



Journal of Official Statistics vol. 37, 2 (June 2021)

- Preface**.....p. 257-260
Francesca Di Iorio, Emanuele Baldacci, Dario Buono, Luca di Gennaro Splendore, Duncan Elliott, Rebecca Killick, Tiziana Laureti, Monica Pratesi and Natalie Shlomo
- A structural Equation Model for Measuring Relative Development of Hungarian Counties in the Years 1994–2016**.....p. 261-287
Klaudia Máténé Bella and Ildikó Ritzlné Kazimir
- Measuring and Communicating the Uncertainty in Official Economic Statistics**..... p. 289--316
Gian Luigi Mazzi, James Mitchell and Florabela Carausu
- The Evolution of the Italian Framework to Measure Well-Being**
..... p. 317–339
Fabio Bacchini, Barbara Baldazzi, Rita De Carli, Lorenzo Di Biagio, Miria Savioli, Maria Pia Sorvillo and Alessandra Tinto
- Improving Time Use Measurement with Personal Big Data Collection – The Experience of the European Big Data Hackathon 2019**..... p. 341-365
Mattia Zeni, Ivano Bison, Fernando Reis, Britta Gauckler and Fausto Giunchiglia
- A Diagnostic for Seasonality Based Upon Polynomial Roots of ARMA Models**.....p.367-394
Tucker McElroy
- Applying Machine Learning for Automatic Product Categorization**..... p. 395-410
Andrea Roberson
- A Product Match Adjusted R Squared Method for Defining Products with Transaction Data**.....p. 411-432
Antonio G. Chessa
- Variance Estimation after Mass Imputation Based on Combined Administrative and Survey Data**..... p. 433-459
Sander Scholtus and Jacco Daalmans
- Assessing and Adjusting Bias Due to Mixed-Mode in Aspect of Daily Life Survey**.....p. 461-480
Claudia de Vitiis, Alessio Guandalini, Francesca Inglese and Marco D. Terribili
- Measuring the Accuracy of Aggregates Computed from a Statistical Register**.....p. 481-503
Giorgio Alleva, Piero Demetrio Falorsi, Francesca Petrarca and Paolo Righi
- A Hybrid Technique for the Multiple Imputation of Survey Data**.....p. 505- 531
Humera Razzak and Christian Heumann

Preface

1. Introduction to the Special Issue on the New Techniques and Technologies

The fast technological progress society has been observing for several years has led to substantial changes in every aspect of work and daily life. Obviously, scientific research has also been strongly affected. One of the main aspects of this process is the availability of huge amounts of data suitable to be collected and analysed by both established and frontier methods. Many disciplines have benefited from this recent availability of data, sometimes leading to new fields of study, especially in the economic and social sciences. Studies on time-use exploiting data from mobile devices, social network analyses, consumer/user profiling, and new classification methodologies are just a few examples. Thanks to these advances, production of official statistics has already gained and is likely to gain more benefits in the future.

One of the immediate consequences of the development of these new technologies is the possibility to reduce the update delay of archives and official registers and to enhance their integration. National Statistical Offices (NSOs) – which have always used archives and registers for their surveys – have seized the opportunity offered by new technologies to make the best use of constantly updated information.

Techniques for data matching are developing rapidly, specifically in order to obtain variables and indicators based on more than one source. Subsequently, NSOs have been required to redesign their surveys to take into account new integration opportunities.

Some of the consequent advantages are obvious: the possibility to provide prompt official information, as in the case of flash forecasts; a reduced need for revision; an exploration of new data previously difficult to obtain; the introduction of new indicators; greater accuracy in data collection; and lower costs. Such innovations as disclosure issues, dissemination of new information, definition of new evaluation procedures and guidelines, and measurement and communication of uncertainty in official statistics have also entailed new prominent challenges.

The New Techniques and Technologies (NTTS) conferences, organised by Eurostat every three years from 1992 to 2001, and then every two years since 2003, have witnessed this evolution. The issues of Big Data in all related aspects, archive linking, and new source integration have been addressed in an increasing number of contributions, available in two NTTS special issues of *Journal of Official Statistics* (2015, 31:2; 2018, 34:4).

The program of the 2019 Conference, to which nearly 150 papers were submitted, fully reflects the extent of the pervasive entry of new technologies into official statistics. Twelve out of 32 parallel sessions refer to topics directly or indirectly related to big data and machine learning, acquisition and processing of mobile data, statistical matching, data integration, multisource statistics, administrative data, and registers.

Following the conference call for papers, the present JOS special issue embodies the positive outcome of NTTS 2019. The articles published in this volume well represent the great variety of topics and operational and methodological proposals discussed during the conference. These contributions are basically of two types: (1) articles that aim to present methodological approaches and results from several (European) projects, and (2) methodological proposals for official statistics production.

1.1. Dissemination

With regard to dissemination, the contribution by Máténé Bella and Ritzlné Kazimir, “*A Structural Equation Model for Measuring Relative Development of Hungarian Counties in the Years 1994–2016*”, presents an investigation on the economic development of the Hungarian regions at NUTS3 level. The model assumes that the development depends on observable economic variables, as well as on other non-economic, social, demographic, cultural, and infrastructural measures, evaluated based on regional indicators published by the Hungarian Central Statistical Office. Main findings constitute a starting point to investigate how the regional development has changed in Hungary after the regime change, and how these variables are influenced by the country’s integration into the global value chain.

The article by Mazzi, Mitchell and Carausu, “*Measuring and Communicating the Uncertainty in Official Economic Statistics*”, illustrates the results achieved so far in the context of the Eurostat Comunikos Project (COMmunicating UNcertainty In Key Official Statistics). The article reviews different categorisations of data uncertainty, distinguishes sampling from non-sampling errors, and introduces a novel typology. The importance of measuring and communicating data uncertainty is emphasised as it is difficult to measure some sources of uncertainty itself, especially those relevant for administrative and big data sets.

The study by Bacchini, Baldazzi, De Carli, Di Biagio, Savioli, Sorvillo and Tinto, “*The Evolution of the Italian Framework to Measure Well-Being*”, points out some of the outcomes of a joint project between the Italian Statistical Institute (Istat) and the National Council for Economics and Labor (CNEL) that aims to establish a multi-dimensional framework for measuring “equitable and sustainable well-being” (Bes). Since 2013, Istat has published an annual report on well-being, disseminating constantly updated Bes indicators, thereby allowing more effective communication. Bes indicators are taken into account in Italian government budget documents, as established by law. The Italian Bes experience may be considered to be one of the most relevant practices at the European level, and displays the potential to become a benchmark for other countries.

The article “*Improving Time Use Measurement with Personal Big Data Collection – The Experience of the European Big Data Hackathon 2019*”, by Zeni, Bison, Reis, Gauckler and Giunchiglia, describes the main contents of the Big Data satellite event of the NTTS Conference 2019. The article illustrates a pilot study on a system, called *i-Log*, which allows to collect personal big data from smartphones’ internal sensors to be used for time-use measurement. The *i-Log* system permits to gather heterogeneous types of data, thereby creating new possibilities for urban studies. Responses are used to train machine-learning algorithms, and allow the system to learn from the user’s habits and generate new so called “time-diary answers”. This new information could be used to assess the quality of existing answers, or to fill the gaps when the user does not provide any response.

1.2. Methodological Proposals

In “*A Diagnostic for Seasonality Based Upon Polynomial Roots of ARMA Models*”, McElroy presents a new methodological approach for seasonal adjustment. In particular, the study suggests connecting the concept of seasonality to a mathematical definition regarding the oscillatory character of the moving average (MA) representation coefficients, and defines a new seasonality diagnostic based on autoregressive (AR) roots. This procedure may help to assess dynamic versus stable forms of seasonality, and to deal with arbitrary seasonal periods, for both raw and seasonally adjusted data. Uncovering over-adjustment can be achieved using the proposed extension of the AR diagnostic to an MA diagnostic. The method is illustrated through simulation studies and empirical evidence.

The two articles by Roberson and by Chessa are both examples of how the new acquisition and data processing methods may be usefully applied in the context of official statistics.

“*Applying Machine Learning for Automatic Product Categorization*” by Roberson shows how machine-learning techniques have rapidly become one of the most used tools for classification. The U.S. Bureau of Census has developed a procedure based on the proposed methodology to address the challenges of high respondent burden and low survey response rates. The presented strategy automatically classifies goods and services – with an accuracy rate of over 90% – based on product information and text descriptions associated for each sold item to the Universal Product Code. Similarly, the same classification strategy may be applied to barcode (GTIN, Global Trade Item Number) product identification. GTINs is affected by the relaunch problem, when disappearing and re-introduced items have to be linked in order to capture possible price changes.

The study by Chessa, “*A Product Match Adjusted R Squared Method for Defining Products with Transaction Data*”, is focused on a method that groups GTINs into strata (‘products’). The method balances the need of “homogeneity” of GTINs within products with the “degree” to which products can be “matched” each month with respect to a base comparison period. The proposed technique combines explained variance in product prices with product match over time, so that different stratification schemes may be ranked according to the combined measure.

Two articles are concerned with the multiple imputation of missing data and new methods of administering questionnaires in social surveys.

Scholtus and Dalmaans, in “*Variance Estimation after Mass Imputation Based on Combined Administrative and Survey Data*”, present a simulation study on artificial data and an application to the Dutch census of 2011, in order to evaluate the variance of estimated frequency tables based on mass imputation, when data may be available from both administrative sources and a sample survey. The motivating application for this article is the Dutch virtual population census, for which it has been proposed to use mass imputation to estimate tables involving educational attainment. The article also discusses a more general bootstrap method for variance estimation.

In “*Assessing and Adjusting Bias Due to Mixed-Mode in Aspect of Daily Life Survey*”, De Vitiis, Guandalini, Inglese and Terribili verify the usefulness of a mixed mode (MM) approach, that is, the use of different collection techniques in the same survey, to both contrast declining response and coverage rates and reduce surveys cost. The mode effect

can be addressed at different levels: at the design phase the best collection instruments to contain the measurement error must be defined, and at the estimation phase the mode effect must be measured and corrected. The article presents the evaluation of the mode effect in the ISTAT mixed-mode survey “Aspects of Daily Life – 2017”, in an experimental context.

The article by Alleva, Falorsi, Petrarca and Righi “*Measuring the Accuracy of Aggregates Computed from a Statistical Register*” proposes the global mean squared error (GMSE) as a measure of accuracy in the production of official statistics, based on a massive integration of administrative archives and survey data. In particular, by focusing on the estimation of a population total from measurements calculated at the unit level, and considering the main sources of uncertainty that may affect registers, the authors suggest a feasible calculation strategy for the GMSE. The performance of the proposed strategy is verified through a simulation study.

“*A Hybrid Technique for the Multiple Imputation of Survey Data*” by Razzak and Heumann discusses imputation techniques in surveys in which categorical variables suffer from a large number of missing values, and complex dependencies may exist. Since the implementation of multiple imputation (MI) techniques or chained equations (MICE) may be problematic in such situations, the authors propose a method to make dependent on categorical variables, previously imputed using latent class models, chained MI equations for continuous variables. A simulation study and a survey data example articulately describe the good performance of the proposed method with respect to alternative techniques.

2. Acknowledgments

On behalf of the Editorship, we would like to warmly thank all the authors who have submitted their manuscripts for our consideration. In addition, we would like to thank all the referees for taking the time to review all the manuscripts. Over 50 referees have contributed to the current special issue and their comments and suggestions greatly helped to improve the original versions of the manuscripts.

As a matter of fact, the participants in the NTTS conference and the readership of this special issue may belong to different fields: they may be, in fact, either academics or official statistics producers. This circumstance determines the fruitfulness and the added value of the NTTS conference, also mirrored in the variety of themes hosted by Journal of Official Statistics, thus allowing the diffusion of new proposals and the presentation of specific experiences.

Francesca Di Iorio
Guest Editor

Emanuele Baldacci
Dario Buono
Luca di Gennaro Splendore
Duncan Elliott
Rebecca Killick
Tiziana Laureti
Monica Pratesi
Natalie Shlomo
Guest Associate Editors

A structural Equation Model for Measuring Relative Development of Hungarian Counties in the Years 1994–2016

Klaudia Máténé Bella¹ and Ildikó Ritzlné Kazimir¹

Relative development of Hungarian counties is described generally by the GDP per capita indicator, but this figure does not cover the knowledge gap on the liveability of the regions. The other frequently used method is the indicator systems, but it does not emphasize the structure of causes and consequences of the regional development, and so, it does not provide information on which factors are more likely to be the causes or, reversely, the consequences of the different regional development. To overcome the shortcomings of the above-mentioned methods, we created a structural equation model (SEM) at NUTS 3 level for years 1994–2016 based on the LISREL estimation procedure. The applied model can be classified into experimental statistics, but it uses data only from official statistics, namely the regional indicators published by the Hungarian Central Statistical Office. The model assumes that the economic development depends on observable economic indicators, and it determines the regional development as well. In addition, the regional development is also explained by non-economic, social, demographic and cultural and infrastructural indicators. The variable selection and the classification into causes and consequences was a three-step process, and the factors were classified by analysis of correlations, cross-correlations and Granger-causality. The results of estimation provided basis for a deeper analysis; how the regional development has changed in Hungary after the regime change, and how these variables were influenced by the country's integration into the global value chain.

Key words: Regional development; latent variable; relative development; NUTS 3; well-being.

1. Introduction

Relative development of Hungarian counties (at NUTS 3 level) is described generally by the GDP per capita indicator. We argue that the regional development is a more complex phenomenon because of its multidimensional nature. Comparing counties on the basis of GDP per capita is a very simple approach, because this figure does not cover the knowledge gap on the liveability of regions. In order to solve this problem, the indicator system is widely used to describe regional development.

In Hungary, Government Regulation 105/2015 (IV. 23) of “Classification of the beneficiary municipalities and the system of classification” defines, in its Annex 1, what kind of indicator groups and what kind of indicators within it should be used to calculate a

¹ Corvinus University of Budapest, Department of Statistics, Fővám tér 8, Budapest 1093, Hungary. Emails: matenebellaklaudia@gmail.com and ildiko.kazimir@gmail.com

Acknowledgments: The research was presented at the Conference on New Techniques and Technologies for Official Statistics (NTTS 2019) in Brussels, 12–14 March 2019. This publication/research has been supported by the European Union and Hungary and co-financed by the European Social Fund through the project EFOP-3.6.2-16-2017-00017, titled “Sustainable, intelligent and inclusive regional and city models”.

composite indicator, namely a development composite indicator for all municipalities. The calculated composite indicator that takes into account four sets of indicators (socio-demographic, housing and living conditions, local economy and labour market, infrastructure and environment) serves as a basis to decide which municipalities are entitled to development funds. Using this legal methodology or other economically reasonable indicators helps determine regional development of Hungarian counties.

Our point is that it is a useful methodology for analysing municipalities or small areas, but it requires a lot of detailed data, which are not available in the official statistics.

Most research projects focusing on regional development and well-being use micro data that are based on surveys and/or census. Unfortunately, these micro data are not available for every year.

Our goal is to find another methodology to define the regional development because statistical offices only publish regional data (indicators of the social and economic environment) at NUTS 2 or 3 level.

Furthermore, we argue that it is important to distinguish the factors that might contribute to the development of a region as causes from the factors that might be observed as consequences. Intuitively, a high activity ratio in the labour market and a high industry production of a region can be identified as causes of regional development, while other indicators such as the number of theatre performances or children enrolled in infant nurseries can be considered as consequences.

Our concept was to create a model based estimation on regional data at NUTS 3 level that is able to show the change of relative development of counties over time together with the causes and consequences. We wanted to answer the following questions: Does the ranking of counties according to the relative development differ from the ranking by GDP per capita? How has Hungary's integration into the global value chain influenced the development of counties from 1994 until 2016? How have the rankings changed as a result of subsidies from the European Union during the period reviewed? In which county (county seats) would people choose to live, according to the relative development?

In order to answer these questions, a structural equation model was constructed. Following [Pietrzak's \(2017\)](#) concept, we argue that the development of counties is a latent variable. A linear regression cannot be made because the dependent variable is unknown. However, a special factor analytic method can provide a solution to this problem. The unobserved dependent variable is influenced by determinants and in turn has an effect on the indicators. Using the LISREL estimation procedure (structural equation model), it is possible to quantify the relative development level of counties. This method is used often to calculate the hidden economy, but it has been already proved that this is a useful method to quantify other latent variables. We suggest that the complex analyses of causes and consequences leads to a reliable picture of the relative development of Hungarian counties between the years 1994–2016. Our model is based on official statistical data, but our methodology can be classified as experimental statistics.

2. Literature Review

The modelling of latent variables with causes and indicators is a commonly used method in psychometry ([Bollen 1989a](#)) or in the estimation of a hidden economy ([Frey 1984](#); [Frey and](#)

Pommerehne 1984; Leandro and Schneider 2018). The alternative methods for estimation of a latent variable can be classified into three groups. The original method for modelling of latent variables is improved by Kofler and Menges (1976) and called “soft modelling” where the distributions of variables are unknown. From “soft-modelling”, the factor analytic method emerged, the LISREL model (linear interdependent structural relationship) presented by Jöreskog (1969) that is a combined model of two previous models and the generalisations of the MIMIC (multiple indicators multiple causes) approach. The measurement model links the latent variable to observed indicators. The structural equation model specifies the causal relationships among the unobserved variables.

The idea that regional development as a latent variable can be analysed with structural equation modelling is not new. Cziráky et al. (2006) proposed a multivariate statistical approach to analyse regional development in Slovenia and Croatia. They suggested four latent variables (economic, structural, social and demographic) to capture regional development, and used several observed indicators at municipality level (see Table 1). The Croatian data came from the 2001 national census, while the Slovenian data had more sources. In their work, a structural equation model was constructed with the mentioned latent and observed variables. For each municipality, the latent variables were calculated,

Table 1. The model structure of Cziráky et al.

Latent development variables	Observed indicators	
	in Slovenia	in Croatia
Economic	Income per capita (in SIT)	Income per capita (in HRK)
	Social aid per capita (in thousands SIT)	Population share making income (%)
	Density (inhabitants per km ²)	Municipality income per capita (in thousands HRK)
	Number of cars per 100 inhabitants	Employment/population ratio
	–	Age index (65+ = (0 ; 20))
Structural	–	Education (share of high-school graduates in total population)
	Employment/population ratio	Employment/population ratio
	Social aid per capita (in thousands SIT)	Social aid per capita (in thousands HRK)
	Share of agricultural population	Share of agricultural population
Social	–	Population share making income (%)
	Share of agricultural population	Share of agricultural population
	Density (inhabitants per km ²)	Education (share of high-school graduates in total population)
Demographic	Students share per 1,000 inhabitants	Age index (65+ = (0 ; 20))
	Age index (65+ = (0 ; 14))	Population trend
	Population trend	Density (inhabitants per km ²)
	–	Vitality index (live births over number of deceased)

and according to the results, the municipalities were clustered to show the regional development at NUTS 2 level in both countries. They offered their model as a better alternative to the GDP/PPS rule.

Pietrzak (2017) applied a structural equation model to estimate regional development in Poland. He considered the regional development as a latent variable and used four observed variables such as gross domestic product per capita (PLN per inhabitant), investment outlays per capita (PLN per inhabitants), employment rate by age (%) and expenditure on R&D activity in relations to GDP (%). However, his model contained only one measurement part with a few variables and ignored the relationship between the regional development and the causes or consequences. In a later work, Pietrzak et al. (2017) showed a similar model to estimate the quality of the entrepreneurial environment at regional level in Poland. Twelve observed variables were used in this structural equation model and the quality of entrepreneurial environment at regional level was considered as a latent variable. The factors affecting the latent variables were categorised along the following dimensions: socio-economic development, social development and the effectiveness of the juridical system (see Table 2).

Cziráky et al. (2006) proposed the SEM with four latent variables at municipality level, but this is difficult to use because of the lack of such a detailed data set. Pietrzak (2017) defined one latent variable (regional development) based on data at NUTS 2 level.

The aim of our work was to build a SEM that is able to capture not only the regional development as a latent variable, but also the causes and consequences. Cziráky et al. (2006) mentioned that there was a relationship among latent variables, but this relationship

Table 2. Pietrzak's model structure.

Dimension	Sub-area	Observed indicators
Socio-economic development	Fundamental economic conditions	Gross domestic product per capita Investment outlays per capita
	Effectiveness of labour market	Employment rate by age Unemployment rate (according to Labour force survey)
Social development	Innovativeness	Expenditure on R&D activity in relation to GDP
	Demography	Ratio of balance of permanent migration person at working age
	Poverty and social exclusion	At-risk-of-poverty rate after social transfers People in households benefiting from the social assistance at domicile as percentage of the total population Average monthly available income per capita in private households.
	Education	Adult persons participating in education and training Children covered by pre-school education as percentage of the total number of children at the age 3–5
Effectiveness of juridical system	–	Rate of detectability of delinquents of ascertained crimes

was not directly analysed. Because data at municipality level are available only every five years based on census and micro-census, we had to choose a different approach to calculate the relative regional development of counties every year. Regional data at NUTS 3 and NUTS 2 level are published by the Hungarian Central Statistical Office every year, thus regional development as a latent variable could be calculated at the NUTS 3 level. In this case, the cluster analysis is unnecessary for the regional units.

3. Data

In the early 1990s there was a major social and economic recession in Hungary because of the regime change. It took a long time for the economy to adapt to the market economy environment. This was true for the statistical system as well. A new statistical framework developed continuously and it included – among others – the labour force survey and the ESA 95. The Hungarian Business Corporations Act was passed in 1990 and only after that year could corporations be formally established, and public enterprises transformed into the new legal forms. Statistical surveys on corporations were delayed for obvious reasons. The regional data between 1990 and 1993 are incomplete, or calculated according to another methodology, or not available at all. For these reasons, there is a consensus that the regional data between 1990 and 1993 are not suitable for the analysis, because of an inhomogeneity in the time series, especially in the case of variables relating to the labour market. The statistical system was stabilised in 1994, regional data became more reliable and is consistent with current methodology.

As a first step, we collected 51 annual variables at territorial level (for 19 counties and the capital, Budapest) for the period of 1994–2016. The source of data was the regional statistics tables (STADAT) of the Hungarian Central Statistical Office between 2000 and 2016 and the Regional Yearbooks between 1994 and ([Hungarian Central Statistical Office, 2019a, 2019b, 2019c, 2019d, 2019e, 2000, 1999, 1998, 1997, 1996, 1995, 1994](#)). Regional data covered such areas as demography, labour market, capital market, production and corporations, retail and tourism, infrastructure, education, culture and social benefits.

In the next step, variable selection was used. As regional development is a latent variable, it is impossible to use a Granger causality test ([Granger 1969](#)), a statistical hypotheses test for specifying whether it is worth using one of the time series to predict the other, explained in more detail below and to calculate correlations between regional development and other variables. However, if it is supposed that GDP per capita has a high correlation with regional development, the Granger causality test can be applied. Correlations can be calculated between variables and GDP per capita, but only in order to select variables for the model.

[Table 3](#) presents the highest correlation coefficients between variables and GDP per capita for the period 1994–2016 for each dimension. The GDP per capita is called x_5 in our model. (Variables in bold are the selected variables for the final model.)

We analysed the cross-correlations among variables, and selected variables for each dimension that have the highest correlations with most of the other variables. The [Table 4](#) presents the cross-correlation of selected variables. The average cross-correlations among the variables is 0.65; it is impossible to improve this value because of the lack of other relevant variables.

Once selected, the variables should be listed on the causes or on the consequences side. We considered that the long-run productivity is determined by the labour and capital

Table 3. The correlations between selected variables and GDP per capita.

