figure 5, where $c_{jt} + u_{jt}, j = 1,2$, is depicted for inflows (solid line) and outflows (dashed line) in logarithmic scale $\times$ 100, as annotated on the left axis. The mentioned phase shift between the cycles becomes evident here. At most of the visible peaks of WTA inflows, the outflows are rising and reach their highest value a few months later. Before the building up of credits beginning in 2005, the outflows dropped, while the credits were used up afterwards during the 2008/09 crisis, where outflows peaked again. The trending behavior of transitory overtime hours from the GSOEP, which is used as the trend in both, in- and outflows from WTA, is shown in hours per week as the thin dash-dotted line with annotation at the right axis. It shows a flattening growth from below 0.5 hours per week to over one hour until 2010, and has diminished slightly over the recent years.

The trend and log-scale cycles are combined multiplicatively to yield the net flow on WTA, which is the relevant statistic measuring the effect on hours worked per period. We compute this effect as

$$\Delta WTA_t \approx \exp\left(\mu^z_{3t}\right)\left(\gamma^z_{1t} + c^z_{1t} + u^z_{1t} - \gamma^z_{2t} - c^z_{2t} - u^z_{2t}\right). \tag{9}$$

This overall effect is plotted in Figure 6, where also the seasonal patterns are assessed. The overall increase in the scale of the fluctuations over time is partly due to the increased overall importance of WTA corresponding to the upward trend of gross flows described above, while the cyclical patterns from Figure 5 are closely reflected by the overall net flows.

As for the results of paid and unpaid overtime hours, further processing of the data is performed within the working-time measurement concept to yield quarterly results which are partly decomposed for several groups of employees. These are published by the IAB in the form of working time components tables, and also enter the publication of national accounts by the German Federal Statistical Office.

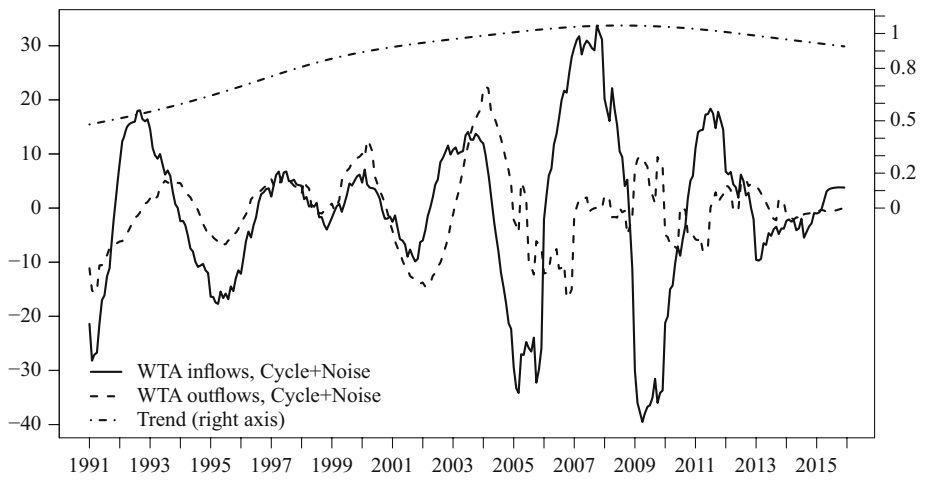


Fig. 5. Cyclical and noise components of in- and outflows (left axis) and trend in flows on working time accounts (right axis). The cycle, noise and trend figures are obtained by the state smoother. Cycles and trends are combined multiplicatively according to (9) to obtain estimated WTA net flows.