Dimension	Variable	Correlation	t-statistic	p-value
Demography	<i>Population density</i>	0.57	14.92	0.00
	Population of cities with more than 100,000 people	0.55	14.37	0.00
Labour market	<i>Population</i>	0.48	11.75	0.00
	Average monthly net earnings	0.86	36.33	0.00
	Activity ratio	0.59	15.84	0.00
	<i>Employed</i>	0.59	15.81	0.00
Capital market	<i>Unemployment rate</i>	-0.31	-7.02	0.00
	Registered capital of corporations with foreign owners	0.74	23.30	0.00
	<i>Number of corporations with foreign owners</i>	0.61	16.65	0.00
Production/ economic activity	Production in industry	0.80	28.56	0.00
	Number of registered corporations	0.69	20.59	0.00
	<i>Investment per capita</i>	0.65	18.68	0.00
	Tourist arrivals (international) in accommodation establishment	0.65	18.54	0.00
Justice	Number of retail shops	0.51	12.85	0.00
	<i>Number of new dwellings</i>	0.30	6.91	0.00
Infrastructure	<i>Crime</i>	0.44	10.52	0.00
	Public water utility gap	0.83	32.94	0.00
Education	Cable TV subscriptions	0.71	22.11	0.00
	Number of students in universities	0.61	16.78	0.00
Health	<i>Hospital beds per 10,000 people</i>	0.10	2.30	0.02
Culture	Number of theatre performances	0.63	17.62	0.00
	Children enrolled in infant nurseries	0.59	15.85	0.00

Table 4. The cross-correlations between the variables.

	y1	x4	x2	y3	x3	y2	y9	y4	x1	y5	y8	y7	y6	x5
y1	1.00													
x4	0.18	1.00												
x2	0.23	0.51	1.00											
y3	0.74	0.44	0.38	1.00										
x3	0.44	0.64	0.56	0.65	1.00									
y2	0.89	0.36	0.34	0.91	0.61	1.00								
y9	0.92	0.28	0.34	0.89	0.50	0.93	1.00							
y4	0.90	0.15	0.22	0.77	0.48	0.93	0.86	1.00						
x1	0.45	0.78	0.59	0.47	0.74	0.53	0.45	0.41	1.00					
y5	0.88	0.40	0.38	0.89	0.65	0.97	0.91	0.92	0.57	1.00				
y8	0.87	0.31	0.23	0.76	0.58	0.93	0.82	0.95	0.54	0.93	1.00			
y7	0.94	0.27	0.31	0.86	0.49	0.92	0.99	0.87	0.45	0.91	0.83	1.00		
y6	0.96	0.23	0.32	0.79	0.51	0.93	0.92	0.95	0.49	0.92	0.91	0.93	1.00	
x5	0.56	0.86	0.60	0.74	0.80	0.69	0.65	0.51	0.84	0.72	0.62	0.64	0.60	1.00

inputs. Variables relating to the labour market (activity ratio, average monthly net earnings) and to the capital market (production in industry) are observed and considered as causal variables. Variables relating to the labour market (activity ratio, unemployment rate) and to the capital market (FDI, production in industry) are observed – and according to the Granger causality test can be considered – as causal variables.

The Granger causality test is a statistical hypothesis test for specifying whether one time series is worth using to forecast another. This test was first proposed by Granger in 1969 (Granger 1969).

Generally, we say that z Granger-causes y if

$$E(y_t|I_{t-1}) \neq E(y_t|J_{t-1}) \quad (1)$$

where I_{t-1} contains past information on y and z , and J_{t-1} contains information on past y . When (1) holds, past values of z are useful, in addition to past values of y , for predicting y_t (Wooldrige 2014). Once we assume a linear model and decide how many lags of y should be included in $E(y_t|y_{t-1}, y_{t-2})$, we can test the null hypothesis that z does not Granger-cause y . To be more specific, suppose that y and z are stationary time series. To test the null hypothesis, two steps are necessary. First, the proper lagged values of variable y should be found for univariate autoregression of y .

$$y_t = a_0 + a_1y_{t-1} + a_2y_{t-2} + \dots + a_my_{t-m} + error_t \quad (2)$$

Then, the autoregression model of y is extended by including lagged values of z , when

$$y_t = a_0 + a_1y_{t-1} + a_2y_{t-2} + \dots + a_my_{t-m} + b_1z_{t-q} + \dots + b_qz_{t-q} + error_t \quad (3)$$

All lagged values of z that are individually significant according to their t -statistics are retained in regression when they collectively add explanatory power to the regression according to an F -test (whose null hypothesis is no explanatory power jointly added by the z 's). We would remark that (according to the [User's Guide of Eviews 11 \(2019\)](#), Eviews runs bivariate regressions of the form:

$$y_t = a_0 + a_1y_{t-1} + a_2y_{t-2} + \dots + a_qy_{t-q} + b_1z_{t-1} + \dots + b_qz_{t-q} + error_t \quad (4)$$

$$z_t = a_0 + a_1z_{t-1} + a_2z_{t-2} + \dots + a_qz_{t-q} + b_1y_{t-1} + \dots + b_qy_{t-q} + error_t \quad (5)$$

for all possible pairs of (z, y) series in the group. The reported F -statistics are the Wald-statistics for the joint hypothesis: $b_1 = b_2 = \dots = b_q = 0$ for each equation. The null hypothesis is that z does not Granger-cause y in the first regression and that y does not Granger-cause z in the second regression.

A Granger causality test was performed on the dataset of years 2000–2016, because in the 1990s there was a regime change in the labour market. From 1994 until 1999, the FDI was not measured in Hungary at regional level, thus these data could not be used. The registered capital of foreign-owned corporations is considered as a consequence in this period. The economic transformation was a lengthy process that allowed a gradual inflow of capital.

All of the selected variables are classified into the causes or the consequences side of the model. One example of causes and consequences is shown in [Figure 1](#). The public water utility gap is a good proxy for the infrastructure needed by industry. (The public water utility gap is a drinking water and wastewater infrastructure gap that is the difference

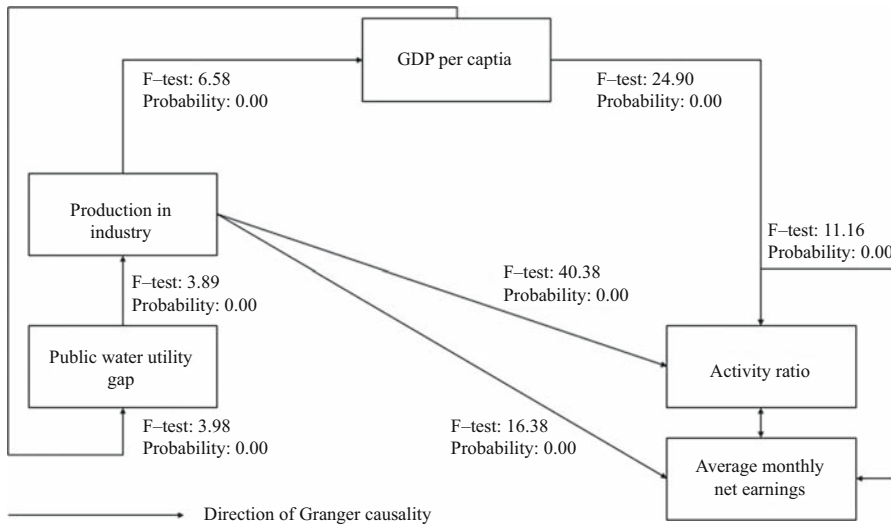


Fig. 1. An example of the structure of causes and consequences between the variables.

between the proportion of the population connected to the public water supply and the population connected to the sewerage network.) Therefore, the public water utility gap is a significant significant indicator of industrial production in Hungary during the selected period. Similarly, production in industry can be a cause of GDP per capita, which can be explained by the relatively high share of industry in the production of goods and services.

The relation between GDP per capita and activity ratio is also interesting. In Hungary there is a trade-off between the activity ratio and employment opportunities. If the job prospects are not positive, job seekers can opt for inactivity in the labour market. Besides this process, the activity ratio influences the average monthly net earnings through the development of the labour market. In addition, the relationship is controversial, because the change in average net earnings induces crossing between active and inactive population.

According to the results of Table 4 and Figure 1 we selected and grouped the variables to causes and consequences side presented in the Table 5.

In the period 1994–1999, the unemployment rate was high in Hungary and privatisation had been completed. The capital inflow from abroad became significant during and after the privatisation period and the capital market was dominated by the foreign capital. This caused a structural change in industrial production. In this context, registered capital of foreign-owned corporations was classified as a variable on the consequences side, while production in industry influenced by the FDI was considered as an observed variable on the causes side.

Of course, the classification is not always clearly defined, as it can be changed from period to period, but indicators such as the existence of big cities, the public water utility gap, the number of retail shops, the number of tourist arrivals in accommodation establishments, the number of students in universities and the number of theatre performances were rather considered consequences of regional development. All variables were standardised before modelling, using SPSS (SPSS Tutorials). Standardised data are linearly transformed data with a mean of zero and a standard deviation of one.

Table 5. The variables on the causes and consequences sides.

Dimensions	Sub-dimensions	Observed indicators	
		Causes side	Consequences side
Production	Labour market	Activity ratio x_2 Average monthly net earnings x_4	–
	Capital market	FDI (2000–2016) Production in industry (1994–2016) x_3 GDP per capita x_5	–
Demographic dimension	Population	–	Population of cities with more than 100,000 people y_1
Economic dimension	Corporations	–	Number of registered corporations y_2 Registered capital of foreign owned corporations (1994–2016) y_3
	Retail and tourism	–	Number of retail shops y_4 Tourist arrivals (international) in accommodation establishment y_9
Structural dimension	Infrastructure	Public water utility gap x_1	Number of cable TV subscriptions y_5
Social dimension	Education	–	Number of students in universities y_8
	Culture	–	Number of theatre performances y_7
	Social	–	Children enrolled in infant nurseries (proxy of female employment) y_6

4. Methods

We constructed a structural equation model to quantify the regional development as a latent variable. We used Jöreskog's concept, as mentioned in the Literature Review (Section 2). For this model, LISREL or IBM AMOS software can be applied. We decided to use IBM AMOS (Analysis of Moment Structure) because with this software a model can be quickly specified, viewed and modified graphically using simple drawing tools. After computations, the program displays the results by the path graph.

4.1. The General Structural Equation Model (SEM)

For the modelling of regional development, Structural Equation Model (SEM) was chosen. The SEM includes two components: the structural model presents the causal dependencies between the latent variables, and the measurement model specifies how the observed variables depend on the unobserved, latent variables. The theoretical description of the SEM model follows the concept of Jöreskog et al. (2016, 344–345).

The next figure shows an example of a general SEM model (see Figure 2).

In Figure 2, there are five non-observed variables, $\xi = (\xi_1, \xi_2, \xi_3)$ and $\eta = (\eta_1, \eta_2)$. There are seven observed explanatory variables ($x = (x_1, x_2, x_3, x_4, x_5, x_6, x_7)$) for the ξ , and four explanatory variables ($y = (y_1, y_2, y_3, y_4)$) for the η . The relationship between the above variables is drawn up with the arrows, and the parameters are on the arrows as well. Some of the parameters are fixed by unit value. For the observed variables the model assigns also error terms $\delta = (\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7)$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$. The $\phi_{21}, \phi_{31}, \phi_{32}$ are the element of covariance matrix between the elements of ξ . The ψ_{21} is the covariance between η_1 and η_2 . This structure is pretty difficult.

The general framework of SEM is described by the following formal model:

$$\eta = \alpha + B\eta + \Gamma\xi + \zeta \tag{6}$$

Equation (6) includes the linear structural relationship between the vectors of the latent dependent ($\eta = (\eta_1, \eta_2, \dots, \eta_m)'$) and the latent independent ($\xi = (\xi_1, \xi_2, \dots, \xi_n)'$) variables. The vector α is the intercept, the matrices B and Γ includes the coefficients and $\zeta = (\zeta_1, \zeta_2, \dots, \zeta_m)'$ is the vector of residuals. It is assumed that the ζ is uncorrelated with ξ , and that the matrix $(I - B)$ is non-singular.

The latent variables (η and ξ) are non-observed, but the variables $x = (x_1, x_2, \dots, x_p)'$ and $y = (y_1, y_2, \dots, y_q)'$ are observed, that such

$$y = \tau_y + \Lambda_y\eta + \varepsilon \tag{7}$$

$$x = \tau_x + \Lambda_x\xi + \delta \tag{8}$$

Where ε and δ are the error term vectors, assumed to be uncorrelated with η and ξ respectively. The vectors τ_y and τ_x are the intercepts in these multivariate regressions, the Λ_y and Λ_x show the coefficient matrices.

The mean vector μ and the covariance matrix Σ of $z = (x', y)'$ can be expressed by the next formulas:

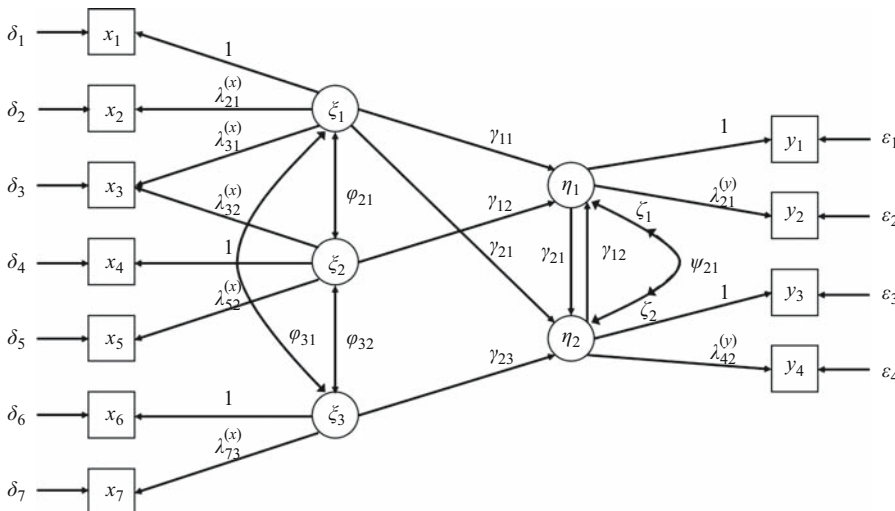


Fig. 2. Path diagram of a general LISREL model (Jöreskog et al. 2016, 344).

$$\mu = \begin{pmatrix} \tau_y + \Lambda_y(I - B)^1(\alpha + \Gamma\kappa) \\ \tau_x + \Lambda_x\kappa \end{pmatrix} \quad (9)$$

$$\Sigma = \begin{pmatrix} \Lambda_y(I - B)^1(\Gamma\Phi\Gamma' + \Psi)(I - B)^1\Lambda_y' + \Theta_\varepsilon & \Lambda_y(I - B)^1\Gamma\Phi\Lambda_x' + \Theta'_{\delta\varepsilon} \\ \Lambda_x\Phi\Gamma'(I - B)^1\Lambda_y' + \Theta_{\delta\varepsilon} & \Lambda_x\Phi\Lambda_x' + \Theta_\delta \end{pmatrix} \quad (10)$$

Where the vector κ is the mean vector of ξ , Φ and Ψ are the covariance matrices of ξ and ζ . The matrices Θ_ε , Θ_δ are the covariance matrices of error terms ε , δ . The $\Theta_{\delta\varepsilon}$ mean is the covariance matrix between ε , and δ . The vector μ and matrix Σ are the functions of elements: κ , α , τ_y , τ_x , Λ_y , Λ_x , B , Γ , Φ , Ψ , Θ_ε , Θ_δ , $\Theta_{\delta\varepsilon}$. These elements can be divided into three kinds:

1. Fixed parameters that have specific values
2. Constrained parameters that are linear or non-linear functions of one or more other variables.
3. Free parameters

The latent variables do not have definite scale, because they are unobservable. The LISREL method has two ways to scale them:

1. If it is a fixed non-zero coefficient between the latent variable and a reference variable, then this reference variable defines the scale for that latent variable.
2. If there is no reference variable for the latent variable with a fixed non-zero coefficient, then the LISREL method standardises the latent variable.

4.2. Specification of the Model

The current model for the period 1994–2016 has two distinct measurement sub-models: The available labour force is observable in the activity ratio and in net monthly earnings. We believed that industrial production is a good indicator of the success of a private company capable of exporting and contributing to the overall production of the region (see [Figure 3](#)).

The regional development can be measured with nine indicators (see [Figure 4](#)), which include the existence of big cities (the population of cities with more than 100,000 people), the number of registered corporations, the registered capital of foreign-owned corporations (production), the number of cable TV subscriptions (infrastructure), the number of children enrolled in infant nurseries (proxy of female employment and the social care system), the number of retail shops and the number of tourist arrivals in accommodation establishments (international) (tourist attractions, beautiful environment, productions), the number of students in universities according to parents' residence (education), the number of theatre performances (culture).

In our model, the structural part consists of two latent variables: the production influences the regional development (see [Figure 5](#)).

The whole model is shown in [Figure 6](#).

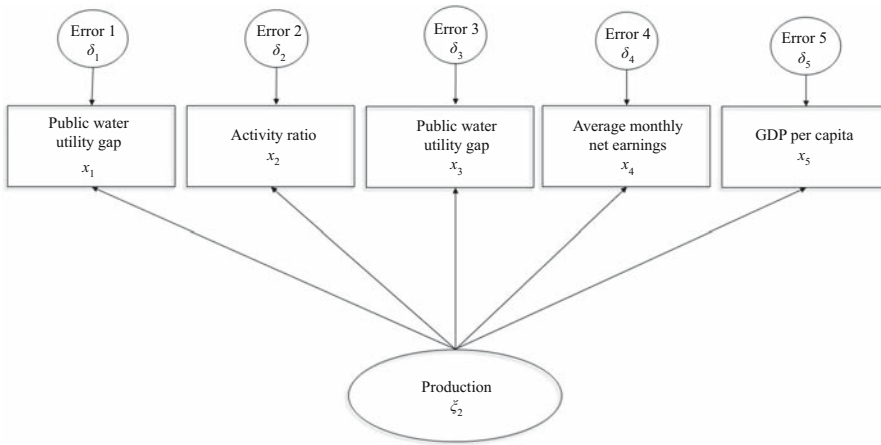


Fig. 3. The side of economic circumstances.

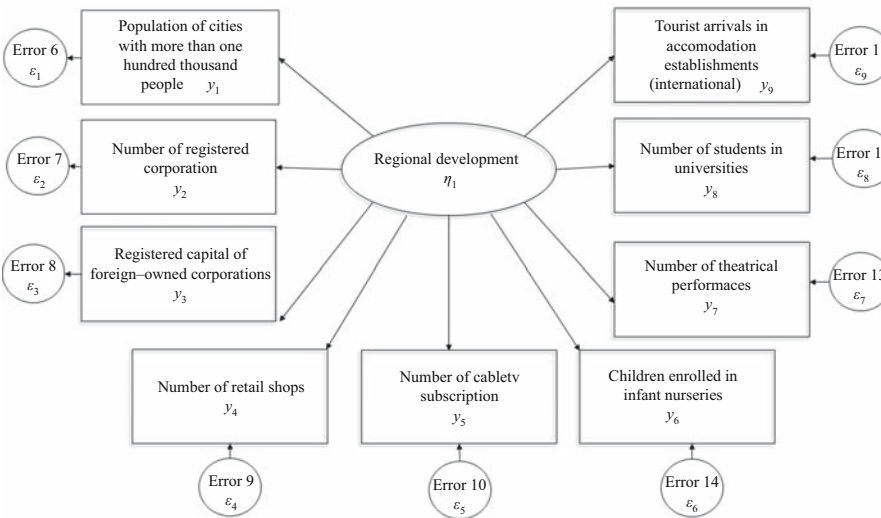


Fig. 4. The side of regional development.

Our model is much simpler than the model in Figure 2. It has only two latent variables. The measurement parts of the model include the fourteen above-tested variables. The linear relationship between the latent variables as follows:

$$\eta_1 = \gamma_{11} \xi_1 + \zeta_1 \tag{11}$$

where η_1 refers to the dependent latent variable, namely regional development, and the latent independent variable is ξ_1 , the production. ζ_1 is the residual.

The measurement model equations for y -variables are

$$y_1 = \lambda_{11}^{(y)} \eta_1 + \epsilon_1 \tag{12}$$

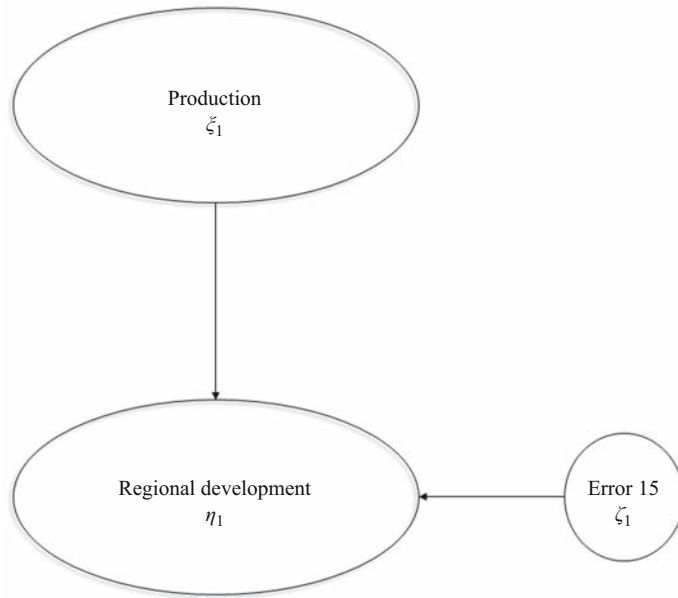


Fig. 5. The relationship between the latent variables.

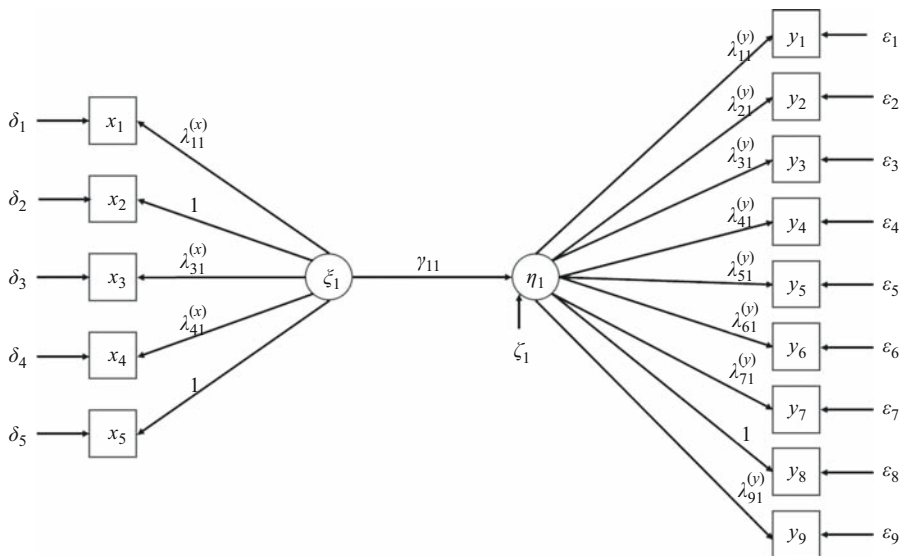


Fig. 6. The structure of the model for the estimation of Hungarian development.

$$y_2 = \lambda_{21}^{(y)} \eta_1 + \varepsilon_2 \tag{13}$$

$$y_3 = \lambda_{31}^{(y)} \eta_1 + \varepsilon_3 \tag{14}$$

$$y_4 = \lambda_{41}^{(y)} \eta_1 + \varepsilon_4 \tag{15}$$

$$y_5 = \lambda_{51}^{(y)} \eta_1 + \varepsilon_5 \tag{16}$$

$$y_6 = \lambda_{61}^{(y)} \eta_1 + \varepsilon_6 \tag{17}$$

$$y_7 = \lambda_{71}^{(y)} \eta_1 + \varepsilon_7 \tag{18}$$

$$y_8 = \eta_1 + \varepsilon_8 \tag{19}$$

$$y_9 = \lambda_{91}^{(y)} \eta_1 + \varepsilon_9 \tag{20}$$

Where $\lambda_{i1}^{(y)}$ and ε_i are the coefficient and error in equation i, respectively. The measurement model equations for x-variables are

$$x_1 = \lambda_{11}^{(x)} \xi_1 + \delta_1 \tag{21}$$

$$x_2 = \xi_1 + \delta_2 \tag{22}$$

$$x_3 = \lambda_{31}^{(x)} \xi_1 + \delta_3 \tag{23}$$

$$x_4 = \lambda_{41}^{(x)} \xi_1 + \delta_4 \tag{24}$$

$$x_5 = \xi_1 + \delta_5 \tag{25}$$

For the estimation, the important parameters are the parameters ϕ_{11} and ψ_{11} which are the variances of ξ_1 and ζ_1 , respectively. The covariance matrices of vectors ε and δ are the diagonal matrices

$$\Theta_\varepsilon = \text{diag}(\theta_{11}^{(\varepsilon)}, \theta_{22}^{(\varepsilon)}, \dots, \theta_{99}^{(\varepsilon)})$$

$$\Theta_\delta = \text{diag}(\theta_{11}^{(\delta)}, \theta_{22}^{(\delta)}, \dots, \theta_{55}^{(\delta)})$$

With 17 unobserved variables (two latent variables and 15 errors) in this model, it is certainly not identifiable. It is necessary to fix the unit of measurement of each unobserved variable by suitable constraints on the parameters.

4.3. Estimation of the Model

Parameter estimation is done by comparing the actual covariance matrices representing the relationships between variables and the estimated covariance matrices of the best fitting model (Table 6).

The calculation of degrees of freedom is necessary for identifying the model, and for computing Chi-square. The Chi-square is detailed in Subsubsection 4.4.2.

The estimated model for the period 1994–2016 is presented graphically as a path model in Figure 7. The latent variables are illustrated in the ellipse and the circle, while the observed variables are shown in the rectangle. On the causes side, production influences regional development significantly through the observed variables such as activity ratio, average net monthly earnings, public water utility gap, production in industry and GDP per capita.

Table 6. *The free parameters of the model.*

Computation of degrees of freedom (Default model)	
Number of distinct sample moments	105
Number of distinct parameters to be estimated	28
Degrees of freedom (105-28)	77
Result (Default model)	
Minimum was achieved	
Chi-square	4110.012
Degrees of freedom (105-28)	77
p-value	0.000

The measurement model presents the relationship between the latent variable (regional development) and its indicators.

4.4. *Testing the Model*

Testing the model is one of the most discussed issues in structural equation models. Three situations can be distinguished according to [Jöreskog et al. \(2016, 495–502\)](#):

- **Strictly confirmatory (SC) situation:** The researcher has specified one single model and has obtained empirical data to test it. The model should be accepted or rejected.
- **Alternative models or competing models (AM) situation:** The researcher has formulated several alternative models, and based on the data, one of the models should be accepted.
- **Model generating (MG) situation:** The researcher has specified an initiative model. If this model does not fit the given data, the model should be modified and tested. The re-specification of each model may be theory-driven and/or data-driven. The goal may be to find a model which not only fits the data well from a statistical point of view, but also has the property that every parameter of the model can be given a meaningful interpretation ([Jöreskog et al. 2016](#)).

Our research relates to the Model-Generating situation, because we have formulated and tested several models in order to find a model that fits the data well and has a meaningful economic interpretation. The evaluation of a model and the assessment of model fit is an important tool for deciding whether the specified model is accepted or rejected. The output from the structural equation program provides information that is useful for model evaluation and assessment of fit. This information can be classified into three groups:

1. Examination of the solution
2. Measures of overall fit
3. Detailed assessment of fit

4.4.1. Examination of the Parameters to Explore Unreasonable Values

The unstandardised regression weights (parameters λ) are significant in all three equations according to the p-value.

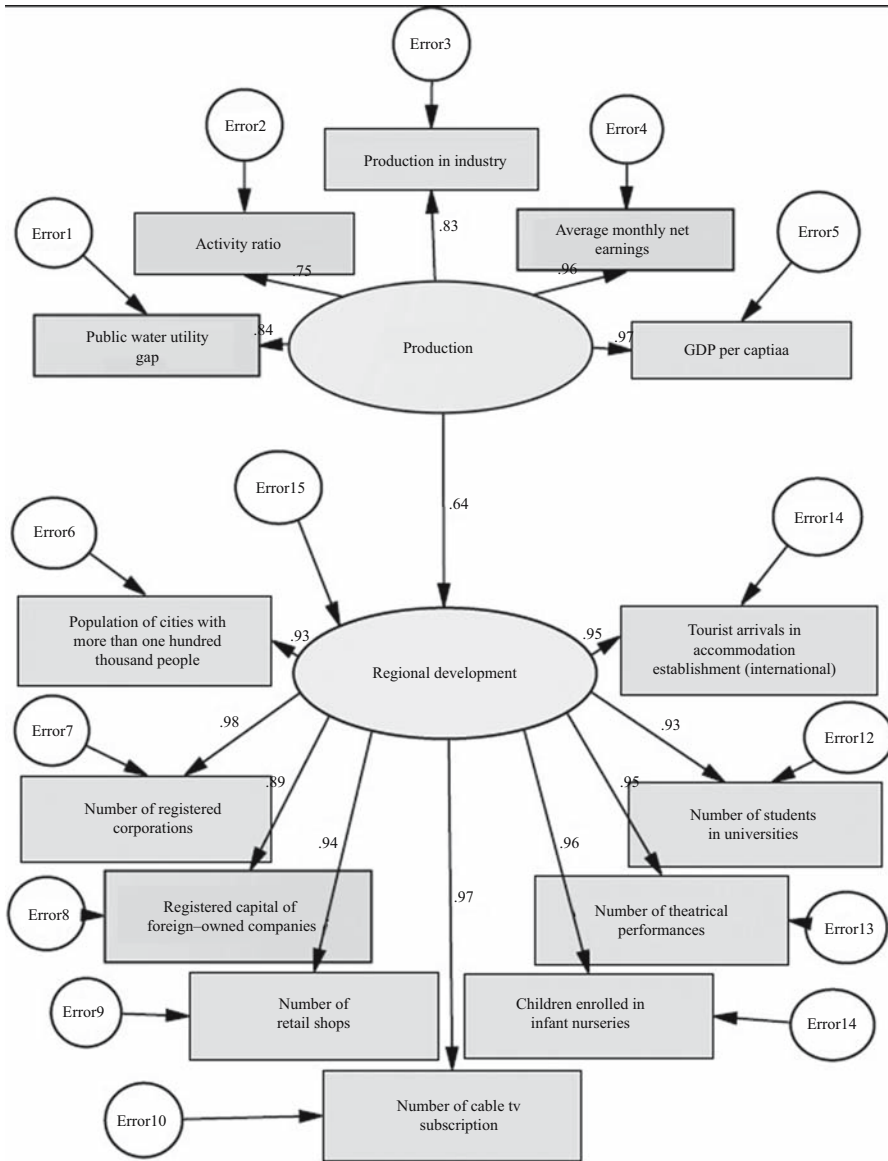


Fig. 7. The path diagram of the estimated model.

4.4.2. Examination of the Measures to Demonstrate Overall Fit of the Model

Several measures are available to test whether the constructed model represents a good fit. We can classify these measures in the following two groups:

1. Chi-square, relative chi-square
2. Fit indices
 - Incremental or relative fit index
 - Absolute fit index
 - Comparative fit index

1. Chi-square is considered a fundamental measure of overall fit of the model to the data. It is a function of the sample size and the difference between the observed covariance matrix and the model covariance matrix. Chi-square is a badness-of-fit measure in the sense that a small chi-square corresponds to a good fit and a large chi-square corresponds to a bad fit. Zero chi-square corresponds to a perfect fit. It is a reasonable measure when the sample size is between 75 and 200, but in more than 400 cases the chi-square is always quite statistically significant.

An old measure of fit is the relative chi-square, namely the chi-square to degree of freedom ratio or χ^2/df proposed by Wheaton et al. (1977). A problem with this fit index is that there is no universally agreed upon standard as to what is a good and a bad fitting model.

2. Fit indices

An **incremental or relative fit index** is analogous to R^2 and so a value of zero indicates having the worst possible model and a value of one indicates having the best possible model. In terms of a formula, it is

$$\frac{\text{Worst possible model} - \text{Our model}}{\text{Worst possible model} - \text{Fit of the Best Possible Model}} \quad (26)$$

The worst possible model is called the *null* (or *independence model* in Amos) and the usual convention is to allow all the variables in the model to have variation but no correlation. (The usual null model is to allow the means to equal their actual value. The degrees of freedom of the null model are $k(k - 1)/2$ where k is the number of variables in the model.) An example of incremental measure of fit is the *Bentler-Bonett Index* or the *Normed Fit Index (NFI)* which was proposed in the literature (Bentler and Bonett 1980). The best model is defined as a model with a χ^2 of zero and the worst model is defined by the χ^2 of the null model. Its formula is from Jöreskog et al. (2016, 501):

$$NFI = \frac{\chi^2(\text{Null Model}) - \chi^2(\text{Proposed Model})}{\chi^2(\text{Null Model})} \quad (27)$$

A value between .90 and .95 is considered marginal, above .95 is good, and below .90 is considered to be a poor fitting model.

A problem with the Bentler-Bonett Index is that there is no penalty for adding parameters. The TLI, *Tucker-Lewis Index* (Bollen (1989b, 2), also called the non-normed fit index or NNFI), another incremental fit index, does have such a penalty. The TLI is computed as follows:

$$TLI = \frac{\chi^2/df(\text{Null Model}) - \chi^2/df(\text{Proposed Model})}{\chi^2/df(\text{Null Model}) - 1} \quad (28)$$

Note that the TLI (and the CFI which follows) depends on the average size of the correlations in the data. If the average correlation between variables is not high, then the TLI will not be very high.

Another incremental measure is the *Comparative Fit Index (CFI)*, which is directly based on the non-centrality measure. Let $d = \chi^2 - df$ where df are the degrees of freedom

of the model. According to Bentler (1990), the *Comparative Fit Index* or CFI equals

$$CFI = 1 - \frac{d(\text{Null Model}) - d(\text{Proposed Model})}{d(\text{Null Model})} \quad (29)$$

Absolute fit index: An absolute measure of fit presumes that the best fitting model has a fit of zero. The measure of absolute fit determines how far the model is from perfect fit. These measures of fit are typically a “badness” measure of fit in that the bigger the index, the worse the fit is.

Root Mean Square Error of Approximation (RMSEA) is an absolute measure of fit based on the non-centrality parameter. Its computational formula is (Schwarz 1978; Raftery 1993):

$$RMSEA = \sqrt{\frac{2 - df}{df(N - 1)}} \quad (30)$$

where N is the sample size and df is the degrees of freedom of the model. If χ^2 is less than df , then the RMSEA is set to zero. Like the TLI, its penalty for complexity is the chi-square to df ratio. The measure is positively biased (i.e., it tends to be too large) and the amount of the bias depends on the smallness of the sample size and df .

Comparative fit index: A comparative measure of fit is only interpretable when comparing two different models. Two examples of this type of measure are the *Akaike Information Criterion (AIC)* and the *Bayesian Information Criterion (BIC)*. Lower values of AIC indicate a better fit and so the model with the lowest AIC is the best fitting model. AIC can be calculated in the following way (Akaike 1973, 1984):

$$AIC = \chi^2 + k(k + 1) - 2df \quad (31)$$

where k is the number of variables in the model and df is the degrees of freedom of the model. The AIC makes the researcher pay a penalty of two for every parameter that is estimated. The BIC increases the penalty as sample size increases, its formula is (Schwarz 1978; Raftery 1993):

$$BIC = \chi^2 + \ln(N) \left[\frac{k(k + 1)}{2} - df \right] \quad (32)$$

where $\ln(N)$ is the natural logarithm of the number of cases in the sample.

After the review of the different measures of overall fit of the model, we present the calculated values of selected indices in Table 7.

Based on the Chi-square test, our model is statistically significant (see the p-value in Table 6). This measure does not give us information whether the model is a “good fitting” model, therefore more indices, namely TLI, CFI and RMSEA are necessary.

The TLI (and CFI) value heavily depends on the average size of the correlations in the data. If the average correlation among variables is not high, then the TLI (and CFI) will not be very high. A TLI (and CFI) value of .95 or higher is desirable (Hu and Bentler 1999). TLI is 0.65 and CFI is 0.705 in our model. These values indicate a “moderately strong model”. It is easy to see that if the fit is medium then the error is not low. The value of zero

Table 7. The unstandardised parameters and their *p*-values.

Equation	Independent variable	Dependent variable	Estimated parameter λ	Standard error SE	Critical ratio for regression weight $CR = \lambda/SE$	<i>p</i> -value
(6)	ξ_1 Production	η_1 Regional development	0.892	0.063	14.183	0
(7)	η_1 Regional development	y_1 Population of cities with more than 100,000 people	1.002	0.026	39.053	0
		y_2 Number of registered corporations	1.054	0.021	49.325	0
		y_3 Registered capital of foreign-owned corporations	0.951	0.029	32.792	0
		y_4 Number of retail shops	1.011	0.025	40.577	0
		y_5 Number of cable TV subscriptions	1.044	0.022	46.843	0
		y_6 Children enrolled in infant nurseries	1.033	0.023	44.447	0
		y_7 Number of theatre performances	1.021	0.024	42.235	0
		y_8 Number of students in universities	1			
		y_9 Tourist arrivals in accommodation establishments (international)	1.022	0.024	42.477	0
(8)	ξ_1 Production	x_1 Public water utility gap	0.888	0.03	29.518	0
		x_2 Activity ratio	1			
		x_3 Production in industry	0.879	0.031	28.722	0
		x_4 Average monthly net earnings	0.905	0.029	31.185	0
		x_5 GDP per capita	1			

of RMSEA indicates the best fit. In our model, this value is 0.338. Hu and Bentler (1999) suggested that 0.08 or a smaller value is a sign of a good fit.

Why do the fit measures suggest a moderate fit? According to [Table 4](#), the average correlation among variables is 0.65, thus using these variables it is impossible to construct a model with better TLI (and CFI) value. We are convinced that all possible official statistical variables available at NUTS 3 level with the highest cross-correlation have been discovered and analysed in our model. We tried to include additional variables in the model, but they had lower cross-correlations, so while the RMSEA improved, the CFI and TLI tended to decline slightly. We came to the conclusion that we have found the model that is statistically significant and has an acceptable fitting on empirical data.

5. Results

The run of the constructed SEM resulted in the factor score weights listed in [Table 9](#).

Using these factor score weights, the standardised value of regional development was calculated for each county and for all years. Then on the basis of these calculated values, the ranking of Hungarian counties was constructed for every year.

In 1994, the regional development of counties reflected the effects of the industrial centres before the regime change. As [Figure 8](#) shows, after the capital, Budapest, the second most developed county was Borsod-Abaúj-Zemplén with the county seat Miskolc. This city was the centre of heavy industry before 1989.

The Hungarian economy integrated into the global value chain rapidly from 2000, due to increasing foreign direct investment by multinational corporations. The manufacturing industry has been completely transformed. The presence of multinational corporations had a impact not only on GDP per capita, but through the local business tax they contributed to the development and investment of local government as well. Later, the increasing industrial concentration also became a self-generating process. The change of relative development of counties resulted in changes in income and so also, in the final demand of households. Nevertheless, the regional distribution of multinational corporations was concentrated and the counties in which big corporations did not settle were lagging behind in development. There was no significant rearrangement between 1994 and 2016 (see [Figure 9](#)).

The southwest of the country is lagging behind in development in relative terms. Counties that could restructure themselves after the regime change, could achieve a better position

Table 8. Measurement of model fit.

Type of fit index	Name of fit index	Calculated value for our model	Optimal target value
Chi-square (p-value:0.000)	Chi-square	4110	
Chi-square	–		
	Chi-square/degree of freedom	4110/77= 55.38	–
Incremental fit index	TLI	0.652	1
Incremental fit index	CFI	0.705	1
Absolute fit index	RMSEA	0.338	0

Table 9. The model factor scores.

Variables	Production	Regional development
Production in industry	0.098	0.001
Average monthly net earnings	0.117	0.001
Number of cable TV subscriptions	0.007	0.163
Number of students in universities	0.003	0.063
Number of retail shops	0.003	0.075
Number of theatre performances	0.004	0.09
Tourist arrivals in accommodation establishments (international)	0.004	0.092
Children enrolled in infant nurseries	0.005	0.117
Public water utility gap	0.104	0.001
Population of cities with more than 100,000 people	0.003	0.065
Number of registered corporations	0.011	0.254
GDP per capita	0.601	0.006
Activity ratio	0.051	0.001
Registered capital of foreign-owned corporations	0.002	0.037

and could gain benefits from the new opportunities after joining the EU in 2004. We highlight the county Győr-Moson-Sopron – this county realised the benefits of Austria's proximity and good infrastructure. Baranya county in the south part of Hungary, on the other hand lost its position, despite the fact that its capital is one of the major university cities.

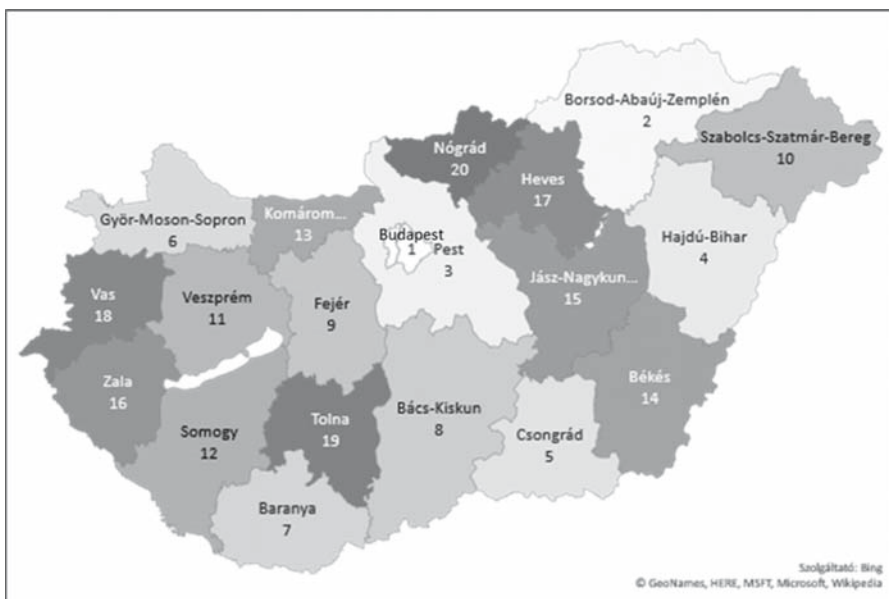


Fig. 8. Ranking of Hungarian counties in 1994 according to regional development based on SEM.

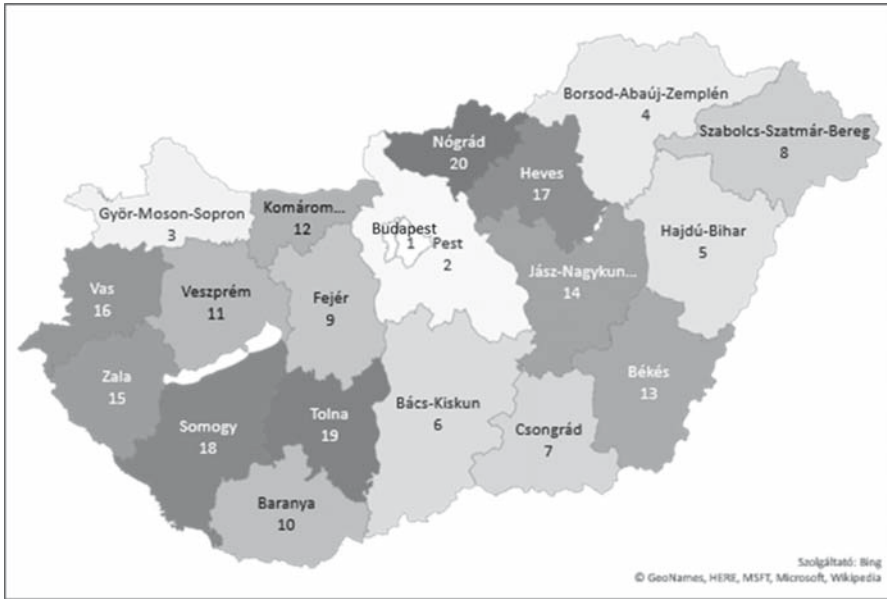


Fig. 9. Ranking of Hungarian counties in 2016 according to regional development based on SEM.

The ranking of the Hungarian counties according to the regional development as a latent variable shows another picture than the ranking of counties according to GDP per capita. The most obvious difference is in the judgement of the position of county Pest (agglomeration of the capital Budapest). According to the latent variable, county Pest was third in the ranking in 1994 and second in 2016, but it was fourteenth/sixth based on GDP

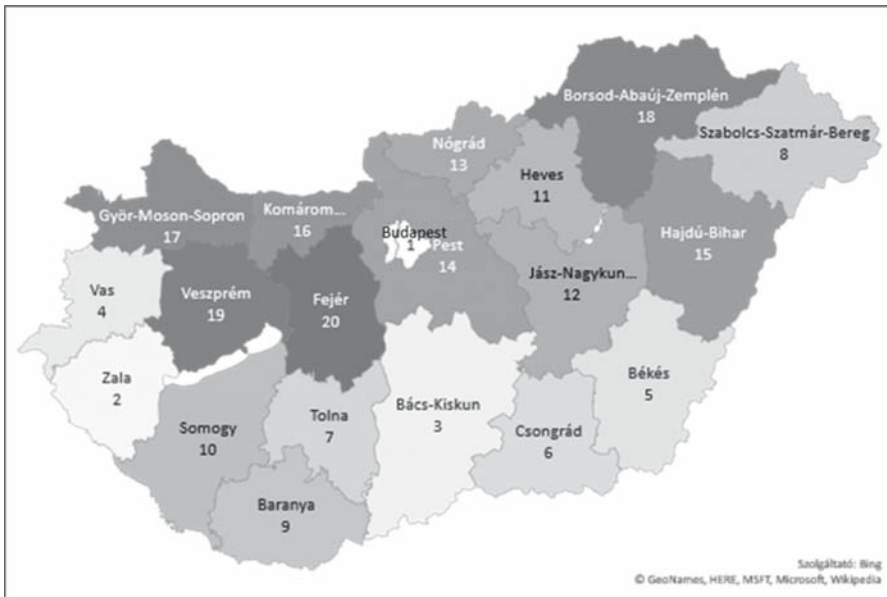


Fig. 10. Ranking of Hungarian counties in 1994 according to GDP per capita.

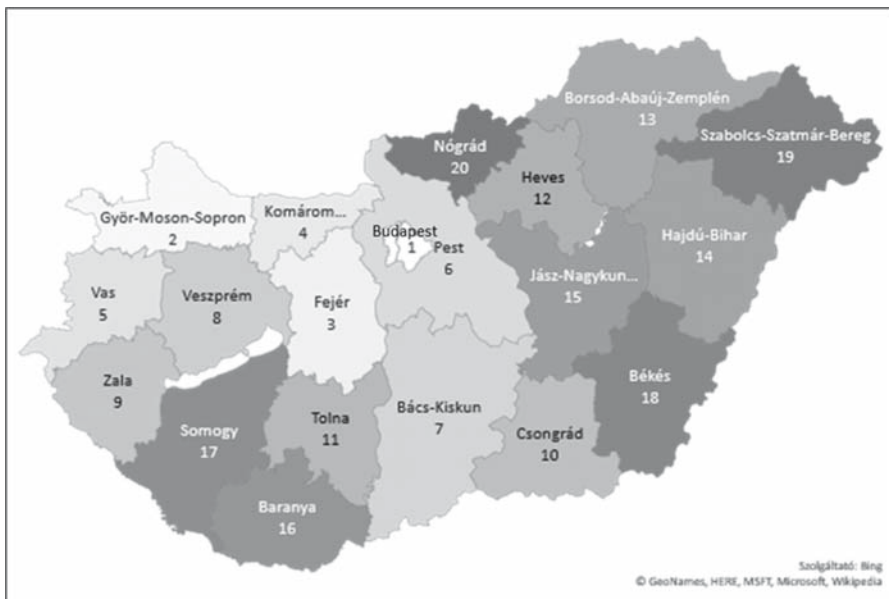


Fig. 11. Ranking of Hungarian counties in 2016 according to GDP per capita.

per capita (see [Figure 10](#) and [Figure 11](#)). The northeast/ southwest parts of Hungary lost their relative position from 1994 to 2016, while the middle of the country and the northwest counties gained in relative position of GDP per capita. The second difference is related to the eastern counties, Borsod-Abaúj-Zemplén, Szabolcs-Szatmár-Bereg, Békés and Hajdú-Bihar. All of them are ranked behind in GDP per capita ranking, while according to the latent variable they belong to the more developed counties. The difference is clearly due to the fact that GDP per capita only takes the production value and population into account, while the SEM uses several other variables. We argue that the ranking based on the latent variable gives a more realistic answer to the question of where it is better to live. The liveability of a county or a county seat is influenced by many factors. We think our model is capable of compressing these factors into one variable while using only official statistics.