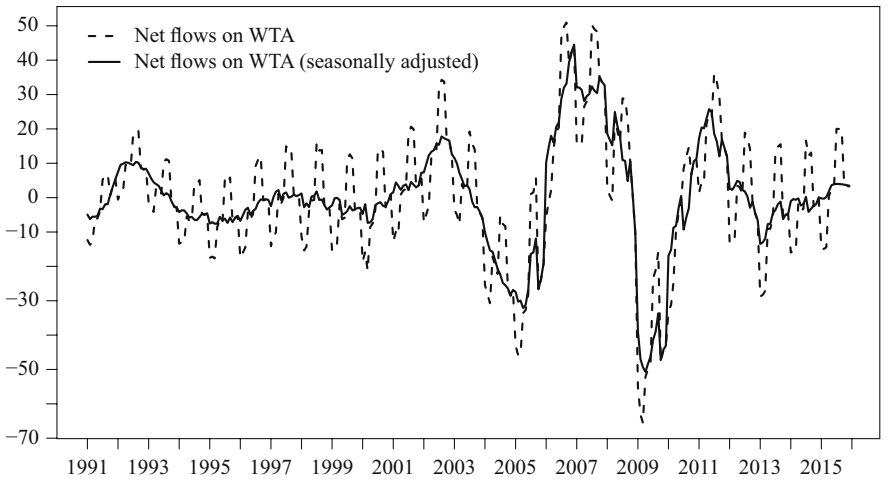


Fig. 6. Working time account net flows in hours per week, computed from smoothed cycles and trends by $\Delta WTA_t \approx \exp\left(\mu^z_{3t}\right)\left(\gamma^z_{1t} + c^z_{1t} + u^z_{1t} - \gamma^z_{2t} - c^z_{2t} - u^z_{2t}\right)$.

6. Conclusion

We have proposed a factor structural time series model and discussed its implementation for possibly high-dimensional problems in official statistics. The model has an intuitive appeal due to its additive, componentwise structure and is quite unrestrictive in its formulation. It is straightforward to apply using a two-step approach with principal components and state-space techniques. In simulation experiments, we found that the two-step approach works reasonably well and that the new method outperforms several competitors in terms of its ability to estimate an unobserved target series. These results show the potential to construct more timely and precise official statistics that use a wide array of recently available data. The empirical application in the article illustrated the usefulness of the method for the measurement of working time components. There, the model was used to construct a time series that is longer, more frequent, and uses more recent information than single survey data sources alone.

The main motivation of the approach was for smoothing latent series using surveys and several other indicators, as is of foremost importance for statistical agencies. However, the methods may reveal their strength also for other tasks such as exploration of the componentwise dynamic properties and co-movements of several macroeconomic time series, as well as forecasting. Additional research may also be concerned with correlated unobserved components models in high dimensions, which allow for a more flexible modeling of spillovers and structural identification.

Appendix A

Properties of the Factor Model in Differences

In this appendix, we show that the data generated by a factor STSM satisfy strong assumptions on time and cross-section dependence when suitably differenced. These

are sufficient to ensure key results of Bai and Ng (2002). Denoting $X_{it} := \Delta\Delta_s y_{it}$, $i = 1, \ldots, N$, and $F_{jt} := \Delta\Delta_s f_{jt}$, $j = 1, \ldots, r$, where y_t is the vector of observed variables and f_t is the compound factor process introduced in the main text, the factor model can be stated as

$$X_{it} = \Lambda_i . F_t + e_{it}.$$

Here, e_{it} is cross-sectionally uncorrelated and independent from F_t by assumption, while both F_{jt} and e_{it} follow strictly stationary linear Gaussian processes with absolutely summable coefficients, as we discuss in the following.

To see the dynamic properties more clearly, a generic element from e_{it} is stated as

$$e_{it} = \Delta\Delta_s \mu_{it} + \Delta\Delta_s \gamma_{it} + \Delta\Delta_s c_{it} + \Delta\Delta_s u_{it},$$

where the I superscript is suppressed for notational simplicity. Since $\mu_{it} = \mu_{i,t-1} + \nu_{i,t-1} + \xi_{i,t-1}$, we have $\Delta\mu_{it} = \nu_{i,t-1} + \xi_{i,t-1}$, while from $\nu_{it} = \nu_{i,t-1} + \zeta_{i,t-1}$, it follows that $\nu_{it} = \nu_{i,t-s} + \zeta_{i,t-1} + \ldots + \zeta_{i,t-s}$. Hence,

$$\Delta\Delta_s \mu_{it} = \Delta_s \nu_{i,t-1} + \Delta_s \xi_{i,t-1} = \zeta_{i,t-2} + \ldots + \zeta_{i,t-s-1} + \xi_{i,t-1} - \xi_{i,t-s-1},$$

where ξ_{it} and ζ_{it} are mutually independent Gaussian iid processes, and hence a finite-order moving average structure is obtained for the differenced trend component, with coefficients straightforwardly obtained from the s nonzero autocovariances.

A similar result is obtained for the seasonal component $\gamma_{it} = -\gamma_{i,t-1} - \ldots - \gamma_{i,t-s+1} + \omega_{i,t-1}$. Applying first differences to both sides of this equation yields $\gamma_{it} = \gamma_{i,t-s} + \omega_{i,t-1} - \omega_{i,t-2}$, and hence

$$\Delta\Delta_s \gamma_{it} = \Delta^2 \omega_{i,t-1},$$

which is again a (over-differenced) finite-order moving average that trivially has absolutely summable coefficients.

Regarding the cycle, Harvey (1991, Sec. 2.5.6) gives the stationary ARMA(2,1) representation for $|\rho_i| < 1$, which leads directly to

$$\Delta\Delta_s c_{it} = \Delta\Delta_s \frac{1 + \theta_i L}{1 - 2\rho_i \cos(\lambda_i) L - \rho_i^2 L^2} \tilde{\kappa}_{i,t-1},$$

where θ_i is a moving average parameter and $\tilde{\kappa}_{it}$ is composed of the two jointly Gaussian iid processes κ_{it} and κ_{it}^*. Since as a stationary ARMA process the fraction expands to a polynomial with absolutely summable coefficients, also the entire expression for $\Delta\Delta_s c_{it}$ shares this property while inheriting stationarity and Gaussianity. The same is true for differenced noise term $\Delta_4 \Delta u_{it}$. Hence, any linear combination of $\Delta\Delta_s \mu_{it}$, $\Delta\Delta_s \gamma_{it}$, $\Delta\Delta_s c_{it}$ and $\Delta\Delta_s u_{it}$ is strictly stationary, Gaussian and has absolutely summable coefficients. The statement is applicable both to the differenced idiosyncratic components e_{it} and to series of the differenced factor process F_t.

The properties of F_t and e_{it} are clearly sufficient to assure Assumptions A (by a law of large numbers drawing on ergodicity of F_t), C (since absolutely summable autocovariances follow from absolutely summable Wold coefficients), and D (due to the independence between e_{it} and F_t) of Bai and Ng (2002), while their Assumption B on the factor loadings has to be imposed additionally to obtain the main results of that article.

Clearly, the squared autocorrelations of e_{it} are also summablein our setup, and hence Bai and Ng (2002, eq. (6)) yields mean-square convergence of estimated F_t to the true values for a given t. Naturally, the consistency holds also for a cumulation of finitely many estimated F_s, $s \leq t$. Hence, also thefactors $\bar{f}_t$ in level are found consistent for a fixed t. The effects of the initial values are lost due to the differencing, however.

Appendix B

The State Space Form

The model given by (3) with measurement scheme (4) can be easily represented in linear state space form which allows to use the techniques described in Durbin and Koopman (2012). We adopt their notation as far as possible and state the system as

$$\begin{pmatrix} \bar{f}_t \\ z_t \end{pmatrix} = z_t \alpha_t + \begin{pmatrix} 0 \\ \varepsilon_t \end{pmatrix}, \quad \varepsilon_t \sim N(0, H_t), \tag{10}$$

$$\alpha_{t+1} = T\alpha_t + R\eta_t, \quad \eta_t \sim N(0, Q), \quad t = 1, \ldots, n. \tag{11}$$