Since joining the European Union in 2004, Hungary has received support from EU funds, which contributed to the development of the counties. It is an interesting question whether this support had any impact on the development of counties and their relative ranking. Due to the lack of regional data on EU subsidies, it is impossible to make a model, but according to the latent variable for regional development, the ranking of counties seems to be quite permanent. Several counties could not improve their position. Reasons for this situation could be the lack of export-oriented domestic and multinational corporations and also the lack of a well-educated labour force.

6. Conclusions

We argue that our model is appropriate for estimating the relative development of Hungarian counties during the examined period of 1994–2016. We defined the causes and

the consequences of regional development in Hungary during this period. The measurement and structural models are specified in IBM AMOS software.

We conclude that the export-oriented, competitive multinational companies selected regions where the input supply was more favourable. This process became a self-generating process, and the gap between developed and less developed regions deepened.

In terms of GDP per capita, the order of counties is significantly different from the ranking based on the latent variable. For this reason, it would be worth considering a more complex set of criteria when designing a support system for the regions, not just focusing on the gross value added.

If there is a big city with at least 100,000 people in the county, then it stands out in cultural, educational, economical, and other opportunities. Therefore, big cities dominate the county's ranking. The liveability of the county depends on the urbanity of the county centre, but it should not be forgotten that meanwhile some small areas within the county may fall off. For this reason, the county cannot be considered homogeneous. Although a district level analysis would be more appropriate, small area data at this level of detail is not available for every year. Therefore, this analysis cannot handle the heterogeneity of counties and reflects the county seat's liveability.

In addition, it is shown that the modelling of the latent variable is appropriate for the estimation of relative change of counties' economic development. GDP per capita only measures the economic performance, while regional development as a latent variable takes into account several important variables from the point of view of well-being as well.

Theoretically, google trends data or Wikipedia data (non-official statistical data) can improve our model, and will be the task of a new project.

7. References

- Bentler, P.M. 1990. "Comparative fit indexes in structural models." *Psychological Bulletin*, 107(2): 238–246. DOI: <http://dx.doi.org/10.1037/0033-2909.107.2.238>.
- Bentler, P.M., and D.G. Bonett. 1980. "Significance tests and goodness of fit in the analysis of covariance structures." *Psychological Bulletin* 88: 588–606. DOI: <https://doi.org/10.1037/0033-2909.88.3.588>.
- Bollen, K.A. 1989a. *Structural Equations with Latent Variables*. New York: Wiley. DOI: <https://doi.org/10.1002/9781118619179>.
- Bollen, K.A. 1989b. "New incremental fit index for general structural equation models." *Sociological Methods and Research* 17: 303–316. DOI: <https://doi.org/10.1177/0049124189017003004>.
- Czirák, D., J. Sambt, J. Rován, and J. Puljiz. 2006. "Regional development assessment: A structural equation approach." *European Journal of Operational Research* 174, Issue 1: 427–442. DOI: <https://doi.org/10.1016/j.ejor.2005.03.012>.
- Frey, B., and W. Pommerehne. 1984. "The hidden economy: State and prospects for measurement." *Review of Income and Wealth* 30: 1–. DOI: <https://doi.org/10.1111/j.1475-4991.1984.tb00474.x>.
- Frey, B.S., and H. Weck-Hanemann. 1984. "The Hidden Economy as an 'Unobserved' variable." *European Economic Review* 26: 33–53. DOI: [https://doi.org/10.1016/0014-2921\(84\)90020-5](https://doi.org/10.1016/0014-2921(84)90020-5).

- Government Regulation 105/2015. (IV. 23) of “Classification of the beneficiary municipalities and the system of classification.” Available at: [https://net.jogtar.hu/-jogszabaly?docid = a1500105.kor](https://net.jogtar.hu/-jogszabaly?docid=a1500105.kor) (accessed May 2019).
- Granger, C.W.J. 1969. “Investigating Causal Relations by Econometric Models and Cross-spectral Methods.” *Econometrica* 37 (3): 424–438. DOI: [doi:10.2307/1912791](https://doi.org/10.2307/1912791). JSTOR 1912791.
- Hungarian Central Statistical Office. 2019a. 6.1. *Population, vital events*. Available at: http://www.ksh.hu/stadat_annual_6_1 (accessed May 2019).
- Hungarian Central Statistical Office. 2019b. 6.2. *Society*. Available at: http://www.ksh.hu/stadat_annual_6_2 (accessed May 2019).
- Hungarian Central Statistical Office. 2019c. 6.3. *General Economic Indicators*. Available at: http://www.ksh.hu/stadat_annual_6_3 (accessed May 2019).
- Hungarian Central Statistical Office. 2019d. 6.4. *Economic branches*. Available at: http://www.ksh.hu/stadat_annual_6_4 (accessed May 2019).
- Hungarian Central Statistical Office. 2019e. 6.5. *Environment*. Available at: http://www.ksh.hu/stadat_annual_6_5 (accessed May 2019).
- Hungarian Central Statistical Office. 2000. *Regional Statistical Yearbook of Hungary, 1999*.
- Hungarian Central Statistical Office. 1999. *Regional Statistical Yearbook of Hungary, 1998*.
- Hungarian Central Statistical Office. 1998. *Regional Statistical Yearbook of Hungary, 1997*.
- Hungarian Central Statistical Office. 1997. *Regional Statistical Yearbook of Hungary, 1996*.
- Hungarian Central Statistical Office. 1996. *Regional Statistical Yearbook of Hungary, 1995*.
- Hungarian Central Statistical Office. 1995. *Regional Statistical Yearbook of Hungary, 1994*.
- Hungarian Central Statistical Office. 1994. *Regional Statistical Yearbook of Hungary, 1993*.
- Hu, L.T., and P.M. Bentler. 1999. “Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives.” *Structural Equation Modeling: A Multidisciplinary Journal* 6: 1–55. DOI: <https://doi.org/10.1080/10705519909540118>.
- Jöreskog, K. 1969. “A general approach to confirmatory maximum likelihood factor analysis.” *Psychometrika* 183–202. DOI: <https://doi.org/10.1007/BF02289343>.
- Jöreskog, K., U. Olsson, and F. Wallentin. 2016. *Multivariate Analysis with LISREL*. Springer International Publishing Switzerland. DOI: <https://doi.org/10.1007/978-3-319-33153-9>.
- Kofler, E., and G. Menges. 1976. *Entscheidungen bei unvollständiger Information*. Berlin Heidelberg New York: Springer-Verlag.
- Leandro, M., and F. Schneider. 2018. “Shadow Economies Around the World: What Did We Learn Over the Last 20 Years?” IMF Working Paper. Available at: <https://www.imf.org/en/Publications/WP/Issues/2018/01/25/Shadow-Economies->

- [Around-the-World-What-Did-We-Learn-Over-the-Last-20-Years-45583](#) (accessed May 2019).
- Pietrzak, M.B. 2017. "Structural equation modelling of regional economic development in polis voivodeships in the years 2010–2014." The 11th International Days of Statistics and Economics, Prague, September 14–16, 2017. Available at: https://msed.vse.cz/msed_2017/article/182-Pietrzak-MichalBernard-paper.pdf (accessed May 2019).
- Pietrzak, M.B., A.P. Balcerzak, A., Gajdos, and L. Arendt. 2017. "Entrepreneurial environment at regional level: the case of Polish path towards sustainable socio-economic development. Entrepreneurship and Sustainability Issues." Entrepreneurship and Sustainability Center. 2017, 5 (2), pp. 190–203. Available at: <http://jssidoi.org/jesi/article/143>, <https://hal.archives-ouvertes.fr/hal-01703293> (accessed May 2019).
- Raftery, A.E. 1993. "Bayesian model selection in structural equation models." In *Testing structural equation models*, edited by K.A. Bollen and J.S. Long, 163–180. Newbury Park, CA: Sage Publications.
- Schwarz, G. 1978. "Estimating the dimension of a model." *The Annals of Statistics* 6: 461–464. Available at: https://projecteuclid.org/download/pdf_1/euclid.aos/1176344136 (accessed May 2019).
- SPSS Tutorials. 2019. Available at: <https://www.spss-tutorials.com/z-scores-what-and-why/> (accessed May 2019).
- User's Guide of Eviews 11. 2019. Available at: http://www.eviews.com/help/helpintro.html#page/content%2Fgroups-Granger_Causality.html%23 (accessed May 2019).
- Wheaton, B., B. Muthén, D.F. Alwin, and G.F. Summers. 1977. "Assessing reliability and stability in panel models". *Sociological methodology*, edited by D.R. Heise, 84–136. San Francisco: Jossey-Bass.
- Wooldrige, J.M. 2014. Introduction to Econometric. *Europe, Middle East and Africa Edition*. Andover: Cengage Learning.

Received June 2019

Revised October 2019

Accepted June 2020

Measuring and Communicating the Uncertainty in Official Economic Statistics

Gian Luigi Mazzi¹, James Mitchell², and Florabela Carausu³

Official economic statistics are uncertain even if not always interpreted or treated as such. From a historical perspective, this article reviews different categorisations of data uncertainty, specifically the traditional typology that distinguishes sampling from nonsampling errors and a newer typology of Manski (2015). Throughout, the importance of measuring and communicating these uncertainties is emphasised, as hard as it can prove to measure some sources of data uncertainty, especially those relevant to administrative and big data sets. Accordingly, this article both seeks to encourage further work into the measurement and communication of data uncertainty in general and to introduce the *Comunikos* (COMmunicating UNcertainty In Key Official Statistics) project at Eurostat. *Comunikos* is designed to evaluate alternative ways of measuring and communicating data uncertainty specifically in contexts relevant to official economic statistics.

Key words: Measurement error; data revisions; official economic statistics; data communication.

1. Introduction

Official economic statistics are inevitably uncertain or, put another way, subject to “errors”, even if not always interpreted or treated as such. Data uncertainty can affect the economic historian’s view of the past and policymaker’s decisions in the present.

Statistical (or measurement) “error” is commonly defined as the difference between the estimate produced by the statistical office and the “true” population value, which is typically unobserved. As Eurostat (2015) states: “Accuracy refers to the closeness of computations or estimates to the exact or true values that the statistics were intended to measure”. But, as Groves and Lyberg (2010) discuss, the notion of a *true* value is itself subject to debate. Accordingly, to define, yet alone measure, “uncertainty” remains elusive. For a general discussion of uncertainty as a concept see Van der Bles et al. (2019). Nevertheless, the idea that the “error” is the difference between the estimated and actual value (even if this itself is unobserved) is at least widely agreed; for example see Dungan et al. (2002) and Fuller (1987).

¹Freelance consultant. Email: glmazzi@pt.lu

²Federal Reserve Bank of Cleveland, 1455 E 6th St, Cleveland, Ohio, 44114, USA. Email: James.Mitchell@clev.frb.org

³GOPA Luxembourg, Rue Luxembourg, Bereldange, 7240, Luxembourg. Email: florabela.carausu@gopa.lu

Acknowledgments: Thanks to the Editor (Francesca Di Iorio), an Associate Editor, two referees, Dario Buono, George Kapetanios, Edwin de Jonge and Massimiliano Marcellino for helpful comments. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Cleveland or the Federal Reserve System.

Accordingly, statisticians and statistical offices have sought to categorise and communicate data uncertainties in various ways. This reflects a long history, dating back at least to [Kuznets \(1948\)](#) and [Morgenstern \(1950\)](#), that emphasises the uncertainty of economic statistics. For example, Principle 12 of Eurostat's European Code of Practice states that "Sampling errors and non-sampling errors are measured and systematically documented according to the European standards" (see [Eurostat 2017](#)); and also "Revisions are regularly analysed in order to improve statistical processes". Despite this, arguably, statistical errors have received limited attention in the seventy years since Kuznet's discussion of 'margins of errors' and Morgenstern's comprehensive catalogue of errors that affect economic data; for example see [Bagus \(2011\)](#). Certainly, as [Manski \(2015\)](#), [Manski \(2018\)](#) and [Van der Bles et al. \(2019\)](#) emphasise, headline statistical estimates tend to be presented as point estimates, arguably conveying a misleading degree of reliability in these data.

This lack of communication of economic data uncertainty is common across national statistical offices and, in turn, in the media when they disseminate statistical office data. Over more recent years, following the encouragements of Manski and others, several statistical authorities and organisations have started investing in identifying ways to measure and communicate data uncertainty; these include: the use of fan charts at the Bank of England and the Riksbank to communicate historical GDP data uncertainties; work by CBS Netherlands on 'Visualising uncertainty' and on the inventarisation of uncertainty sources; and UK Government Statistical Service guidance on 'Communicating Uncertainty and Change' (see [ONS 2018a](#)).

Measuring uncertainty is a complex and challenging task that can involve the use of sophisticated statistical and econometric techniques (classical or Bayesian) and subjective judgement to quantify the data uncertainties. However, as challenging as the quantification of data uncertainties *per se*, is how to communicate them – ideally in a way that is both 'comprehensive', in terms of capturing fully the uncertainties, but also 'understandable' so that different users and readers of these data correctly infer and interpret the uncertainties communicated to them.

Accordingly, in late 2018 Eurostat launched the *Comunikos* (COMmunicating UNcertainty In Key Official Statistics) project. *Comunikos* explores and analyses tools for measuring and communicating data uncertainties. The aim is to formulate proposals and make recommendations for the most appropriate ways to measure and communicate uncertainties for official statistics. As the risk of misleading or indeed confusing users is arguably high, but providing clear and 'accurate' uncertainty measures may enhance the relevance and credibility of official statistics, *Comunikos* investigates the *pros* and *cons* of communicating uncertainties to users of official statistics. In particular, it shows the potential for state-space models to measure uncertainty and produce confidence intervals or densities for official statistics, mainly for intra-annual time-series data ([Kapetanios et al. 2021](#)); analyses methods to measure uncertainty in cross-sectional data focusing on poverty measures ([Laureti and Benedetti 2020](#)) and considers verbal, quantitative and visual tools to communicate uncertainty in official statistics reports ([De Jonge 2020](#)). The aim of *Comunikos* is that by providing additional uncertainty information, users of official statistics will be able to make better decisions, in particular at times of heightened data uncertainty that we might expect to occur precisely (e.g. at business cycle 'turning points') when users are most interested in the data.

In this article, to help further the *Comunikos* project agenda and more generally encourage work measuring and communicating data uncertainty, we provide a methodological review and categorisation of uncertainty measures and their sources for economic statistics. We focus on quantitative economic data. Despite this qualification, the discussion below also has relevance for qualitative, ordinal data, although we refer to [Piccolo and Simone \(2019\)](#) for a more specific discussion of uncertainties for ordinal data. We exploit the fact that statisticians commonly categorise uncertainties to reflect non-sampling and sampling errors. Non-sampling errors apply to administrative records and surveys, including censuses, whereas sampling errors apply only to sample surveys. In principle, therefore, the *total* uncertainty associated with statistical output comprises both sampling error and non-sampling error. Though, in practice, the measurement of the total survey error is difficult, given the complexity of estimating and quantifying both sampling and, in particular, non-sampling errors. As [Boumans \(2012\)](#) discusses, while sympathetic to Morgenstern's call for the use of errors to accompany economic statistics, [Kuznets \(1950\)](#) emphasised the challenges implied by [Morgenstern \(1950\)](#) for measurement. [Kuznets \(1950\)](#) argued that economic statistics are better thought of as the products of evolving institutions, rather than making analogies, as Morgenstern did, with scientific data from controlled experiments. [Groves and Lyberg \(2010\)](#) discuss the conceptual history of total survey error over more than seventy years. The total survey error of an estimate is considered as an indicator of data quality.

As there are various ways in which these *errors* can be classified, a recent distinction and the proposed typology of [Manski \(2015\)](#) is also introduced and discussed. This is helpful in explaining how the different types of uncertainty can be communicated quantitatively – and ideally they then comprise part of the *statistical output*. Other frameworks, and proposals to classify errors have also been proposed including [Morgenstern \(1950\)](#) and [Verma et al. \(2010\)](#). [Morgenstern \(1950\)](#), as drawn on throughout this review, distinguishes ten sources of uncertainty:

- (1) lack of designed experiments;
- (2) hiding of information, lies;
- (3) the training of observers;
- (4) errors from questionnaires;
- (5) mass observations;
- (6) lack of definition and classification;
- (7) errors of instruments;
- (8) the factor of time;
- (9) observations of unique phenomena;
- (10) interdependence and stability of errors.

These are nested in the classification below. It is also common, as discussed in [Van der Bles et al. \(2019\)](#), to distinguish 'aleatory uncertainty', due to the fundamental indeterminacy or randomness in the world, from 'epistemic' uncertainty. Epistemic uncertainty is arguably what matters for statistical data that generally, but not always, seek to measure past or present (via a nowcast) phenomena. That is, our focus is on numbers that we currently do not know but could, at least in theory, know if only the information set were more complete. In contrast, 'aleatory uncertainty' generally relates to future events which we cannot know for certain.

This article is structured as follows. In Subsections 2.1 and 2.2 we review the traditional typology of data uncertainty that distinguishes sampling from nonsampling errors. In each case, we further break down the sources of sampling and nonsampling uncertainties, emphasising the importance of nonsampling uncertainties for administrative and big data sources. Throughout the importance of measurement – of the elements comprising uncertainty – is emphasised. As without measurement, it is hard if not impossible to gauge the consequences and importance of these uncertainties. Section 3 then introduces the new typology of uncertainties of [Manski \(2015\)](#). It is argued this confers some conceptual advantages when it comes to measuring and quantifying the different elements of total uncertainty. Section 4 then discusses the consequences, known and unknown, of data uncertainties. Section 5 concludes. In so doing it discusses the future agenda by repeating the call of [Manski \(2015\)](#) for more empirical research to study the effects of data uncertainty on the public's understanding, interpretation and use of official statistics.

2. The Traditional Typology of Data Uncertainties

2.1. Sampling Errors

Unlike the natural sciences, as [Morgenstern \(1950\)](#) emphasised, official economic statistics are not produced via repeated experiments. Instead, surveys are often run by national statistical institutes to measure the economic variables of interest. In Subsection 2.2.1, we turn to consider the growing use of administrative and big data sources in official statistics.

Sampling error is the most commonly reported measure of statistical uncertainty. This is because, unlike nonsampling errors discussed in Subsection 2.2 below, sampling error can be quantitatively estimated for many – but not all – sample surveys. Sampling error is the uncertainty or variability in an estimate that results from using a sample from a population rather than conducting a census or complete enumeration of the population.

If a sample from the population is chosen randomly, for example, then each random sample will involve sampling some different units and imply that each sample will produce different sample estimates. When there is great variation among the samples drawn from a given population (i.e. there is greater variability in the population), the sampling error is high. Then there is a larger chance that the survey estimate is far from the true population value. In a census when the entire population is surveyed there is no sampling error, but nonsampling errors still exist.

As summarised by the Office for National Statistics (ONS) in the United Kingdom (see [ONS 2017b](#)), standard errors are typically influenced by a number of factors that include:

- the survey sample size – a larger sample size will reduce standard errors;
- the variability in the population – when measuring a more variable characteristic, standard errors will be larger;
- the survey sample design – for example, any stratification or clustering used; and
- the estimation method used.

2.1.1. Measures of Sampling Error

Measures of sampling error associated with an estimate are typically based on estimates of the standard error and the mean squared error. The latter can be decomposed into the sum

of the square of the sampling bias and the sampling variance. The standard error is often used to compute the coefficient of variation or margin of error, both of which are related measures of the amount of uncertainty in the estimate.

Standard errors to measure sampling error can be computed – given the actual or assumed nature of the survey sample design. When sampling biases are zero or close to zero the standard error can be taken to represent total sampling error. This assumes that “population uncertainty” does not exist, which in practice is a strong often unrealistic assumption. Population uncertainty (see [Plumper and Neumayer 2012](#)) arises from the reality that ‘random samples’ from a given population may not be random, when there is uncertainty about who forms part of the population. As discussed in [Plumper and Neumayer 2012](#)), oversampling and sample selection corrections can be used to tackle population uncertainty.

There exist many approaches (e.g. see [Goedeme 2013](#)) to estimate standard errors. For an “easy-reading” discussion see [Peters \(2001\)](#). *Direct* estimators, which rely on analytic variance formulae, can be distinguished from *indirect* or resampling methods, like the bootstrap. Bootstrap methods involve taking a large number of draws from the original sample to mimic the actual sampling process: the sampling distribution of the target statistic across these bootstrap draws then measures the uncertainty. Whichever approach is used, the sampling process and the estimation procedure should ideally be acknowledged; for example see [Goedeme \(2013\)](#) who considers the complexities involved in estimating standard errors specifically for index numbers from complex surveys, and the discussion in [Seljak \(2006\)](#). For a detailed analysis of how uncertainty can be measured for mid-year population estimates given knowledge of the sampling processes and an understanding of the qualitative importance of different sources of uncertainty, see [ONS \(2017a\)](#).

As another example, [ONS \(2017a\)](#) consider how the bootstrap can be used to quantify the main uncertainties associated with UK migration data. Their method acknowledges the different data sources (including census, survey and administrative data) that are used to measure migration. Also see [Mevik \(2004\)](#) for a detailed study of sampling errors, from the Norwegian Business Tendency Survey, that contrasts “design-based” measures of the standard error with “model-based” ones that again make use of the bootstrap. For a general discussion of design-based versus model-based methods see [Koch and Gillings \(2006\)](#).

2.1.2. Challenges Measuring and Quantifying Sampling Errors

As discussed, standard errors can be computed – given the actual or assumed nature of the survey sample design. For statistical estimates of variables like GDP, the complexities involved in measuring the components, whether on the income or expenditure side, mean that it is not obvious what the survey design is and how this can lead to analytical expressions for the standard error.

To quote [ONS \(2019\)](#): “The estimate of GDP . . . is currently constructed from a wide variety of data sources, some of which are not based on random samples or do not have published sampling and non-sampling errors available. As such, it is very difficult to measure both error aspects and their impact on GDP. While development work continues in this area, like all other G7 national statistical institutes, we don’t publish a measure of the sampling error or non-sampling error associated with GDP”. This quotation is, in fact, remarkably reminiscent of the discussion in [Kuznets \(1948, p.176\)](#): “The treatment of

margins of error is most difficult for the national income and product statistics. The totals are a composite of a great variety of data, which differ in reliability from sector to sector of the economy. The margin of error in the composite totals is thus a complex amalgam of errors in the parts whose magnitude is not easily determined”.

But, as [Kuznets \(1948\)](#) emphasised, this complexity should not imply that attempts to measure these margins of error should not be made. [Kuznets \(1948\)](#), in fact, sought to quantify the uncertainties in GDP via expert judgement – famously concluding that there was a 10% margin of error associated with GDP.