For simplicity of exposition we assume that $l \geq s - 1$, so that l lags of all components have to be included in the state vector to make the measurement equation (4) representable in state space form. Hence, the state vector α_t holds the components μ_{it}^I, μ_{jt}^C, ν_{it}^I, ν_{jt}^C, γ_{it}^I, γ_{jt}^C, $(\tilde{c}_{it}^I, \tilde{c}_{it}^{I,*})$, $(\tilde{c}_{jt}^C, \tilde{c}_{jt}^{C,*})$, u_{it}^I and u_{jt}^C, each for $i = 1, \ldots, N_z$ and $j = 1, \ldots, r$, along with l lags of each component. More precisely,

$$\begin{aligned} \alpha_t' = (&\mu_{1t}^I, \ldots, \mu_{N_z,t}^I, && \mu_{1t}^C, \ldots, \mu_{r,t}^C, \\ &\nu_{1t}^I, \ldots, \nu_{N_z,t}^I, && \nu_{1t}^C, \ldots, \nu_{r,t}^C, \\ &\gamma_{1t}^I, \ldots, \gamma_{N_z,t}^I, && \gamma_{1t}^C, \ldots, \gamma_{r,t}^C, \\ &\tilde{c}_{1t}^I, \tilde{c}_{1t}^{I*}, \ldots, \tilde{c}_{N_z,t}^I, \tilde{c}_{N_z,t}^{I*}, && \tilde{c}_{1t}^C, \tilde{c}_{1t}^{C*}, \ldots, \tilde{c}_{r,t}^C, \tilde{c}_{r,t,}^{C*} \\ &u_{1t}^I, \ldots, u_{N_z,t}^I, && u_{1t}^C, \ldots, u_{r,t}^C, \\ &\mu_{1,t-1}^I, \mu_{2,t-1}^I, && \ldots \text{lagged components} \ldots, u_{r,t-l}^C)' \end{aligned}$$

is the $m := 6(N_z + r)(l + 1)$-dimensional state vector. Accordingly, the $6(N_z + r) \times (l + 1) \times 6(N_z + r)(l + 1)$ transition matrix is given by

$$T = \begin{pmatrix} \tilde{T} & 0 & \ldots & 0 \\ I & & & \vdots \\ & \ddots & & \vdots \\ 0 & & I & 0 \end{pmatrix}, \quad \text{where} \quad \tilde{T} = \begin{pmatrix} T_\mu & T_{\mu\nu} & 0 & 0 & 0 \\ 0 & T_\nu & 0 & 0 & 0 \\ 0 & 0 & T_\gamma & 0 & 0 \\ 0 & 0 & 0 & T_c & 0 \\ 0 & 0 & 0 & 0 & T_u \end{pmatrix}$$

is a $6(N_z + r) \times 6(N_z + r)$ matrix with $T_\mu = T_\nu = T_{\mu\nu} = I_{N_z+r}$. Moreover, $T_u = 0_{N_z+r}$ and T_c is a $2(N_z + r) \times 2(N_z + r)$ block diagonal matrix with ith block given by

$$T_c^{(i,i)} = \rho_i \begin{pmatrix} \cos\lambda_i & \sin\lambda_i \\ -\sin\lambda_i & \cos\lambda_i \end{pmatrix},$$

for $i = 1, \ldots, N_z + r$. Here, ρ_i and λ_i correspond to the individual cycle parameters for $i = 1, \ldots, N_z$, while they correspond to the parameters of the joint cycles, $\rho_i = \rho^C_{i-N_z}$ and $\lambda_i = \lambda^C_{i-N_z}$ for $i = N_z + 1, \ldots, N_z + r$. The transition innovation covariance matrix Q is block diagonal with block element given by Σ^I_ξ, Σ^C_ξ, Σ^I_ζ, Σ^C_ζ, Σ^I_ω, Σ^C_ω, $\Sigma^I_\kappa \otimes I_2$, $\Sigma^C_\kappa \otimes I_2$, Σ^I_u and Σ^C_u, respectively, while R is a vertical stacking of an identity and l quadratic zero matrices that selects the contemporaneous states.