However, it is possible to provide data-based and quantitative indications of “transitory” statistical uncertainties associated with GDP estimates by analysing historical revisions. National statistical offices and central banks accordingly often now publish realtime data vintages and analyse the implied revisions (e.g. see [Croushore and Stark 2001](#)). Other sources of uncertainty, for example due to limitations of the survey methodology, are not represented; and methodological work on measuring non-sampling errors continues (e.g. see [Manski 2016](#)).

There is also a long tradition (dating back at least to [Stone et al. 1942](#)) of exploiting the fact that for some variables there are multiple measures – albeit perhaps ones based on different sampling approaches. In particular, as a leading example, GDP can be estimated by the production, expenditure and income approaches. In principle, all three of these measures should be equal; but they are not in practice, given that they are calculated from different samples. But comparison of these approaches, assuming they all seek to measure the latent variable “true GDP”, can be used to produce so-called balanced or reconciled estimates of “true GDP” that also quantify the “statistical” or “measurement” error, as it is commonly referred to in this literature (e.g. see [Smith et al. 1998](#); [Aruoba et al. 2016](#)). Another cross-country example of how measurement errors can be quantified by comparing alternative estimates is how a specific country’s trade balance statistics can be compared with estimates from their trading partners: one country’s exports are another country’s imports.

2.2. Nonsampling Errors

It is more challenging to categorise, and certainly to measure and quantify, nonsampling errors for official statistics. Nonsampling errors stem from the design, data collection and processing methods used. As also seen in the typology of [Morgenstern \(1950\)](#), these errors often stem from lack of knowledge of the “nature of the data” given that the data are typically not measured by designed experiments. Nonsampling errors affect administrative (such as census) data as well as survey-based statistics. In general, sampling errors decrease as the sample size increases, but non-sampling errors increase as the sample size increases.

A common typology of nonsampling errors (e.g. see [Biemer and Lyberg 2003](#); [Eurostat 2019](#); [US Census Bureau 2018](#); [Statistics Canada 2017](#); [Australian Bureau of Statistics 2013](#); [National Science Foundation 2018](#)) is to decompose the nonsampling errors into the five elements listed in the typology below:

- (1) **Specification error:** Survey questions often cannot and/or do not perfectly measure the concept which they are intended to measure. For example, if asked to report whether they have a disability, respondents may have different subjective views of

what constitutes a *disability* and accordingly they provide different answers; as another popular example, the number of patents does not perfectly measure the quantity of invention in a macroeconomy. As emphasised in Manski's typology below, there can also be *classification* errors (perhaps reflecting conceptual uncertainties), for example, reflecting whether to classify some expenditure component of GDP as investment or consumption. Difficulties, and therefore errors, can also arise when, for instance, classifying economic activity to different industries. Economic activity is rarely confined to one specific industry.

- (2) **Coverage (or frame) error:** Coverage error occurs when the sample (frame) is inaccurate or incomplete, as a unit in the sample is erroneously excluded or included (e.g. duplicated), leading to *under* or *over* coverage errors. These errors make the survey less representative of the underlying population. The correction of coverage errors can be expensive, involving survey redesign and undertaking new surveys. A specific source of error that might be interpreted as stemming from a coverage error (albeit one known to the statistical office), arises in a mixed frequency data environment when statistical offices use temporal disaggregation methods to interpolate missing data at the higher frequency using observed data on higher frequency *indicator* variables. In effect, the temporal disaggregation methods fill in the gaps left by the incomplete survey evidence at the higher frequency. For example, monthly estimates of GDP are not commonly published by national statistical offices. So temporal disaggregation methods (from univariate models such as [Chow and Lin \(1971\)](#) to multivariate dynamic extensions, such as [Mitchell et al. \(2005\)](#) and [Frablet et al. 2011](#)) have been used to estimate monthly GDP based on the monthly movements of a range of observed indicator variables believed to relate to (unobserved) monthly GDP. Importantly, these methods impose the constraint that the interpolated monthly estimates for GDP add up to the quarterly totals published by the statistical office.
- (3) **Nonresponse error:** Nonresponse error occurs when not all units of the sample respond to the survey. This leads to a difference between the statistics computed from the collected or observed data and those that would be computed if there were no missing values. Two types of nonresponse can be delineated: (1) unit nonresponse, that is, when no data are collected about a population unit; and (2) item nonresponse, that is, when data on some but not all the survey data items are collected for a given population unit. Nonresponse can cause nonresponse bias (as well as nonresponse variance) when the observed sample differs systematically from those who do not respond (the unobserved sample). For example, complete or partial nonresponse is often more likely among lower-income or less-educated respondents or firms facing serious financial difficulties. The nonresponse rate can usually be accurately measured – as the ratio of the number of completed surveys to the total number of sample units. In turn, response rates therefore indicate the proportion of sample units that respond to the survey. However, these nonresponse rates do not help the user of the statistic directly infer, for example, the standard error of the estimate. They are therefore of limited direct use, as ideally the user would be provided with an estimate of the standard error. Subsection 2.2.1 below considers recent work in econometrics that has sought to quantify nonresponse errors directly.

- (4) **Unit-level measurement error:** response error and interviewer error: Measurement errors stem from what is observed or measured by the survey differing from the actual values for the sample units. Measurement errors, as defined here, relate to the accuracy of measurement at the unit level. In turn, measurement errors can be broken down into response errors and interviewer errors. Response errors arise when respondents knowingly or unknowingly provide inaccurate responses. These errors might arise due to inherent cognitive biases (e.g. a tendency for a respondent to give an answer that they believe is correct or will please the interviewer) and poorly designed survey questionnaires that lead to misunderstandings about what is being asked. Interviewer errors arise when the person undertaking the survey, whether on purpose or not, records incorrect responses or consciously or unconsciously influences the respondent with the effect that they provide inaccurate responses.
- (5) **Processing error:** Processing errors include errors in recording, checking, coding and preparing survey data. They can include interpolation and extrapolation errors for missing or, what are believed to be, *inconsistent* data. In some contexts (some of) these errors can be measured and quantified. [Van der Loo et al. \(2017\)](#) consider the variance caused by data cleaning. They note that survey data sets, $X = \{x_i\}_{i=1}^n$, often suffer from missing values, outliers and incorrect values that preclude the applicability of a simple estimator $f(X)$ such as the sample mean. Although, as touched on below, robust estimators may be more appropriate. So data editing processes are used to transform the raw data set X into a new data set, Y , which is then used for estimation. The population estimator is also given by $f(Y)$. But as Y is a transformation, the variance of the estimator after editing is no longer simply the variance of the new data set but should also reflect the *extra* variance induced by the data editing processes. This extra variance may comprise estimation uncertainties as well as sampling variance. As [Van der Loo et al. \(2017\)](#) explain, in general – in real-life practical examples of interest to official statistics – it is hard to obtain analytical expressions for this composite uncertainty; and they therefore suggest a computational approach to measuring the variance that uses the bootstrap. But in other contexts isolating and removing processing errors (without a warning or help from the statistics office) simply by inspecting a published time-series can be challenging. It amounts to having to define and then isolate *outliers*. This raises identification challenges, since an outlier could be due to variability in the sampling processes rather than a processing (or measurement) error. Interestingly, given that some statistical estimators are more robust to outliers (or more generally to uncertainty whatever the source) than others, the use of robust estimators may offer promise when communicating data in the presence of uncertainties. For example, the median rather than the mean offers a robust measure of central tendency.

2.2.1. Measures of Nonsampling Error

As emphasised, nonsampling errors are typically hard to measure and quantify. But, as [Manski \(2015\)](#) emphasises, this does not justify ignoring them.

Statistically, it can be helpful to consider that nonsampling errors can be classified into two groups: random errors and systematic errors. Random errors are the unpredictable (ideally, independently and identically distributed) errors. They generally cancel out if a

large enough sample is used. They lead to increased variability in the statistic, but no bias. In turn, systematic errors are errors that accumulate. For example, if there is an error in the survey or questionnaire design, this causes errors in respondents' answers, often leading to biases. We first consider, in Subsubsection 2.2.2 how measures of nonsampling error may be computed using partial identification methods. Then, in Subsubsection 2.2.3, we consider the scope to measure nonsampling errors for administrative and big data.

2.2.2. Partial Identification and Measures of Survey Nonresponse

Recent work on “partial identification” (Manski 2016) has shown how, with access to the underlying micro data, more could be said about nonsampling uncertainties (for aggregated data) – and in particular nonsampling errors due to survey nonresponse – than at present is commonplace.

The basic idea is that in the presence of missing data, sample statistics can still be computed. But to measure the nonsampling errors, due to missing data, these statistics can be computed taking into account all the values that the missing data might take. This delivers interval rather than point estimates. The approach of Van der Loo et al. (2017), who as mentioned uses the bootstrap, can also be interpreted within this framework.

In simple terms, Manski (2016) sets out how if one lets $P(y|z = 1)$ denote the distribution of random variable Y for those units who report y (denoted, $z = 1$; $z = 0$ otherwise), then from the law of total probability

$$P(y) = P(y|z = 1)P(z = 1) + P(y|z = 0)P(z = 0) \tag{1}$$

The sample evidence reveals $P(z)$ and the observables $P(y|z = 1)$ when $P(z = 1)$. But the sample evidence is uninformative on $P(y|z = 0)$. Therefore the sample evidence reveals that $P(y)$ lies in the identification region

$$H[P(y)] = [P(y|z = 1)P(z = 1) + \gamma P(y|z = 0)P(z = 0), \gamma \in \Gamma_Y] \tag{2}$$

where Γ_Y denotes the set of all probability distributions on the set Y . As discussed by Manski (2016), the notion of the identification set can then be used for meaningful inference. For example, suppose the statistics office is interested in quantifying the probability that Y falls within some interval or set, B i.e. $P(y \in B)$. Then, again by the law of total probability,

$$P(y \in B) = P(y \in B|z = 1)P(z = 1) + \gamma P(y \in B|z = 0)P(z = 0) \tag{3}$$

and the empirical evidence reveals $P(y \in B|z = 1)$, $P(z = 1)$ and $P(z = 0)$. But it does not reveal $P(y \in B|z = 0)$. However $P(y \in B|z = 0)$ must lie between 0 and 1. This yields the sharp bound on $P(y \in B)$:

$$P(y \in B|z = 1)P(z = 1) \leq P(y \in B) \leq P(y \in B|z = 1)P(z = 1) + P(z = 0) \tag{4}$$

If the statistician is willing and/or able to make assumptions on the nature of the nonresponse, that restrict $P(y|z = 0)$ within some probability space (e.g. to a specific set of density functions), then these bounds can be made tighter.

2.2.3. Nonsampling Error: Administrative and Big Data

The increasing availability and use of administrative and big data, including from new data sources (such as the internet and social media), raises both new challenges and opportunities for the measurement and quantification of uncertainties, especially those coming from nonsampling errors. The hope is to exploit some data source that provides the “true” estimate; comparison of other estimates, say from surveys, with this “true” estimate then provides a clear way to measure the statistical or measurement “error” of the other survey-based estimate.

But many of these administrative and big data sources were designed for purposes other than official data collection. Indeed, many of the data sources are from private companies, raising challenges as to data ownership and privacy. This has prompted experimental research into, for example, the use of apps (see [Gromme et al. 2017](#)) to measure directly data from the population (the “citizens”) rather than rely on third party data. As emphasised by [Hand \(2018\)](#), analysis of administrative data presents new statistical challenges not least that these data are, by definition, typically not random samples but so-called non-probability samples.

As stressed by [Kapteyn and Ypma \(2007\)](#) and [Abowd and Stinson \(2013\)](#), while administrative data and indeed big data sets in general offer the prospect of fewer non-response errors than traditional surveys, they still likely suffer from uncertainties in particular due to “measurement error”. They may not measure exactly the concept a researcher is interested in. And since administrative databases typically link data from different sources there is the possibility of mismatching, due to imperfect linkage information (e.g., errors in social security numbers). [Abowd and Stinson \(2013\)](#) therefore emphasise the errors that are present in all data sources; and, in the tradition of [Stone et al. \(1942\)](#), they specify a so-called prior weight vector used to define the “truth” as a weighted average of both the administrative and the survey data.

Therefore new sources of uncertainties, but also opportunities to reduce these, arise from this practice of matching administrative or big data with existing, more traditional, sources of data collected by the statistics office. Matching involves combining information available in distinct sample surveys about the same target population. For example, work by [Lui et al. \(2011\)](#) sought to match firm-level qualitative survey data from the Confederation of British Industry, a business organisation in the United Kingdom (that provides information on a range of variables not posed in official surveys) with those same firms’ responses to official surveys from the ONS. This sort of matched data set offers the prospect of both better understanding the nature and statistical properties of the non-official data and of cross-checking the accuracy of the new data.

[Kapteyn and Ypma \(2007\)](#) provide a framework to model the errors in administrative data due to mismatching, based on a comparison of the administrative and survey data when estimates from both are available. [Conti et al. \(2012\)](#) also consider the measurement of uncertainty in statistical matching. [Conti et al. \(2012\)](#) set out a model that can be used to estimate the joint distribution of variables observed in separate and independent surveys. Consider two surveys that deliver random variables Y and Z , with observations y and z , respectively. Both surveys also provide information on a set of additional random variables, X , with observations x . Let the two (known) marginal conditional distributions then be denoted, $F(y|x)$ and $F(z|x)$. In the spirit of “partial identification” as set out by [Manski](#)

(2016), Conti et al. (2012) measure uncertainty as the set of probability distributions of the random vector $(Y, Z|X)$ compatible with $F(y|x)$ and $F(z|x)$. Again a bounds-based approach is proposed as a way to quantify the uncertainty. Recent work by Oberski et al. (2017) extends analysis to estimate the extent of measurement errors in administrative data that measures the errors in administrative data, allowing both the administrative data and the survey data to be simultaneously subject to measurement errors.

Coverage errors, as discussed in Subsection 2.2, remain a concern for administrative and big data data. The administrative population is often a proxy for the target population. For example, the employment register in the Netherlands also contains employees that work but do not live in the Netherlands; but it misses Dutch inhabitants who have a job abroad. And with big data, often the observations cannot be identified/linked to a member of the target population; it is then not directly possible to find out how representative the data, often based on non-probability samples, are.

“Hiding of information”, one of the elements in the typology of errors listed by Morgenstern (1950), may afflict at least some administrative data. For example, when replying to the tax authorities, individuals or businesses may in a sense, as Morgenstern puts it, “deliberately lie” or at least obscure the truth. As Morgenstern (1950) writes, there is a long history of apparently venerable institutions falsifying or at least obscuring “facts” for strategic or political purposes. A more recent example of this typology of error is when Greek public finance data were investigated by the European Commission in 2010. In the final report of the European Commission’s investigation (see European Commission 2010) it was recognised that there were political pressures to obstruct accurate data collection. Nevertheless, growing use at statistical offices of individual and firm-level tax return data in the national accounts is delivering improvements both in terms of timeliness and accuracy; that is, the use since 2017 of Value Added Tax data in measuring GDP in the United Kingdom means that turnover data from 630,000 businesses now inform GDP estimates (see ONS 2018b), providing wider coverage.

Some of Morgenstern’s other elements in his typology are also probable sources of uncertainty for big data and administrative data. His “mass observations” may well involve errors that likely accumulate and do not necessarily cancel out. And his “errors of instrument” are also likely to become more important as economic statistics are increasingly collected by machines (e.g. scanners) rather than human beings (via surveys and questionnaires).

Measurement of these uncertainties for administrative and big data is in its infancy. Chambers (2014) considers how “model-based thinking” can help measure nonignorable nonresponse in surveys; and how adaptive surveys can be used to select a sample of nonrespondents to interview or survey at a second wave of the survey so as to minimise the nonresponse bias. ONS (2018a) recommend that for administrative data measures of coverage and completeness, editing rates and imputation rates should be measured and used as quantitative estimates of data uncertainty. And the aforementioned approach of Morgenstern (1950) is attractive in measuring uncertainties without having to assume that either the administrative data or the survey data are accurate. Hand (2018) provides a recent discussion; and calls for research to establish what the ‘generally accepted theory’ might be for the analysis of administrative data. The first of his challenges is, in fact, to consider how to define and communicate uncertainty for administrative data, given that the sources of uncertainty in administrative data are many and diverse, and may not include sampling variation.

3. Manski's (2015) Typology of Data Uncertainties

Manski (2015) re-interprets sampling and nonsampling uncertainty as comprising three elements:

- (1) “transitory” statistical uncertainty;
- (2) “permanent” statistical uncertainty;
- (3) conceptual uncertainty.

Transitory statistical uncertainty stems from publication of early data releases that are revised over time as new information arrives. For example, for many years the Office for National Statistics (ONS) in the United Kingdom published its first – so-called “preliminary” – quarterly GDP estimates around 27 days after the end of the quarter. Because this timeliness was achieved by basing their estimate on 44% of the sample, it is (and should be) no surprise to see the ONS revise these preliminary estimates as more sampling information subsequently became available to them. Interestingly, in the summer of 2018 ONS shifted back its production – so that the new so-called *first* estimate is now available *only* at about 40 days. But this delay *buys* the ONS a higher sampling fraction, and should therefore reduce transitory statistical uncertainty.

Permanent statistical uncertainty arises due to data incompleteness (e.g. nonresponse) or the inadequacy of data collection (e.g. sampling uncertainty due to a finite sample) which does not diminish over time. Therefore, permanent statistical uncertainty comprises elements of both sampling and nonsampling errors, as delineated in the typology above.

Conceptual uncertainty arises from a lack of understanding about what the statistics measure. It arises not from the statistics themselves, as with transitory and permanent statistical uncertainty, but from how the statistics are interpreted. Conceptual uncertainty is of course not a new element in many typologies of the sources of uncertainty, albeit it is one that is often ignored – given the challenges in measuring it. In fact, discussion of conceptual uncertainty again dates back to Morgenstern (1950), who discussed uncertainties arising due to a “lack of definition and classification”. As Morgenstern (1950, 35) explains, “the theoretical characteristics of, say, an industry or a ‘price’, are less well established than those of a wave length”.

It is perhaps helpful to begin to break down conceptual uncertainty by considering the following components or sub-elements:

- conceptual uncertainty due to different definitions and classifications. This overlaps with the eighth element in the typology of errors of Morgenstern (1950): “the factor of time”. Such uncertainties arise as economic data are often measured on a discrete basis with observations attributed to a specific window of time – so there are *errors* especially when there are changes in classification (e.g. changes in the definition of an industry or changes to the characteristics of a specific product);
- conceptual uncertainty due to differences in the compilation process (e.g. direct estimates based on surveys or administrative data versus indirect estimates using temporal disaggregation techniques); and
- conceptual uncertainty due to seasonal adjustment.

Conceptual uncertainty need not produce “errors” in the usual statistical sense – as [Morgenstern \(1950\)](#) explains. But differences of definition, for example, clearly result in uncertainties, revisions to estimates and doubts as to the use and interpretability of data and their comparability.

3.1. A Bayesian Approach

In principle, at a formal level, the approach of [Draper \(1995\)](#) offers a methodological way of understanding some aspects of conceptual uncertainty. If we consider these aspects as part of the “model” used to measure the underlying variable, then [Draper \(1995\)](#) provides an approach to think about both uncertainty about the form of the model (so-called structural uncertainty) and the parameters of the model (so-called parametric uncertainty). This motivates a Bayesian approach.

Bayesian Model Averaging (BMA) offers a conceptually elegant means of dealing with model uncertainty. BMA is an application of Bayes’ theorem, Model uncertainty is incorporated into the theorem by treating the set of models S as an additional parameter and then integrating over S , where $S \equiv \{S_i, i = 1, \dots, N\}$ with N models, and the models S_i are defined as continuous density functions $g_{it}(y_t)$ for the variable of interest y_t .

The posterior density estimate of the variable of interest y_t given ‘data’ Ω_t , $p_t(y_t | \Omega_t)$, is then defined as the weighted average of the densities $g_{it}(y_t) = \Pr(y_t | S_{it}, \Omega_t)$, where the weights w_{it} are the different model’s (or the different estimates’) posterior probabilities, $w_{it} = \Pr(S_{it} | \Omega_t)$:

$$p_t(y_t | \Omega_t) = \sum_{i=1}^N w_{it} g_{it}(y_t); (t = 1, \dots, T), \quad (5)$$

where $w_{it} \geq 0$ and $\sum_{i=1}^N w_{it} = 1$. $p_t(y_t | \Omega_t)$, or, for expositional ease suppressing dependence on the ‘data’ Ω_t when defining the posterior probabilities, equivalently $p_t(y_t)$, is the combined density estimate.

A Bayesian approach, due to its ability to handle multiple sources of uncertainty, also offers promise as a way to provide an integrated measure of total uncertainty – that integrates out uncertainty about sampling and nonsampling errors or transitory and permanent uncertainty. In addition, priors can be used to acknowledge if and when there is additional information that can be used to guide the data in the right direction.

3.2. Measures of Transitory, Permanent and Conceptual Errors

An advantage of the typology of [Manski \(2015\)](#) is that when it comes to actually measuring and quantifying the three elements of total uncertainty, the first element – transitory statistical uncertainty – when relevant, is at least usually measurable. It is measurable by analysing revisions to the statistics as more information becomes available as the statistic is revised. As elaborated on in Subsection 3.2.1, measurement and quantification of transitory statistical uncertainties is now facilitated by the relatively wide availability of real-time (vintage) data sets for macroeconomic variables such as GDP. These real-time data sets let one measure the revisions between successive estimates. Many authors have proposed models of data revisions – using real-time GDP data sets –

to model and forecast this “transitory” GDP data uncertainty (e.g. see [Jacobs and Van Norden 2011](#); [Cunningham et al. 2012](#); [Kishor and Koenig 2012](#); [Galvao 2017](#)).

In turn, measurement of the permanent and conceptual uncertainties is again challenging, as it is for sampling errors (at least for variables like GDP) and nonsampling errors. But attempts can still be made to communicate (at least some of) these uncertainties. A famous example of how uncertainties can be communicated even for a variable like GDP, which as discussed, is usually subject to multiple surveys precluding direct estimates of the data’s standard error, are the fan charts published by the Bank of England; see [Figure 1](#) for an example.

[Figure 1](#) provides an illustrative example of what these fan charts look like, taken from the Bank of England’s *Inflation Report*. Importantly, from [Figure 1](#) we see that the Bank seeks to quantify both future uncertainties but also past or historical data uncertainties. This is emphasised in the words that accompany the fan chart pictures in the Bank’s publications: “(t)o the left of the first vertical dashed line, the centre of the darkest band of the fan chart gives the Committee’s best collective judgement of the most likely path for GDP growth once the revisions’ process is complete” ([Bank of England 2007](#), 39). As the Bank of England explains, these fan charts should be interpreted as “the MPC’s best collective judgement of the most likely path for the mature estimate of GDP growth, and the uncertainty around it, both over the past and into the future.” [Figure 1](#) reveals that the fan becomes progressively narrower as one looks further back in time, as the data revisions’ process is more complete and fewer future revisions are expected to older

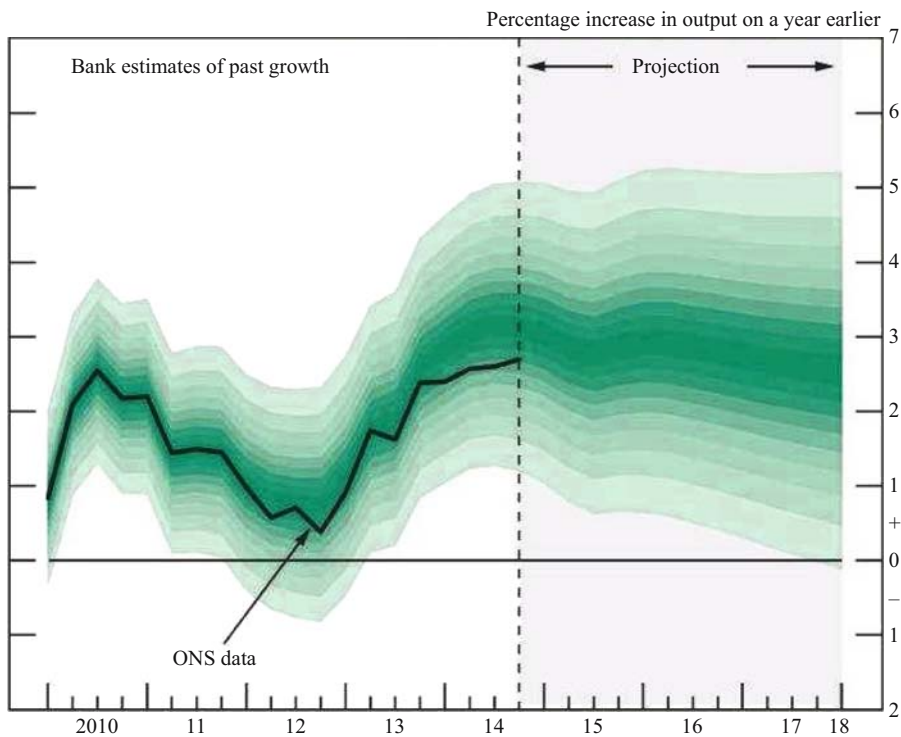


Fig. 1. Illustrative Bank of England fan chart –GDP data uncertainty.

estimates. The ONS's latest estimate of GDP growth is shown in [Figure 1](#) by the solid black line. [Cunningham and Jeffery 2007](#) provide an explanation of the data revisions model, used by Bank staff, that along with judgement from the Bank's Monetary Policy Committee helps shape the form of these backcast fan charts. Their model exploits historical patterns in ONS revisions and information from qualitative business surveys. The Bank assume that data uncertainty is determined by a Gaussian probability density function (see [Bank of England 2007](#)); and the mean of this probability density function does not always equal or have to equal the ONS's latest GDP estimate. This enables the Bank to quantify biases, as well as the variance around the point estimate. For a detailed discussion and an *ex post* calibration analysis of the Bank of England's probabilistic backcasts see [Galvao and Mitchell \(2020\)](#).