The observation matrices Z_t reflect both the observation patterns for the variables and the loading of common components on the individual series. We denote

$$\tilde{\Lambda} = \begin{pmatrix} 0 & \Gamma_\mu & 0 & 0 & 0 & \Gamma_\gamma & 0 & \check{\Gamma}_c & 0 & \Gamma_u \\ I & \Lambda_u & 0 & 0 & I & \Lambda_\mu & \check{I} & \check{\Lambda}_c & I & \Lambda_\mu \end{pmatrix},$$

where the checked matrices reflect the phase shifts of the variables, so that the ith row of $\check{I}$ is $(\cos(\lambda_i\delta_i), \sin(\lambda_i\delta_i)) \otimes I_i.$, the ith row of $\check{\Gamma}_c$ is $(\cos(\lambda_i\delta_i), \sin(\lambda_i\delta_i)) \otimes \Gamma_{c,i}.$ and the ith row of $\check{\Lambda}_c$ is $(\cos(\lambda_i\delta_i), \sin(\lambda_i\delta_i)) \otimes \Lambda_{c,i}.$ which have twice the number of columns as the unchecked quantities. Then, for

$$\begin{aligned} \tilde{M}_t(L) &= \tilde{M}_{t0} + \tilde{M}_{t1}L + \ldots + M_{tl}L^l \\ &= \begin{pmatrix} I & 0 \\ 0 & M_{0t} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & M_{1t} \end{pmatrix} L + \ldots + \begin{pmatrix} 0 & 0 \\ 0 & M_{lt} \end{pmatrix} L^l \end{aligned}$$

the time-varying observation matrices are given by

$$Z_t = (\tilde{M}_{t0}\tilde{\Lambda}, \tilde{M}_{t1}\tilde{\Lambda}, \ldots, \tilde{M}_{tl}\tilde{\Lambda}),$$

which completes the state space representation for the general case with $d_t = 0$.

If constant terms or statistical breaks occur, the transition matrix is enriched by additional diagonal elements of 1, while the observation matrix reflects this by additional columns with corresponding element either set to the constant values, or switching from zero to that constant at a specified period. The state innovation error covariance matrix is unchanged and the matrix R holds additional rows of zeros.

7. References

Anderson, T.W. 1984. *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Bai, J. and S. Ng. 2002. "Determining the Number of Factors in Approximate Factor Models." *Econometrica* 70(1): 191–221. Doi: http://dx.doi.org/10.1111/1468-0262.00273.

Bai, J. and S. Ng. 2004. "A PANIC Attack on Unit Roots and Cointegration." *Econometrica* 72(4): 1127–1177. Available at: http://www.columbia.edu/~sn2294/pub/ecta04.pdf (accessed December 2017).

Bai, J. and S. Ng. 2008. "Large Dimensional Factor Analysis." *Foundations and Trends in Econometrics* 3(2): 89–163. Doi: http://dx.doi.org/10.1561/0800000002.

Bai, J. and S. Ng. 2013. "Principal Components Estimation and Identification of Static Factors." *Journal of Econometrics* 176(1): 18–29. Doi: https://doi.org/10.1016/j.jeconom.2013.03.007.

Banerjee, A., M. Marcellino, and I. Masten. 2014. "Forecasting with Factor-Augmented Error Correction Models." *International Journal of Forecasting* 30(3): 589–612. Doi: https://doi.org/10.1016/j.ijforecast.2013.01.009.

Bernanke, B.S., J. Boivin, and P. Eliasz. 2005. "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach." *The Quarterly Journal of Economics* 120(1): 387–422. Doi: https://doi.org/10.1162/0033553053327452.

Bollineni-Balabay, O., J. van den Brakel, and F. Palm. 2015. "Multivariate State Space Approach to Variance Reduction in Series with Level and Variance Breaks Due to Survey Redesigns." *Journal of the Royal Statistical Society:* Series A (Statistics in Society). Doi: http://dx.doi.org/10.1111/rssa.12117.

Bräuning, F. and S.J. Koopman. 2014. "Forecasting Macroeconomic Variables Using Collapsed Dynamic Factor Analysis." *International Journal of Forecasting* 30(3): 572–584. Doi: https://doi.org/10.1016/j.ijforecast.2013.03.004.

Breitung, J. and J. Tenhofen. 2011. "GLS Estimation of Dynamic Factor Models." *Journal of the American Statistical Association* 106(495): 1150–1166. Doi: https://doi.org/10.1198/jasa.2011.tm09693.