3.2.1. Measurement of Transitory Uncertainty: Real-time Data sets

When considering temporary or transitory uncertainty, the availability of a properly defined and updated real-time database constitutes a powerful tool. By a properly defined and updated real-time database we mean one where vintages are recorded by the date of official release. Commonly, in a real-time database, the data for each vintage (the date at which the data are released) are presented in a different column, with the rows referring to the date for which the time series is measured. The final data value shown in each column is therefore the first release of the data point indicated by the row. This means that real-time databases often take the form of data triangles – as more recent data vintages (columns to the right) provide data over longer (more recent) periods of time. Some real-time databases are based on vintages collected at fixed intervals, such as each week or month. In this case, especially when vintages are collected monthly, multiple releases within the month are lost, with only the latest stored. This can compromise the overall quality of the revisions analysis. Returning to the treatment of transitory uncertainty, comparison of the vertical (or column) data vintages can be used to identify characteristics of the revisions process, such as its bias and persistence. In addition, by considering diagonal data vintages (i.e. data down the diagonals of the data triangle) it is possible to study the different degrees of uncertainty associated with various releases as well as empirically check for the existence of a trade-off between timeliness and accuracy. Finally, by comparing horizontal (or row) data vintages it is possible to study uncertainty at specific points in time, such as those related to a period of crisis, to see if their intrinsic uncertainty is different to observations in more *normal* situations.

The increased accessibility of real-time data sets, in particular in the United States as maintained by the Federal Reserve Banks of Philadelphia and St. Louis, has enabled statisticians and economists to analyse and model data revisions; (for example, see [Croushore and Stark 2001, 2003](#); [Croushore 2011](#); [McCracken and Ng 2016](#)). Real-time databases, especially for GDP, are now available for various countries. This includes the Euro area (as maintained by the European Central Bank; see [Giannone et al. 2012](#)), Switzerland ([Bernhard 2016](#)), the United Kingdom (as maintained by the Bank of England), Canada (as maintained by Statistics Canada), New Zealand (as maintained by the Reserve Bank of New Zealand), and the OECD. These databases thus provide a means to measure and quantify transitory statistical uncertainty.

3.2.2. Measurement of Conceptual Uncertainty: Seasonal Adjustment

Conceptual uncertainty involves both subjective and objective components. Measurement of the former is more challenging; and there has been little or no work on it to date. Measurement, perhaps qualitative, would appear to require the design and use of new surveys to gauge, for example, the public's interpretation of GDP data – do the public understand correctly what GDP measures? For example, experimental evidence from 3000 members of the UK public suggests that a large fraction of the public do not correctly understand what GDP measures; see [Galvao et al. \(2019\)](#). But as the sub-classification of conceptual uncertainty suggests, some aspects of conceptual uncertainty can be measured quantitatively (*ex post*, i.e. after the revision) by examining revisions to statistics.

Revisions to seasonally adjusted real-time data can be decomposed into two separate but related sources. The first source is the application of the method used for seasonal adjustment. As seasonal adjustment involves application of a filter to the underlying series, with the passage of time as new data accumulate the weights attached in the filter to specific observations change and there are revisions to the seasonally adjusted estimates; see [Wallis \(1982\)](#) for further details and analysis. [Burrige and Wallis \(1985\)](#) discuss how the variance of the seasonally adjusted series can be calculated when the seasonal adjustment filter is recast as an optimal filtering problem in an unobserved components framework.

The second source of revisions is that for many series, like GDP, the unadjusted data are themselves revised by the statistics office. Attempts to quantify these revisions, paralleling the literature that has examined the revisions properties of output gap estimates (cf. [Orphanides and Van Norden 2002](#)), have involved recursive real-time application of the seasonal adjustment filter to the real-time unadjusted series. For example, [Mehrhoff \(2008\)](#) considers the empirical quantification of both these sources of uncertainty for selected German time series, using the real-time database of the Deutsche Bundesbank.

Additional uncertainties arise when calculating seasonally adjusted estimates for aggregated variables, such as Euro area GDP. This is because Euro area GDP involves the aggregation of GDP data for the member countries. The question then arises of when the series should be seasonally adjusted. One can distinguish two approaches. The “direct” approach consists of seasonally adjusting the raw data of the aggregate itself. The “indirect” approach consists of seasonally adjusting the raw data corresponding to the sub-components (national GDP) and then aggregating. In general, the direct and indirect adjustment for an aggregate series are not identical. Only, under certain conditions, for so-called uniform seasonal adjustment filters, such as X-11, does the order of seasonal adjustment and aggregation not matter. But when the filters differ, as they will when an optimal signal extraction method is used like an unobserved components model, the order is crucial. For further discussion see [Ghysels \(1997\)](#). For completeness we note that one can also distinguish a “multivariate” or simultaneous approach that has certain optimality properties, although since it is computationally demanding and requires hard choices to be made about the appropriate information set this approach is rarely considered; see [Geweke \(1978\)](#).

4. Consequences of Uncertainty

While the impact of sampling errors (in the first typology) or transitory statistical uncertainty (in Manski's typology) can be measured and quantified – at least for some

variables – as emphasised, measuring and quantifying nonsampling errors and permanent statistical uncertainties is much harder. This means it is not generally possible to measure quantitatively, for a specific variable, the relative importance of the different elements or components of total survey error as delineated in the two typologies. However, perhaps in part subjectively formed, attempts can still be made to communicate the total error, as the Bank of England's fan charts illustrate.

It is an open question whether it is better to try and communicate these data uncertainties or not; and if so, how? Is it best to communicate data uncertainties quantitatively, like in the Bank's fan charts, or qualitatively perhaps via textual caveats and qualifications that emphasise that the data are uncertain?

Understanding, and certainly measuring, the consequences of uncertainty requires a cross-disciplinary approach, involving the intersection of psychology, behavioural and decision science and statistics. Consistent with the conclusions of [Manski \(2015\)](#), it requires new empirical research to study the effects of uncertainty – and its communication or lack of – on users' understanding, interpretation and use of statistics. A start is made in [Galvao et al. \(2019\)](#) who conduct randomised controlled experiments and a targeted expert survey in order to assess if and how the UK public and experts interpret and understand GDP data uncertainty. They find that the majority of the UK public do understand that there is uncertainty inherent in GDP point estimates. But communicating uncertainty information to them improves their understanding of why data revisions happen but does not decrease trust in the data. Their results indicate that it is particularly helpful to communicate uncertainty information quantitatively via intervals, density strips and bell curves.

In the absence of (to date) a body of published research both across countries and economic variables on this neglected issue, here three ways of understanding and measuring the consequences of data uncertainty are discussed. First, we review the growing literature, especially in economic statistics, that has sought to analyse and model data revisions. Secondly, we provide a case-study illustrating how for GDP growth the size of data revisions – of transitory statistical uncertainty – varies both across time and countries. Thirdly, we emphasise how the effects of uncertainty, particularly of transitory statistical uncertainty, relate to the trade-off between the timeliness and accuracy of statistics.

4.1. Revisions: Real-Time Data Analysis

As [Croushore \(2011\)](#) reviews, over the last 15 years there has been a growing literature, especially in applied macroeconomics, on if and how data revisions matter. [McKenzie \(2006\)](#) delineates seven reasons for “revisions” including updated sample information, correction of errors, benchmarking, updated base period for constant price estimates and changes in statistical methodology.

Research has examined the properties of data revisions, how structural macroeconomic modelling is affected by data revisions, how data revisions affect forecasting, the impact of data revisions on monetary policy analysis, and the use of real-time data when nowcasting. This research has been supported by the increased, but still imperfect, availability of realtime data sets by central banks and statistical offices. Importantly, as [Croushore \(2011\)](#)

concludes, until these data sets became more widely available most economists thought that data revisions were likely to be small and did not matter. But this view has been shown to be misplaced by real-time research: data revisions are often found to be large and have important implications, including for policymakers like central banks.

The general framework often used to measure and analyse the properties of these data revision uncertainties is twofold. First, studies typically report the mean (or bias) of the revisions and test if these are statistically significant. Secondly, to provide more information on the nature of the uncertainties and the ensuing revisions, studies discriminate between *news* and *noise* revisions following the approach of [Mankiw and Shapiro \(1986\)](#).

Tests for whether revisions are *news* or *noise* are based on so-called forecast efficiency regressions:

$$(y_t^{new} - y_t^{old}) = \beta_0^{news} + \beta_1^{news} y_t^{old} + \varepsilon_t \quad (6)$$

$$(y_t^{new} - y_t^{old}) = \beta_0^{noise} + \beta_1^{noise} y_t^{new} + \varepsilon_t \quad (7)$$

where y_t^{new} denotes the latest or new estimate of variable y at time t , and y_t^{old} denotes the previous or older estimate of variable that is revised.

The null hypothesis that data revisions add information (they contain *news*) implies $\beta_1^{news} = 0$. If data revisions remove the measurement error (*noise*) in the initial release then $\beta_1^{noise} = 0$. If data revisions reduce noise, then data revisions are predictable. For additional details on the application of these tests see [Clements and Galvão \(2010\)](#) and references therein.

4.2. Case-study on GDP: Cross-country Comparisons Measuring Data Revisions

To illustrate the importance, or otherwise, of transitory statistical uncertainties we review recent cross-country comparisons, from existing studies, that have sought to compare GDP data revision *errors* across countries. These papers build on the pioneering work of [Mankiw and Shapiro \(1986\)](#) and [Faust et al. \(2005\)](#). [Faust et al. \(2005\)](#) found that in the G7 economies, revisions to GDP announcements are large – many revisions in quarterly GDP growth are over a full percentage point at an annualised rate. Moreover, they found that while US GDP revisions are largely unpredictable, as predicted by the *news* model, for Italy, Japan and the United Kingdom, about half the variability of subsequent revisions can be accounted for by information available at the time of the preliminary announcement – so there was evidence for *noise*.

While studies that measure and then quantify data uncertainties, due to data revisions, are a helpful method to measure transitory statistical uncertainties, as ever we should recall the last of the ten sources of uncertainty listed in the typology of [Morgenstern \(1950\)](#): “the interdependence and stability of errors”. Measures of data uncertainty based on historical revisions measure just that: “historical” data revisions. They are therefore only a good guide to current data uncertainties to the extent that we expect history to repeat itself. If the statistical office, for example, has improved its measurement processes over time we might well expect current data uncertainties to be less than historical ones.

We should emphasise that different ways of producing and estimating GDP across countries no doubt affect the balance or relative importance of the different sources of GDP data uncertainty.

In more recent work, [OECD \(2015\)](#) compares GDP revisions across OECD countries. [OECD \(2015\)](#) uses the mean revision to measure the importance of data revisions. [Figure 2](#), taken from [OECD \(2015\)](#), provides estimates of the mean revisions for a range of OECD countries, importantly using different measures of the *outturn* – the *final* estimate against which the first GDP estimate is compared. Of course, as data revisions are an ongoing process the true or final estimate is never in reality observed, so an assumption has to be made. In applied macroeconomic studies, it is common to take the $t + 2$ year or $t + 3$ year as the *final* estimate. This is based on the assumption that revisions after this date are more unpredictable often reflecting, for example, benchmarking revisions.

[Figure 2](#) shows that most countries make upward revisions to their initial GDP estimates and this is so across different measures of the outturn. This implies that countries tend to underestimate GDP growth in their early estimates. An obvious exception is the United States that initially overstates GDP. Belgium, France, Germany, Italy, Norway, Korea and the United Kingdom make the lowest mean revisions.

However when these revisions are tested for bias, using statistical significance tests (see [Figure 3](#)), [OECD \(2015\)](#) concludes that short-term revisions (up to five months and after one year) are random and centered around zero for most countries. Focusing again on the Year-on-Year (YoY) growth rates, only Belgium, Australia, Norway, Denmark and the Netherlands experience statistically significant revisions. However, there is more evidence for bias in the longer term.

These results are supported by the more recent cross-country results reported in [Walton \(2016\)](#); see [Figure 4](#). [Figure 4](#) again shows that most countries (again with the notable exception of the United States) have made upwards revisions to their early GDP estimates, and this result holds across alternative measures of the outturn. This reinforces the finding that data uncertainties matter – and that these data uncertainties can involve mean (bias) terms as well as variance components. An important question for future research is why some countries have high(er) revisions and why some have low(er) and whether, for example, institutional factors and different statistical systems help explain the observed cross-country variations. Most OECD countries release their first GDP estimates around 30 to 45 days after the end of the quarter; so (large) cross-country differences in timeliness

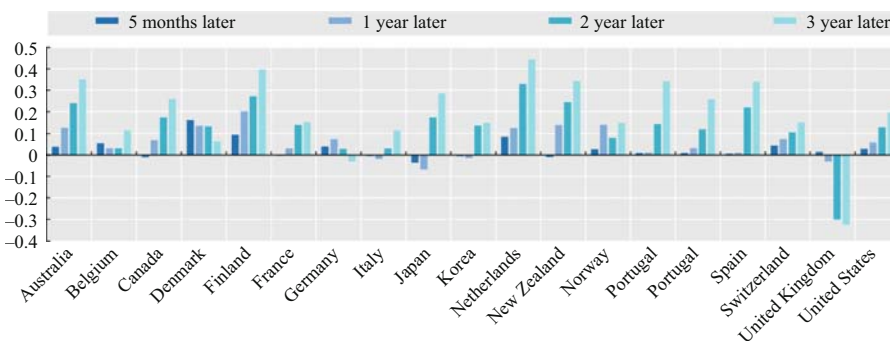


Fig. 2. OECD (2015) cross-country comparison of the importance of GDP revisions: Mean revision to first published YoY quarterly GDP growth rates (in percentage points) from 1994Q4. Source: OECD (2015). Reproduced with permission of the OECD.

Country	5 months later		1 year later		2 year later		3 year later	
	QoQ	YoY	QoQ	YoY	QoQ	YoY	QoQ	YoY
Australia	0.04	0.04	0.02	0.12	0.07	0.24	0.11	0.35
Belgium	0.03	0.06	0.04	0.03	0.03	0.03	0.07	0.11
Canada	0.01	-0.01	-0.01	0.07	0.03	0.17	0.05	0.26
Denmark	0.16	0.16	0.08	0.13	0.10	0.13	0.09	0.06
Finland	0.05	0.10	0.07	0.20	0.08	0.27	0.13	0.40
France	-0.02	0.00	-0.02	0.03	0.00	0.14	0.02	0.15
Germany	0.04	0.04	0.03	0.07	0.04	0.03	0.02	-0.03
Italy	0.02	-0.01	-0.01	-0.02	0.01	0.03	0.04	0.11
Japan	-0.02	-0.04	-0.01	-0.07	0.00	0.17	0.10	0.29
Korea	0.00	-0.01	0.00	-0.01	0.04	0.14	0.07	0.15
Netherlands	-0.01	0.09	0.06	0.12	0.07	0.33	0.10	0.44
New Zealand	0.01	-0.01	0.00	0.14	0.06	0.24	0.11	0.34
Norway	-0.03	0.03	-0.01	0.14	0.00	0.08	-0.01	0.15
Portugal	-0.01	0.01	-0.13	0.01	-0.04	0.14	0.01	0.34
Spain	0.00	0.01	0.01	0.03	0.01	0.12	0.07	0.26
Switzerland	0.00	0.01	-0.02	0.01	0.03	0.22	0.08	0.34
United Kingdom	0.02	0.04	0.01	0.07	0.02	0.11	0.05	0.15
United States	0.01	0.02	0.01	-0.03	-0.06	-0.30	-0.07	-0.33
Average	0.02	0.03	10.01	0.06	0.03	0.13	0.06	0.20

Statistical significance levels 1% 5% 10%

Fig. 3. OECD (2015) estimates of the bias to GDP revisions: Mean revision and statistical significance at different revision intervals for QoQ and YoY quarterly GDP growth rates (in percentage points) from 1994Q4–2013Q4.

Source: OECD (2015). Reproduced with permission of the OECD.

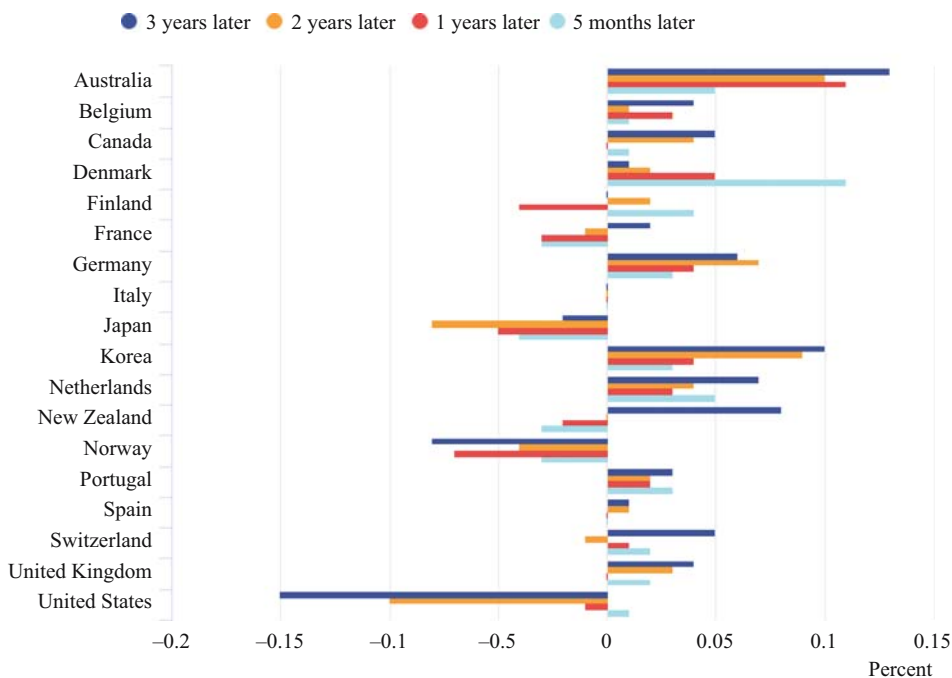


Fig. 4. ONS (taken from Walton 2016) estimates of the bias to GDP revisions for selected OECD and G20 countries: Mean revisions to GDP growth (chain volume measure, quarter on quarter) from 1998Q4–2015Q2. Source: Walton (2016). Reproduced with permission of the ONS.

are likely not a key factor in explaining the observed cross-country variations in data revisions bias.

4.3. Trade-Off Between Timeliness and Accuracy

As discussed, in particular when discussing transitory statistical uncertainty, official measures of variables like GDP, from national statistical offices, are revised as new information is received and methodological improvements are made. So the aforementioned move by the ONS in the United Kingdom to wait 13 days longer before publishing its first GDP estimate means their estimates will now be based on a higher data content than previously. This is expected to deliver more reliable GDP estimates, subject to fewer revisions – emphasising the important trade-off between the timeliness and accuracy of many statistical estimates. That is, statistics with fewer uncertainties can often be produced by delaying publication until more sampling and nonsampling information becomes available. But this delay may impede policy decisions.

As emphasised by [Ruggeri Cannata and Mazzi \(2017\)](#), the choice of the optimal point on the timeliness-accuracy trade-off *curve* has always been open to debate. Decisions made have depended on the characteristics of the statistical system in a given country and users' and policymakers' requests. For example, in the United States, macroeconomic variables tend to be very timely thanks to an efficient and centralised statistical system. By contrast, as [Ruggeri Cannata and Mazzi \(2017\)](#) go on to discuss, in Europe focus was historically more on accuracy although, with the advent of monetary union, pressure (from users and policymakers) for more timely and higher frequency data increased. However, greater decentralisation made it more difficult for Europe to speed up extant production systems. So new statistical systems, such as the construction of flash estimates using econometric methods, were adopted to meet the need for more timely statistics. As national statistical offices now transition to an era of big data, it is increasingly evident that the historical trade-offs between timeliness and accuracy will change, as timely big data are used to complement traditional sources of information. Accuracy may well increase with timeliness.

5. Conclusion

This article emphasises the data uncertainties present in official economic statistics. It accordingly reviews different categorisations of uncertainty, specifically the traditional typology that distinguishes sampling from nonsampling errors and the newer, but complementary, typology of [Manski \(2015\)](#) that distinguishes transitory, permanent and conceptual uncertainties. The importance of nonsampling or permanent uncertainties for administrative and big data sources is explained. Throughout, the article aims to emphasise the importance of measuring and then communicating these uncertainties, as hard as this can prove in some instances. Thereby, the article seeks to introduce and motivate the *Comunikos* project at Eurostat and emphasise the need for more empirical research to minimise public misunderstanding and misuse of official statistics.

To stress once again the relevance of measuring and communicating uncertainty in official statistics, we emphasise the point made by [Manski \(2015, 2019\)](#) and [Van der Bles et al. \(2019\)](#): that reporting official statistics as point estimates projects *incredible certitude*. This may lead to sub-optimal decision making. In other words, this practice may

encourage users to treat statistics as known with certainty. Or they may then make their own, perhaps misleading, (private/subjective) estimates of the degree of uncertainty associated with the point estimates presented to them. In short, in the absence of a body of empirical research seeking to study the impact of data uncertainties on the public's and experts' use and interpretation of official statistics, it is impossible to say, with any confidence, if and how known and unknown data uncertainties do have an impact.

Moreover, as emphasised by [Morgenstern \(1950\)](#) and consistent with a more recent literature in econometrics (cf. [Granger and Pesaran 2000](#)), what surely matters when assessing the importance of uncertainty is how this uncertainty affects decisions. This calls for a joint analysis of how uncertainty matters for decisions made in specific contexts; that is, uncertainty cannot be really understood free from the context in which the uncertain data are used. It calls for empirical and experimental studies following [Kloprogge et al. \(2007\)](#) and [Van der Bles et al. \(2019\)](#) – like [Van der Bles et al. \(2018, 2020\)](#), [Manclossi and Ayodele \(2016\)](#), and [Galvao et al. \(2019\)](#) – that consider, for a given measure of uncertainty, how best this data uncertainty should be communicated.