Burda, M.C. and J. Hunt. 2011. "What Explains the German Labor Market Miracle in the Great Recession?" *Brookings Papers on Economic Activity*, Spring 2011, 273–335. Available at: https://www.brookings.edu/wp-content/uploads/2011/03/2011a_bpea_burda.pdf (accessed December 2017).

Bureau of Economic Analysis. 2017. *Concepts and Methods of the U.S. National Income and Product Accounts*. Technical report. Available at: https://bea.gov/national/pdf/all-chapters.pdf (accessed February 2018).

Commandeur, J., S. Koopman, and M. Ooms. 2011. "Statistical Software for State Space Methods." *Journal of Statistical Software* 41(1): 1–18. Available at: https://www.researchgate.net/publication/260335110_Statistical_Software_for_State_Space_Methods (accessed December 2017).

Croushore, D. 2011. "Frontiers of Real-Time Data Analysis." *Journal of Economic Literature* 49(1): 72–100. Doi: http://dx.doi.org/10.1257/jel.49.1.72.

Durbin, J. 2000. "The State Space Approach to Time Series Analysis and its Potential for Official Statistics." *The Australian and New Zealand Journal of Statistics* 42(1): 1–23. Doi: http://dx.doi.org/10.1111/1467-842X.00104.

Durbin, J. and S.J. Koopman. 2012. *Time Series Analysis by State Space Methods: Second Edition*. Oxford: Oxford Statistical Science Series.

Durbin, J. and B. Quenneville. 1997. "Benchmarking by State Space Models." *International Statistical Review* 65(1): 23–48. Doi: http://dx.doi.org/10.1111/j.1751-5823.1997.tb00366.x.

Eickmeier, S. 2009. "Co-Movements and Heterogeneity in the Euro Area Analyzed in a Nonstationary Dynamic Factor Model." *Journal of Applied Econometrics* 24(6): 933–959. Doi: http://dx.doi.org/10.1002/jae.1068.

Federal Statistical Office. 2008. *National Accounts: Quarterly Calculations of Gross Domestic Product in accordance with ESA 1995 – Methods and Data Sources*. Subject-matter Series 18, S. 23. Available at: https://www.destatis.de/EN/Publications/Specialized/Nationalaccounts/QuarterlyCalculationsGrossDomesticProductAccordance.pdf?__blob=publicationFile (accessed December 2017).

Giannone, D., L. Reichlin, and D. Small. 2008. "Nowcasting: The Real-Time Informational Content of Macroeconomic Data." *Journal of Monetary Economics* 55(4): 665–676. Available at: https://EconPapers.repec.org/RePEc:eee:moneco:v:55:y:2008:i:4:p:665-676 (accessed December 2017).

Harvey, A. and C.-H. Chung. 2000. "Estimating the Underlying Change in Unemployment in the UK." *Journal of the Royal Statistical Society*: Series A (Statistics in Society) 163(3): 303–339. Doi: http://dx.doi.org/10.1111/1467-985X.00171.

Harvey, A. and S. Koopman. 1997. Multivariate Structural Time Series Models. Systematic Dynamics in Economic and Financial Models, (pp. 269–298). Available at: http://personal.vu.nl/s.j.koopman/old/publications/multi_stsm.pdf (accessed February 2018).

Harvey, A.C. 1991. *Forecasting, Structural Time Series Models and the Kalman Filter*. New York: Cambridge University Press.

Helske, J. 2016. *KFAS: Kalman Filter and Smoother for Exponential Family State Space Models*. R package version 1.2.4.

Koopman, S., A. Harvey, J. Doornik, and N. Shephard. 2009. *STAMP 8.2: Structural Time Series Analyser, Modeler, and Predictor*. London: Timberlake Consultants.

Krieg, S. and J.A. van den Brakel. 2012. "Estimation of the Monthly Unemployment Rate for Six Domains through Structural Time Series Modelling with Cointegrated Trends." *Computational Statistics and Data Analysis* 56(10): 2918–2933. Doi: https://doi.org/10.1016/j.csda.2012.02.008.

Moauro, F. and G. Savio. 2005. "Temporal Disaggregation Using Multivariate Structural Time Series Models." *The Econometrics Journal* 8(2): 214–234. Doi: http://dx.doi.org/10.1111/j.1368-423X.2005.00161.x.

Morley, J.C., C.R. Nelson, and E. Zivot. 2003. "Why Are the Beveridge-Nelson and Unobserved Components Decompositions of GDP so Different?" *Review of Economics and Statistics* 85(2): 235–243. Available at: http://www.jstor.org/stable/3211575 (accessed December 2017).

Ohanian, L.E. and A. Raffo. 2012. "Aggregate Hours Worked in OECD Countries: New Measurement and Implications for Business Cycles." *Journal of Monetary Economics* 59(1): 40–56. Doi: https://doi.org/10.1016/j.jmoneco.2011.11.005.

Pfeffermann, D. 1991. "Estimation and Seasonal Adjustment of Population Means Using Data from Repeated Surveys." *Journal of Business and Economic Statistics* 9(2): 163–175.

Pfeffermann, D. and R. Tiller. 2006. "Small-Area Estimation with State-Space Models Subject to Benchmark Constraints." *Journal of the American Statistical Association* 101(476): 1387–1397. Doi: https://doi.org/10.1198/016214506000000591.

Quenneville, B. and C. Gagné. 2013. "Testing Time Series Data Compatibility for Benchmarking." *International Journal of Forecasting* 29(4): 754–766. Doi: https://doi.org/10.1016/j.ijforecast.2011.10.001.

Rünstler, G. 2004. "Modelling Phase Shifts Among Stochastic Cycles." *Econometrics Journal* 7(1): 232–248. Doi: http://dx.doi.org/10.1111/j.1368-423X.2004.00129.x.

Stock, J.H. and M.W. Watson. 2002. "Macroeconomic Forecasting Using Diffusion Indexes." *Journal of Business and Economic Statistics* 20(2): 147–162. Doi: https://doi.org/10.1198/073500102317351921.

Stock, J.H. and M.W. Watson. 2005. "Implications of Dynamic Factor Models for VAR Analysis." NBER Working Paper 11467, NBER. Doi: http://dx.doi.or/10.3386/w11467.

Stock, J.H. and M.W. Watson. 2011. "Dynamic Factor Models." In *The Oxford Handbook of Economic Forecasting*, edited by M. Clements and D. Hendry, 35–60. Oxford: Oxford University Press.

Tiller, R.B. 1992. "Time Series Modeling of Sample Survey Data from the US Current Population Survey." *Journal of Official Statistics* 8(2): 149–166. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/time-series-modeling-of-sample-survey-data-from-the-u.s.-current-population-survey.pdf (accessed December 2017).

Valle e Azevedo, J., S.J. Koopman, and A. Rua. 2006. "Tracking the Business Cycle of the Euro Area: a Multivariate Model-Based Bandpass Filter." *Journal of Business and Economic Statistics* 24(3): 278–290. Available at: http://www.jstor.org/stable/27638878 (accessed December 2017).

Wanger, S. 2013. "Arbeitszeit und Arbeitsvolumen in Deutschland – Methodische Grundlagen und Ergebnisse der Arbeitszeitrechnung." *Wirtschafts- und Sozialstatistisches Archiv* 7(1-2): 31–69. Doi: https://doi.org/10.1007/s11943-013-0127-0.

Wanger, S., R. Weigand, and I. Zapf. 2016. "Measuring Hours Worked in Germany – Contents, Data and Methodological Essentials of the IAB Working Time Measurement Concept." *Journal for Labour Market Research* 49(3): 213–238. Doi: https://doi.org/10.1007/s12651-016-0206-0.

Wood, J. and D. Elliott. 2007. "Methods Explained: Forecasting." *Economic and Labour Market Review* 1(12): 55–58. Doi: https://doi.org/10.1057/palgrave.elmr.1410188.

Received August 2015
Revised December 2016
Accepted October 2017