6. References

- Abowd, J.M., and M.H. Stinson. 2013. “Estimating measurement error in annual job earnings: A comparison of survey and administrative data.” *The Review of Economics and Statistics* 95(5): 1451–1467. DOI: <http://dx.doi.org/10.2139/ssrn.1894690>.
- Aruoba, S.B., F.X. Diebold, J. Nalewaik, F. Schorfheide, and D. Song. 2016. “Improving GDP measurement: A measurement-error perspective.” *Journal of Econometrics* 191(2): 384 – 397. DOI: <http://dx.doi.org/10.1016/j.jeconom.2015.12.009>.
- Australian Bureau of Statistics. 2013. “Types of error.” Available at: [https://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical + language + - + types + of + error](https://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+types+of+error) (accessed December 2019).
- Bagus, P. 2011. “Morgenstern’s forgotten contribution: A stab to the heart of modern economics.” *American Journal of Economics and Sociology* 70(2): 540–562. DOI: <https://doi.org/10.1111/j.1536-7150.2011.00783.x>.
- Bank of England. 2007. *Inflation Report, November 2007*. Available at: <https://www.bankofengland.co.uk/-/media/boe/files/inflation-report/2007/november-2007.pdf?la=en&hash=03AA4FB7EE5CA613952B392EFB00411A7D10195A> (accessed December 2019).
- Bernhard, S. 2016. “A real-time GDP data set for Switzerland.” *SNB Economic Studies* 9/2016. Available at: https://www.snb.ch/n/mmr/reference/economic_studies_2016_09/source/economic_studies_2016_09.n.pdf (accessed December 2019).
- Biemer, P.P., and L.E. Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, NJ: John Wiley Sons. DOI: <https://doi.org/10.1002/0471458740>.
- Boumans, M. 2012: “Observations in A Hostile Environment: Morgenstern on the Accuracy of Economic Observations.” *History of Political Economy* 44: 110–131. DOI: <https://doi.org/10.1215/00182702-1631806>.
- Burridge, P., and K.F. Wallis. 1985. “Calculating the variance of seasonally adjusted series.” *Journal of the American Statistical Association* 80(391): 541–552. DOI: <https://doi.org/10.1080/01621459.1985.10478151>.

- Chambers, R. 2014. "Survey sampling in official statistics - some thoughts on directions." *Proceedings of Statistics Canada Symposium 2014: Beyond traditional survey taking: adapting to a changing world*. Available at: <https://www.statcan.gc.ca/eng/conferences/symposium2014/program/14251-eng.pdf> (accessed December 2019).
- Chow, G., and A.-I. Lin. 1971. "Best linear unbiased interpolation, distribution, and extrapolation of time series by related series." *Review of Economics and Statistics* 53(4): 372–75. DOI: <https://doi.org/10.2307/1928739>.
- Clements, M.P., and A.B. Galvão. 2010. "First announcements and real economic activity." *European Economic Review* 54: 803–817. DOI: <https://doi.org/10.1016/j.eurocorev.2009.12.010>.
- Conti, P.L., D. Marella, and M. Scanu. 2012. "Uncertainty analysis in statistical matching." *Journal of Official Statistics* 28: 69–88. Available at: <https://www.scb.se/contentassets/ca21efb41fee47d293bbe5bf7be7fb3/uncertainty-analysis-in-statistical-matching.pdf> (accessed December 2019).
- Croushore, D. 2011. "Frontiers of real-time data analysis." *Journal of Economic Literature* 49: 72–100. DOI: <https://doi.org/10.1257/jel.49.1.72>.
- Croushore, D., and T. Stark. 2001. "A real-time data set for macroeconomists." *Journal of Econometrics* 105(1): 111–130. DOI: [https://doi.org/10.1016/s0304-4076\(01\)00072-0](https://doi.org/10.1016/s0304-4076(01)00072-0).
- Croushore, D., and T. Stark. 2003. "A real-time data set for macroeconomists: Does the data vintage matter?" *The Review of Economics and Statistics* 85: 605–617. DOI: <https://doi.org/10.1162/003465303322369759>.
- Cunningham, A., J. Eklund, C. Jeffery, G. Kapetanios, and V. Labhard. 2012. "A state space approach to extracting the signal from uncertain data." *Journal of Business and Economic Statistics* 30: 173–80. DOI: <https://doi.org/10.2139/ssrn.1080189>.
- Cunningham, A. and C. Jeffery. 2007. "Extracting a better signal from uncertain data." *Bank of England Quarterly Bulletin 2007Q3*. Available at: <https://www.bankofengland.co.uk/-/media/boe/files/quarterly-bulletin/2007/extracting-a-better-signal-from-uncertain-data.pdf?la=en&hash=58933100C07CD43ED04DC670F9D6721154AD4F87> (accessed December 2019).
- De Jonge, E. (2020). "Communicating uncertainties in official statistics – A review of communication methods." Available at: <https://ec.europa.eu/eurostat/documents/3888793/12135507/KS-TC-20-011-EN-N.pdf/b2ee60fe-ed5b-ad3e-05bd-56f6b87dfec0?t=1608645519272> (accessed May 2021).
- Draper, D. 1995. "Assessment and propagation of model uncertainty." *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1): 45–97. DOI: <https://doi.org/10.1111/j.2517-6161.1995.tb02015.x>.
- Dungan, J., D. Gao, and A. Pang. 2002. "Definitions of uncertainty." Available at: <ftp://ftp.cse.ucsc.edu/pub/reinas/papers/white.pdf> (accessed December 2019).
- European Commission. 2010. "Report on Greek government deficit and debt statistics." Available at: https://ec.europa.eu/eurostat/documents/4187653/6404656/COM_2010_report_greek/c8523cfa-d3c1-4954-8ea1-64bb1e59b3a (accessed October 2020).
- Eurostat. 2015. "Eurostat statistics explained." Available at: <https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Accuracy> (accessed December 2019).
- Eurostat. 2017. "European statistics code of practice: revised edition 2017." Available at: <https://ec.europa.eu/eurostat/web/quality/principle12> (accessed December 2019).

- Eurostat. 2019. "ESS Standard for Quality Reports." Available at: https://unstats.un.org/unsd/dnss/docs-nqaf/eurostat-esqr_final.pdf (accessed December 2019).
- Faust, J., J.H. Rogers, and J. H. Wright. 2005. "News and noise in G-7 GDP announcements." *Journal of Money, Credit and Banking* 37(3): 403–420. DOI: <https://doi.org/10.1353/mcb.2005.0029>.
- Frare, C., M. Marcellino, G.L. Mazzi, and T. Proietti. 2011. "EUROMIND: a monthly indicator of the Euro Area economic conditions." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2): 439–470. DOI: <https://doi.org/10.1111/j.1467-985x.2010.00675.x>.
- Fuller, W. 1987. *Measurement Error Models*. John Wiley and Sons: New York. DOI: <https://doi.org/10.1002/jae.3950030407>.
- Galvao, A.B. 2017. "DSGE models and data revisions." *Journal of Econometrics* 196: 215–232. DOI: <https://doi.org/10.1016/j.jeconom.2016.09.006>.
- Galvao, A. and J. Mitchell. 2020. *Real-Time Perceptions of Historical GDP Data Uncertainty, EMF Research Papers 35, Economic Modelling and Forecasting Group, Warwick Business School*. Available at: <https://ideas.repec.org/p/wrk/wrkemf/35.html> (accessed September 2020).
- Galvao, A. J. Mitchell, and J. Runge. 2019. "Communicating data uncertainty: experimental evidence for UK GDP." Economic Statistics Centre of Excellence Discussion Paper 2019-20. Available at: <https://EconPapers.repec.org/RePEc:nsr:escoe:escoe-dp-2019-20> (accessed December 2019).
- Geweke, J. 1978. "The temporal and sectoral aggregation of seasonally adjusted time series." In *Seasonal Analysis of Economic Time Series* edited by A. Zellner. Washington, DC: US Department of Commerce, Census Bureau. Available at: <https://www.census.gov/ts/papers/Conference1978/Geweke1978.pdf> (accessed December 2019).
- Ghysels, E. 1997. "Seasonal adjustment and other data transformations." *Journal of Business and Economic Statistics* 15: 410–418. DOI: <https://doi.org/10.1080/07350015.1997.10524719>.
- Giannone, D., J. Henry, M. Lalik, and M. Modugno. 2012. "An Area-Wide Real-Time Database for the Euro Area." *Review of Economics and Statistics* 94: 1000–1013. DOI: https://doi.org/10.1162/rest_a_00317.
- Goedeme, T. 2013. "How much Confidence can we have in EU-SILC? Complex Sample Designs and the Standard Error of the Europe 2020 Poverty Indicators." *Social Indicators Research* 110(1): 89–110. DOI: <https://doi.org/10.1007/s11205-011-9918-2>.
- Granger, C.W.J., and M.H. Pesaran. 2000. "Economic and statistical measures of forecast accuracy." *Journal of Forecasting* 19: 537–560. DOI: [https://doi.org/10.1002/1099-131x\(200012\)19:7%3C537:aid-for769%3E3.3.co;2-7](https://doi.org/10.1002/1099-131x(200012)19:7%3C537:aid-for769%3E3.3.co;2-7).
- Gromme, F. Ustek-Spilda, E. Ruppert, and B. Cakici. 2017. "Citizen data and official statistics: Background document to a collaborative workshop." Available at: <https://pdfs.semanticscholar.org/4bc0/6a6504c5086a2e81a299abd95ef42617b77b.pdf> (accessed December 2019).
- Groves, R.M., and L. Lyberg. 2010. "Total survey error: past, present and future." *Public Opinion Quarterly* 74: 849–879. DOI: <https://doi.org/10.1093/poq/nfq065>.

- Hand, D.J. 2018. “Statistical challenges of administrative and transaction data.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181(3): 555–605. DOI: <https://doi.org/10.1111/rssa.12315>.
- Jacobs, J.P.A.M. and S. van Norden. 2011. “Modeling data revisions: Measurement error and dynamics of “true” values.” *Journal of Econometrics* 161: 101–109. DOI: <https://doi.org/10.1016/j.jeconom.2010.04.010>.
- Kapetanios, G., M. Marcellino, F. Kempf, G.-L. Mazzi, J. Eklund, and V. Labhard. 2021. “Measuring and communicating uncertainty in official statistics: State of the art and perspectives.” Statistical working papers collection, Eurostat. Available at: <https://ec.europa.eu/eurostat/cros/system/files/ks-tc-20-006-en-n.pdf> (accessed May 2021).
- Kapteyn, A., and J.Y. Ypma. 2007. “Measurement error and misclassification: A comparison of survey and administrative data.” *Journal of Labor Economics* 25(3): 513–551. DOI: <https://doi.org/10.1086/513298>.
- Kishor, N.K., and E.F. Koenig. 2012. “VAR estimation and forecasting when data are subject to revision.” *Journal of Business and Economic Statistics* 30: 181–190. DOI: <https://doi.org/10.1198/jbes.2010.08169>.
- Klopogge, P., J. van der Sluijs, and A. Wardekker. 2007. “*Uncertainty Communication*, Utrecht: Copernicus Institute.” Available at: http://www.nusap.net/downloads/reports/uncertainty_communication.pdf (accessed December 2019).
- Koch, G.G., and D.B. Gillings. 2006. *Inference, Design-Based vs. Model-Based*. American Cancer Society. DOI: <https://doi.org/10.1002/0471667196.ess1235.pub2>.
- Kuznets, S. 1948. “National income: a new version.” *Review of Economics and Statistics* 30(3): 151–179. DOI: <https://doi.org/10.2307/1926746>.
- Kuznets, S. 1950. “Review of *On the accuracy of economic observations*, by Oscar Morgenstern.” *Journal of the American Statistical Association* 45: 576–79. DOI: <https://doi.org/10.2307/2280732>.
- Laureti, T., and I. Benedetti. 2020. “Measuring and communicating uncertainty of poverty indicators at regional level.” Statistical working papers collection, Eurostat. Available at: <https://ec.europa.eu/eurostat/documents/3888793/12137895/KS-TC-20-010-EN-N.pdf/6745684c-c989-b3e5-33ae-7bd7dd89bf92?t=1608634421113> (accessed May 2021).
- Lui, S., J. Mitchell, and M. Weale. 2011. “Qualitative business surveys: signal or noise?” *Journal of the Royal Statistical Society Series A* 174(2): 327–348. DOI: <https://doi.org/10.1111/j.1467-985x.2010.00667.x>.
- Manclossi, S., and V. Ayodele. 2016. “Users understanding of uncertainty measures to describe data quality.” Presented at The 21st Government Statistical Service (GSS) Methodology Symposium Methodology: the key to the door of innovation. Available at: <https://gss.civilservice.gov.uk/wp-content/uploads/2016/07/4.4.3-Silvia-Manclossi-Users-understanding-and-use-of-uncertainty-measures-to-descripdf> (accessed December 2019).
- Mankiw, N.G., and M.D. Shapiro. 1986. “News or noise: An analysis of GNP revisions.” *Survey of Current Business (May 1986)*, US Department of Commerce, Bureau of Economic Analysis: 20–25. DOI: <https://doi.org/10.3386/w1939>.

- Manski, C.F. 2015. "Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern." *Journal of Economic Literature* 53(3): 631–653. DOI: <https://doi.org/10.1257/jel.53.3.631>.
- Manski, C.F. 2016. "Credible interval estimates for official statistics with survey nonresponse." *Journal of Econometrics* 191(2): 293–301. DOI: <https://doi.org/10.1016/j.jeconom.2015.12.002>.
- Manski, C.F. 2019. "Communicating uncertainty in policy analysis." *Proceedings of the National Academy of Sciences* 116(16): 7634–7641. DOI: <https://doi.org/10.1073/pnas.1722389115>
- McCracken, M.W., and S. Ng. 2016. "FRED-MD: A Monthly Database for Macroeconomic Research." *Journal of Business and Economic Statistics* 34(4): 574–589. DOI: <https://doi.org/10.1080/07350015.2015.1086655>.
- McKenzie, R. (2006), "Undertaking revisions and real-time data analysis using the OECD main economic indicators original release data and revisions database." OECD Statistics Working Papers, 2006/02, OECD Publishing, Paris.
- Mehrhoff, J. 2008. "Sources of revisions of seasonally adjusted real time data." Paper prepared for the Meeting of the OECD Short-term Economic Statistics Working Party (STESWP), 23–24 June 2008, Paris. France. Available at: <http://www.oecd.org/sdd/fin-stats/40671433.pdf> (accessed December 2019).
- Mevik, A.-K. 2004. "Uncertainty in the Norwegian Business Tendency Survey." Available at: https://www.ssb.no/a/english/publikasjoner/pdf/doc_200410_en/doc_200410_en.pdf (accessed December 2019).
- Mitchell, J., R. Smith, M. Weale, S. Wright, and E.L. Salazar. 2005. "An Indicator of Monthly GDP and an Early Estimate of Quarterly GDP Growth." *Economic Journal* 115(501): 108–129. DOI: <https://doi.org/10.1111/j.0013-0133.2005.00974.x>.
- Morgenstern, O. 1950. *Introduction to survey quality*, Princeton University Press.
- National Science Foundation 2018. "Data accuracy." Available at: <https://nsf.gov/statistics/2018/nsb20181/report/sections/appendix-methodology/data-accuracy> (accessed December 2019).
- Oberski, D.L., A. Kirchner, S. Eckman, and F. Kreuter. 2017. "Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models." *Journal of the American Statistical Association* 112(520): 1477–1489. DOI: <https://doi.org/10.1080/01621459.2017.1302338>.
- OECD 2015. "Revisions of quarterly GDP in selected OECD countries." *Statistics Brief—July 2015 – No 22 by Zwijnenburg*. Available at: <https://www.oecd.org/sdd/na/Revisions-quarterly-GDP-selected-OECD-countries-OECD22.pdf> (accessed December 2019).
- ONS 2017a. "Methodology for measuring uncertainty in ONS local authority midyear population estimates: 2012 to 2016." Available at: <https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/methodologyformeasuringuncertaintyinonslocalauthoritymidyearpopulationestimates20> (accessed December 2019).
- ONS 2017b. "ONS methodology working paper series no. 9 – Guide to calculating standard errors for ONS Social Surveys." Available at: <https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/ons->

- methodologyworkingpaperseriesno9guidetocalculatingstandarderrorsforonssocialsu (accessed December 2019).
- ONS 2018a. “Communicating quality, uncertainty and change.” Available at: <https://gss.civilservice.gov.uk/policy-store/communicating-quality-uncertainty-and-change/> (accessed December 2019).
- ONS 2018b. “VAT turnover data in National Accounts: background and methodology.” Available at: <https://www.ons.gov.uk/economy/grossdomesticproductgdp/methodologies/vatturnoverdatainnationalaccountsbackgroundandmethodology> (accessed December 2019).
- ONS 2019. “Gross domestic product (GDP) QMI.” Available at: <https://www.ons.gov.uk/economy/grossdomesticproductgdp/methodologies/grossdomesticproductgdpqmi> (accessed December 2019).
- Orphanides, A., and S. van Norden. 2002. “The unreliability of output-gap estimates in real time.” *The Review of Economics and Statistics* 84: 569–583. DOI: <https://doi.org/10.1162/003465302760556422>.
- Peters, C.A. 2001. “Statistics for Analysis of Experimental Data.” In *Environmental Engineering Processes Laboratory Manual*, edited by S.E. Powers. Champaign, IL: AEESP. Available at: https://www.princeton.edu/~cap/AEESP_Statchap_Peters.pdf (accessed December 2019).
- Piccolo, D., and R. Simone, 2019. “The class of cub models: statistical foundations, inferential issues and empirical evidence.” *Statistical Methods and Applications* 28: 389–435. DOI: <https://doi.org/10.1007/s10260-019-00461-1>.
- Plumper, T., and E. Neumayer. 2012. “Population and sample uncertainty.” *EPSA 2013 Annual General Conference Paper 166*. Available <https://ssrn.com/abstract=2224688> (accessed December 2019).
- Ruggeri Cannata, R., and G.-L. Mazzi. 2017. “The trade-off between timeliness and accuracy: the perspective of a statistical agency.” In *Handbook on Rapid Estimates*, edited by G.L. Mazzi. Eurostat and United Nations, : 123–131. DOI: <https://doi.org/10.2785/317290>.
- Seljak, R. 2006. “Estimation of standard error of indices in the sampling business surveys.” *Proceedings of Q2006 European Conference on Quality in Survey Statistics*. Available at: <https://ec.europa.eu/eurostat/documents/64157/4374310/16-ESTIMATION-OF-STANDARD-ERROR-OF-INDICES-IN-THE-SAMPLING.pdf/5721f63b-9ac8-4767-b557-dd30b0378336> (accessed December 2019).
- Smith, R.J., M.R. Weale, and S.E. Satchell. 1998. “Measurement Error with Accounting Constraints: Point and Interval Estimation for Latent Data with an Application to U.K. Gross Domestic Product.” *Review of Economic Studies* 65(1): 109–134. DOI: <https://doi.org/10.1111/1467-937x.00037>.
- Statistics Canada 2017. “Non-sampling error.” Available at: <https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch6/nse-endaie/5214806-eng.htm> (accessed December 2019).
- Stone, R., D.G. Champernowne, and J. Meade. 1942. “The precision of national income estimates.” *Review of Economic Studies* 9(2): 111–125. DOI: <https://doi.org/10.2307/2967664>.

- US Census Bureau 2018. "Measures of nonsampling error (chapter 11)." Available at: <https://www.census.gov/ts/papers/Conference1978/Geweke1978.pdf> (accessed December 2019).
- Van der Bles, A.M., D. Spiegelhalter, S. Dryhurst, A. Freeman, M. Pearson, and J. Park. 2018. "Determining and facilitating the clearest ways to visualize uncertainty around estimates, time series and curves." Report to M2D. Available at: <https://wintoncentre.maths.cam.ac.uk/projects/communicating-uncertainty/> (accessed June 2020).
- Van der Bles, A.M., S. van der Linden, A.L.J. Freeman, and D.J. & Spiegelhalter. 2020. "The effects of communicating uncertainty on public trust in facts and numbers." *Proceedings of the National Academy of Sciences* 117(14): 7672–7683. DOI: <https://doi.org/10.1073/pnas.1913678117>.
- Van der Bles, A.M., S. van der Linden, A. Freeman, J. Mitchell, A.B. Galvao, L. Zaval, and D. Spiegelhalter. 2019. "Communicating uncertainty about facts, numbers and science." *Royal Society Open Science* 6: 1–42. DOI: <https://doi.org/10.1098/rsos.181870>.
- Van der Loo, J. Pannekoek, and L. Rijnveld. 2017. "Computational estimates of data-editing related variance." United Nations Economic Commission for Europe, Conference of European Statisticians, Work Session on Statistical Data Editing (The Hague, Netherlands, 24/26 April 2017). Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2017/mtg2/Paper_7_DataEditingBootstrap.pdf (accessed December 2019).
- Verma, V., G. Betti, and F. Gagliardi. 2010. "An assessment of survey errors in EU-SILC." *Eurostat: Methodologies and working papers*. Available at: <https://ec.europa.eu/eurostat/documents/3888793/5848229/KS-RA-10-021-EN.PDF/2b5b31ad-3973-48ad-9a87-8645c1d2c16c> (accessed December 2019).
- Wallis, K.F. 1982. "Seasonal adjustment and revision of current data: linear filters for the X-11 method." *Journal of the Royal Statistical Society A* 145: 74–85. DOI: <https://doi.org/10.2307/2981422>.
- Walton, A. 2016. "International comparison of GDP revisions." *Office for National Statistics* : 1–34. Available at: <https://www.ons.gov.uk/economy/grossdomesticproductgdp/articles/internationalcomparisonofgdprevisions/2016-04-28> (accessed December 2019).

Received June 2019

Revised December 2019

Accepted November 2020

The Evolution of the Italian Framework to Measure Well-Being

Fabio Bacchini¹, Barbara Baldazzi¹, Rita De Carli¹, Lorenzo Di Biagio¹, Miria Savioli¹, Maria Pia Sorvillo¹, and Alessandra Tinto¹

Recently, a new approach for measuring well-being was developed by eighteen European countries in the wake of the “Beyond GDP movement” started in the 1990 and continued by the Stiglitz Commission. Among these European economies, eleven of them use measures of well-being for monitoring public policy. The Italian Statistical Institute (Istat) jointly with the National Council for Economics and Labor (CNEL) developed a multi-dimensional framework for measuring “equitable and sustainable well-being” (Bes) and since 2013 Istat publishes an annual report on well-being. The Bes framework is continuously updated to take into account new challenges: the exploitation of new data sources, to produce better indicators; new ways for making the communication more effective and foster public awareness; the inclusion of well-being indicators in the budget documents, as established by law. Especially for the latter, the Italian Bes can be considered a forerunner and, more generally, the Italian experience is one of the most relevant at the European level, showing potential of become a benchmark for other countries. This article illustrates the development of the Italian Bes, focusing on its recent progresses and challenges.

Key words: Multidimensional well-being; beyond GDP; SDGs; composite indices; policy evaluation; official statistics.

1. Introduction

As reported by M. Wolf (Financial Times, 30 May 2019) in March 1809, leaving the US presidency, Thomas Jefferson wrote that “the care of human life and happiness, and not their destruction, is the first and only legitimate object of good government”. Echos of this brilliant intuition are scattered across history, recall Bob Kennedy’s famous speech “GDP measures everything except that which is worthwhile”. But only in recent years has it been translated into a suitable set of indicators useful for setting and monitoring the policy agenda.

Along this path an important step was made in 1990 when the United Nations launched the Human Development Report (HDR), which laid the foundations for the definition and measurement of the concept of development, embracing non-income related dimensions (UNDP 1990). The HDR laid on the “capability approach” of Amartya Sen and Martha Nussbaum, focusing not on how much a nation produces, but how people who live there

¹ Italian National Institute of Statistics (Istat), Via Cesare Balbo, 16 – 00184 Rome, Italy. Emails: bacchini@istat.it, baldazzi@istat.it, decarli@istat.it, lorenzo.dibiagio@istat.it, savioli@istat.it, and tinto@istat.it.

Acknowledgments: We are grateful to the editor of this special issue and the two anonymous reviewers for helpful comments and suggestions. We are grateful for the support from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 770643 MAKSWELL.

are doing (Sen 1989; Nussbaum and Sen 1993). Well-being, poverty and inequality must be assessed in the space of capacity, that is the real opportunities that people have to live the life they value. Economic resources and material goods should be understood as means for the realization of functioning, that is people's real achievements. The approach is people-centered: the individual is an end and a means for development. In the same period the World Bank's World Development Report introduced an international poverty line based on 'a dollar a day' and identified just over a billion people – a fifth of the world population at the time – as living in extreme poverty (World Bank 1990).

Ten years later, in 2000, the Millennium Development Goals (MDGs) were proposed by the UN, combining the stimulus of previous years' UN conferences and the OECD's ambition to agree on an international set of indicators to measure development progress and well-being. Later on in 2007, when the "Istanbul Declaration" (OECD 2007) was signed by the UN, the World Bank and the European Commission to highlight the need to measure the progress of societies going "beyond the GDP", several initiatives aimed at measuring well-being with economic, social and environmental statistical indicators were already in place worldwide. Finally, on September 2015 the UN General Assembly adopted the 2030 Agenda for Sustainable Development, in which the Global Goals strive to end poverty, protect the planet and ensure prosperity for all (UN-DESA 2016).

In this long-standing effort, the work of the Stiglitz Commission (Stiglitz et al. 2009) represented another important pillar, stressing how important it is to propose well-being as a multidimensional phenomenon, with different dimensions measured on a micro or macro population level (i.e., households, regions, countries) across time. In line with the proposal presented by the Stiglitz Commission, the OECD first developed a framework for measuring well-being in 2011, as part of the broader Better Life Initiative (OECD 2013). The OECD's initiative inspired and urged several national statistical offices (NSOs), government departments and international organizations to set structured initiatives for the measurement of well-being through extensive collection of social, environmental and economic indicators.

The use of well-being indicators to shape and assess public policies is a further step along the road that has been implemented in a heterogenous way across countries, as OECD (Exton and Shinwell 2018) and the Horizon 2020 project MAKSWELL (Making Sustainable development and WELL-being frameworks work for policy, Tinto et al. (2018) have documented.

Among the international experience, the Italian initiative named "equitable and sustainable well-being" Benessere equo e sostenibile (*Bes*), can be considered as a forerunner presenting a measurement system up to date and a notable example for the inclusion of well-being indicators in the budget documents. *Bes*, initially run by the Italian National Institute of Statistics (Istat) together with the National Council for Economics and Labor (CNEL), refers to 130 indicators, available at national and regional level, organized across 12 domains. The indicators are updated two times a year, once jointly with the dissemination of the annual report on well-being (in the 7th edition, Istat 2019). The *Bes* framework is continuously updated to take into account new challenges such as the exploitation of new data sources, to produce better indicators. Most importantly, the most recent Budget law (L. 163/2016) assigned to the well-being indicators to measure the way in which public policy tries to foster citizens' well-being. Italy is the first country in

the European Union to include well-being as one of the objectives of the Government's economic and social policy. Istat contributed to the selection of the subset of 12 Bes indicators that are now included in the budget documents, accounting for its updating and real-time estimation.

The increasing importance of well-being indicators in the political debate requires new ways for making the communication more effective and fostering public awareness. Bes has developed different strategies for dissemination. Presentation of the data is based on an in-depth analysis of the dashboard of several indicators. This approach is considered best suited for studying complex multidimensional phenomena (Stiglitz et al. 2018b). At the same time, Istat has experimented with the use of composite indices, aggregating indicators by domain, in line with the wake of other international experiences, and especially of *Human Development Index* (HDI) by the UN (UNDP 1990, 2010). Throughout the text, in accordance to the existing literature (Saisana and Tarantola 2002; Harvey 2020), we will refer to an individual indicator (something that measures a specific concept) as an "indicator", while we will refer to an aggregated (composite) index (a single score made by mathematically combining several other scores) as an "index". All in all, the Bes framework can be considered as a reference point for public debate and a good example to be considered at national and international level.

This article aims to share the experience of the evolution of the Bes, focusing on its recent progresses in dissemination and its practical use to improve the discussion on the budget plan. In particular, Section 2 will describe the current framework for measuring well-being in Italy, starting from a historical account and later concentrating on the maintenance and development process. Section 3 will explain how Istat is addressing the issue of the communication of the results and the different approaches adopted, with pros and cons. Section 4 will analyze the use of well-being indicators for policy evaluation and Section 5 will conclude.

2. The Current Framework for Measuring Well-Being in Italy

2.1. The Beyond GDP Movement in Italy

In 2003 the network "Sbilanciamoci!", comprising 49 Italian organizations and civil society networks working on public spending and economic policy alternatives, proposed a composite index to measure development, the Quality of Regional Development Index (QUARS). The proposed framework was set up along seven dimensions of development at regional level (Environment, Economy and Labor, Rights and Citizenship, Health, Education and Culture, Equal Opportunities, Participation). QUARS compared the "quality of development" at regional level (Lazio and Piemonte), and provincial level (provinces of Trento and Ascoli Piceno) and at municipal level (Arezzo and Cascina). The proposal of seven dimensions represented a novelty in the Italian debate, while the comparison of different territories using also social and environmental indicators stemmed from a previous experience launched years earlier by the national daily business newspaper "Il Sole 24 Ore" (Il Sole 24 ore 2019).

Other experiences reinforced the attention on social and environmental issues. In the 1990s Legambiente (*League for the Environment*) and Ambiente Italia (*Environment*

Italy) started to publish the “Urban Ecosystem”, a summary index on the environmental quality of the provincial capitals. In 2010, Confartigianato, a confederation of artisans and small enterprises, composite index based on information related to quality of life, cultural resources and the environment, together with GDP.

Istat contributed to this research effort to expand expanding the availability and use of social indicators in Official Statistics (Sabbadini and Maggino 2018) by means of the introduction of the “Multipurpose Survey System” in the 1990s. This step was important because it gave the Italian community the opportunity to reinforce the dissemination of social indicators even before the Istanbul Declaration, which stressed the important role of official statistics as a key provider of data useful for monitoring the progress of societies.

2.2. The Bes Project

In the wake of international and national experiences, Istat, together with the National Council for Economics and Labor (CNEL), launched an inter-institutional initiative in December 2010 aimed at developing a multi-dimensional approach for the measurement of “equitable and sustainable wellbeing” (Bes). The proposal was in line with the recommendations issued by the OECD and the Stiglitz Commission (Stiglitz et al. 2009).

The project has been characterized by a participatory process, involving civil society, academia and national experts. All of them were involved in the definition of the framework and in the selection of indicators. This approach led to a wide acceptance of the framework.

As a first step, the attention was focused on defining the Italian well-being, searching for its most notable dimensions. To perform this task, a joint “Steering Group on the Measurement of Progress in Italian Society” was set up. The Steering Group included representatives from enterprises, professional associations, trade unions, environmental groups such as WWF and Legambiente, Italian cultural heritage groups, women’s groups, consumer protection groups and the civil society network. At the same time, between October 2011 and February 2012, Italian citizens were asked for their opinion on the dimensions of well-being elaborated by the Steering Committee, through a dedicated website that included both a short questionnaire and a blog. The questionnaire was filled in by 2,518 people on a voluntary basis. In the same period, a further extensive consultation was set up using the Multipurpose Survey *Aspects of daily life* that reached 45,000 people aged 14 years and over, representative of the population resident in Italy. Respondents were asked to give a score from 0 to 10 to a list of 15 dimensions of well-being.

The results of the consultations and the evidence coming from international experience were the input for the Steering Group to the definition of the domains. The following 12 domains were identified: health; education and training; work and life balance; economic well-being; social relationship; safety; landscape and cultural heritage; environment; subjective well-being; politics and institutions; research and innovation; quality of services.

As a second step of the process, indicators were selected to be included in each domain. For this process Istat started up a Scientific Committee with more than 80 experts in different domains of well-being. The selection of indicators was a crucial step, in the sense that “what we measure” affects “what we do” (Stiglitz et al. 2009). The following criteria were used for the selection:

- form part of the Official Statistics,
- time series availability (starting from 2004),
- sub-national availability (Italian regions – NUTS2),
- clear interpretation of the relationship amid indicator and the well-being evolution (UN–IAEG-MDG 2013),
- both objective and subjective measures were included,
- attention to international comparisons.

The activity of the Scientific Committee on the production of new indicators questions in pre-existing surveys. For instance, questions on trust in institutions and questions on perception of landscape and environment were added in the annual multipurpose survey on *Aspects of daily life*. Through this process, 134 indicators were identified as representing the 12 domains of well-being.

According to the definition of the well-being framework, attention was focused on identifying equity amid social groups and geographic areas of the country, and sustainability for future generations.

Equity and sustainability are cross-cutting characteristics related to all dimensions of well-being. Measuring equity leads to focus on the distribution of well-being across regions, socio-economic groups, gender and age, while the concept of sustainability mainly relates to inter-generational comparison.

The importance of equity and sustainability for the definition of well-being is addressed by international literature. Since the 1980s Amartya Sen has stated the need to consider inequality on information more closely related to living standards (Sen 1989). Furthermore, the recommendations defined by the Stiglitz Commission (Stiglitz et al. 2009) stated that considering inequalities in human conditions is essential for assessing quality of life across countries and social groups. The *How's Life?* report (OECD 2011) dedicated special attention to inequalities as a central element in wellbeing assessment, providing a valuable presentation of multidimensional inequalities related to every dimension. In 2018, the importance of measuring equity and sustainability was further reaffirmed by the OECD-hosted High-Level Group on the Measurement of Economic Performance and Social Progress (HLEG) (Stiglitz et al. 2018a).

These suggestions were identified in the Italian Bes, which presented and analyzed indicators by regions (NUTS2), gender and age groups. At the same time, specific indicators were included in the framework to take into account sustainability for future generations.

2.3. *The Development and Maintenance of the Bes Framework*

After the definition of the framework and the first release of the Istat report in 2013 (Istat 2013), the agenda then focused on the development and maintenance of the framework adopted to measure well-being. The division in charge of well-being has planned an annual review of the indicators, taking into account new information needs and new data sources. The proposed were first discussed with the Scientific Committee and, after the end of its mandate, with the Commission of Users of Statistical Information (CUIS) and with experts in the field. This approach makes it possible to maintain the original inspiration of Bes as a common tool within the community. Following this procedure in 2017, a broad revision of the set of indicators was carried out to improve timeliness and to strengthen the

structure of the Landscape and Cultural Heritage, Research and Innovation and Quality of services domains.

From the beginning, Landscape and Cultural Heritage was characterized — more than other domains — by an exploratory approach. It was not one of the domains proposed by the Stiglitz Report and it remains a unique case in the international panorama of well-being statistics. However, some of the indicators of the original set described macro-trends or context factors, and proved to be quite invariant or unobservable in the short term, hardly fitting for a yearly report. Therefore, the revision carried out in 2017 led to an improvement of the domain by strengthening the representation of short-term trends and identifying new indicators. This led to a substantial reorganisation in which five of the original indicators were discontinued due to quality issues. Four new indicators were introduced, either to replace the discontinued ones (Density and importance of museum heritage for Endowment of cultural heritage; Spread of rural tourism facilities for Quality assessment of Regional programs for rural development), or to expand the coverage of the concept map (Impact of forest fires and Pressure of mining and quarrying activities, referred to the component of the natural landscape).

A general innovation introduced in the Environment domain concerns the way indicators are organized and analyzed, according to the categories of the DPSIR model Driving Force, Pressure, State, Impact, Response (Figure 1). The conceptual scheme breaks down the relationship between the natural system and the anthropic system into successive phases, connected to each other through a causal circuit. Following this pattern, innovations were introduced in order to improve territorial representativeness, to merge indicators that provided information on different aspects of the same phenomenon, and to enrich the information provided on the stress exerted on water resources and waste management.

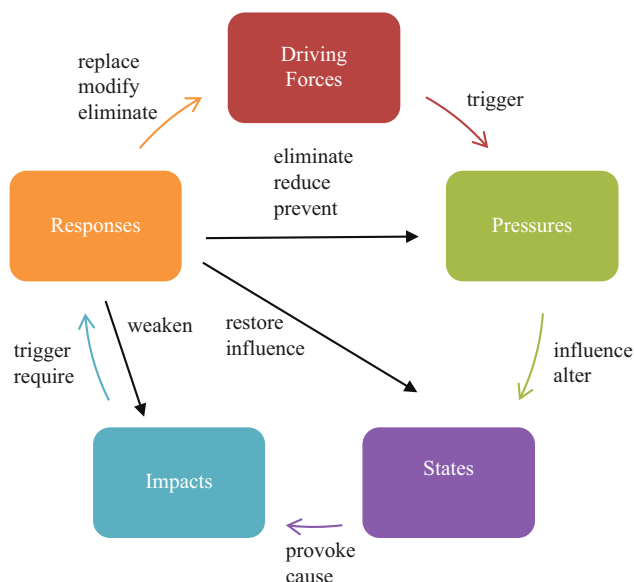


Fig. 1. The causal framework DPSIR “driving force, pressure, state, impact, response” for describing the interactions between society and the environment.

Source: elaboration on Istat (2017).

With regard to the Research and Innovation domain, it was reformulated with the aim of broadening the measurement of the different components that interact with innovation, leaving aside the aspects more closely related to the performance of firms, which are discussed in depth in other Istat publications (See Istat annual reports on competitiveness, <http://www.istat.it/it/competitivita>). The reformulation of the domain is characterized by three components:

1. the maintenance of the information on research and development, with the addition of a measure relating to investments in intellectual property, now included in the national accounts,
2. the introduction of an indicator on employment in cultural and creative industries, and
3. the inclusion of an indicator to measure the country's capacity to attract highly educated young people.

This approach was echoed in the work carried out by Eurostat within ESSnet-Culture, which proposed an estimate of cultural (and creative) employment based on the cross-reference between the classification of economic activities (NACE Rev. 2) and the classification of occupations (ISCO-08). In light of the increased focus on the cultural and creative component, the domain has been renamed "Innovation, research and creativity".

The revision of the Quality of Services domain identified weaknesses due to partial coverage of phenomena and lack of timeliness of some indicators. In order to strengthen its structure, the conceptual scheme was reorganized, at the same time taking into account different typologies of services (social services, infrastructure and mobility) and their main characteristics (Allocation and Accessibility; Effectiveness and Satisfaction). An analytical matrix was used for the analysis, which was also useful for the elaboration of the composite index, based on the coverage of each cell of the matrix.

In 2018, improvements were mainly related to checking for the relevance of the selected domains and to the multidimensional analysis. With regard to the first point, a set of questions was included in the Istat Consumer confidence survey to evaluate the importance of the 12 domains for measuring people's well-being and quality of life. Respondents were asked to evaluate each domain on a scale between 0 and 10 (See [Figure 2](#)). The results confirmed that all 12 domains are considered significant, with average marks between 7.4 (politics and institutions) and 9.5 (health). At the same time, a new section was introduced in the report to present analyses on the multidimensional characteristics of well-being. The first two contributions were devoted to the determinants of subjective well-being and to the vertical inequality.

Finally, in the last edition of 2019 of the report ([Istat 2019](#)) the analysis of indicators by region, gender and age group was accompanied by an analysis of indicators also by educational level, in order to enhance the evaluation of equity.

3. Dissemination of the Results: Dashboard and Composite Indices

Communicating the results is an important step to take into account in order to succeed in redirecting citizens' focus on well-being.

Many different and complementary approaches can be used in the analysis of well-being and in the dissemination of results. For instance, one can either opt for the analytical

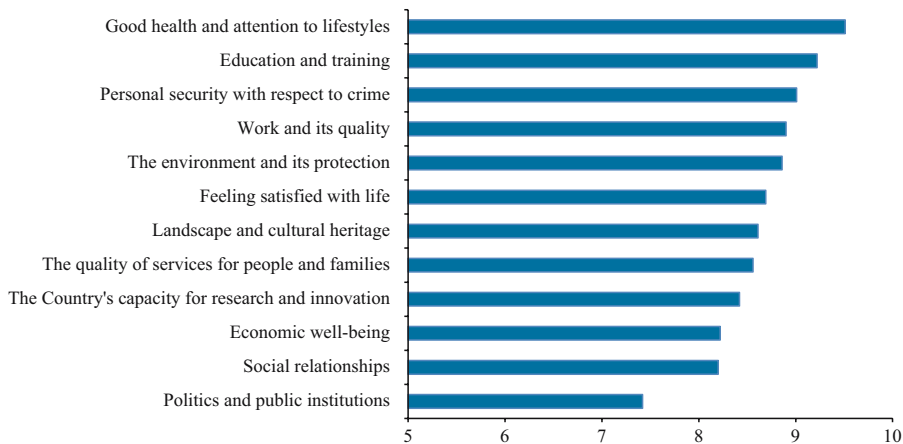


Fig. 2. Average score attributed to the Bes domains (between 0 and 10), Italy. Year 2018. Data from Istat Consumer confidence survey 2018.

comment of a dashboard of several individual indicators or the use of one or more composite indices aggregating conceptually unrelated indicators. While the former has the advantage of providing a detailed picture, it does not allow for easy public communication and easy comparisons across countries and over time (Ciommi et al. 2017; Bleys 2012). On the other hand, composite indices can be very useful for summarizing multi-dimensional realities, for supporting decision-makers, and for the dissemination of findings; even if complex concepts are very difficult (some says impossible) to capture with only one index (“different numbers are useful for different purposes, and local context is important in selecting which numbers matter for what”, Stiglitz et al. 2018b) – there is a long-lasting and never resolved scientific duel between aggregators and non-aggregators: for some hints read on, for an overview see Greco et al. (2019), for a detailed analysis see Sharpe (2004)).

For the Bes initiative (Section 2) Istat has primarily adopted the dashboard approach. Proposed indicators are presented, analyzed and commented yearly in a report on Equitable and Sustainable Well-being in Italy (Bes reports – eight reports published so far since 2013). For each of the 12 well-being domains, a specific chapter is devoted to the analysis of the level of indicators, their evolution over time and the comparison across regions, gender, age and level of education. Since the 2018 report (Istat 2018) Istat has decided to complement these analyses with immediate summary measures, based on the dashboard. We can get a glimpse of temporal trends by counting how many indicators have improved or deteriorated in the latest available year (Figure 3) providing an initial outlook of the evolution of well-being. For example, in 2018 in Italy over 50% of the 115 indicators for which comparison is feasible show an improvement in all areas of the country. Over the last year, in Italy, in the majority of domains over 50% of the indicators improved, while lower values are recorded in the domains Work and life balance (41.7%), Social Relationships (44.4%), Landscape and Cultural Heritage (44.4%) and Environment (46.7%) (Istat 2019). Moreover, we can get a glimpse of the overall representation of relative levels of well-being in the Italian regions by observing the distribution by quintiles

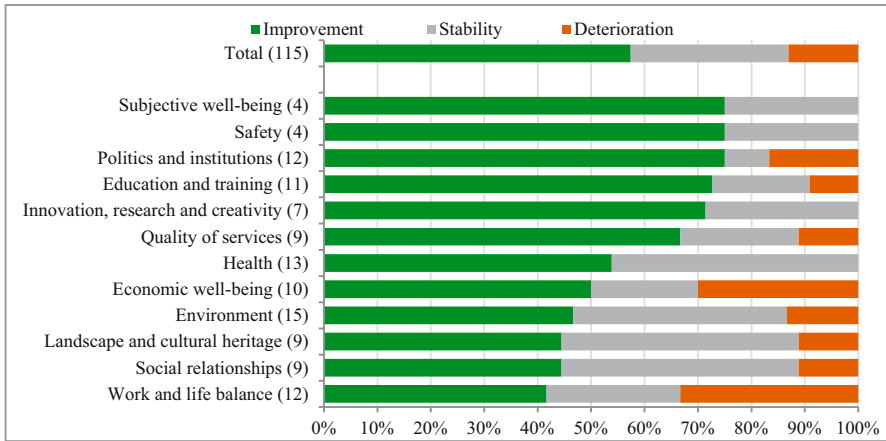


Fig. 3. Trend of Bes indicators: comparison between latest available year (in most cases 2018) and the previous one by domain. Percentage of total comparable indicators. Italy.

Source: Istat (2019).

of the indicators in the most recent available year (Figure 4). This figure shows how the geography of equitable and sustainable well-being reflects the traditional Italian territorial gradient, with the northern regions showing higher levels of well-being compared to the central and southern regions. The provinces of Bolzano and Trento have the highest levels of well-being, with 53.2% and 60.0% of the indicators in the highest quintile respectively, and less than 10% at the opposite extreme in the first quintile.

The lowest levels of well-being are recorded in Calabria and Sicilia, with 56.3% and 52.1% of the indicators falling in the first quintile, respectively (Istat 2019).

Yet, the dashboard approach (even if accompanied by these counting schemes) fails to fulfill the strong demand for a synthesis of all the data, while it is true that composite indices provide an easy tool to compare complex dimensions effectively, also over time,

REGIONS and GEOGRAPHICAL AREAS	Quintiles					Number of available indicators
	I (0-20)	II (20-40)	III (40-60)	IV (60-80)	V (80-100)	
Piemonte	8,4	18,5	28,6	31,9	12,6	119
Valle d'Aosta/Vallée d'Aoste	15,5	12,1	12,9	21,6	37,9	116
Liguria	15,0	20,8	22,5	25,8	15,8	120
Lombardia	12,6	13,5	15,1	22,7	36,1	119
Bolzano/Südtirol	9,9	10,8	11,7	14,4	53,2	111
Trento	9,1	0,9	12,7	17,3	60,0	110
Veneto	9,2	18,3	13,3	31,7	27,5	120
Friuli-Venezia Giulia	5,0	11,8	16,0	29,4	37,8	119
Emilia-Romagna	12,7	18,6	13,6	26,3	28,8	118
Toscana	8,3	12,5	32,5	30,0	16,7	120
Umbria	10,3	22,2	31,6	20,5	15,4	117
Marche	5,9	21,9	27,7	28,6	16,0	119
Lazio	21,7	35,0	18,3	13,3	11,7	120
Abruzzo	12,5	43,3	21,7	17,5	5,9	120
Molise	26,5	31,6	11,1	14,5	16,2	117
Campania	57,5	15,0	10,8	6,7	10,0	120
Puglia	38,3	30,8	12,5	9,2	9,2	120
Basilicata	31,9	31,1	16,0	6,7	14,3	119
Calabria	56,7	13,3	7,5	6,7	15,8	120
Sicilia	52,5	15,8	13,3	9,2	9,2	120
Sardegna	26,1	33,6	13,5	14,3	12,6	119
North	6,5	17,1	14,6	41,5	20,3	123
Center	9,8	26,2	40,2	18,0	5,7	122
South and Islands	48,4	24,6	13,1	11,5	2,5	122

Fig. 4. Bes indicators by region and quintile. Percentage distribution. Latest available year (in most cases 2018). Source: Istat (2019).

facilitating the communication with the general public and promoting accountability (Saisana and Tarantola 2002 and OECD and JRC 2008). They make it possible to measure multidimensional concepts in a way that is usually easier to interpret than finding common trends in many separate indicators. In fact, a complex concept is easier to communicate in the form of a unique number than in the form of an overabundance of indicators (Greco et al. 2019; Saltelli 2007; Stiglitz et al. 2018a). That is why composite indices have seen a general, steep growth in use and impact over the past two decades (Becker et al. 2017; Greco et al. 2019). That is why composite indices have seen a general, steep growth in use and impact over the past two decades (Becker et al. 2017; Greco et al. 2019), even if several criticisms from different angles are still raised against composite indices (Kuc-Czarnecka et al. 2020): first of all the reduction of information they induce is not always desirable (Ravallion 2011, 247; Stiglitz et al. 2018b, chap. 2); secondly, composite indices are considered problematic because their construction involves arbitrary assumptions that have to be carefully assessed, for example about the weighting procedure, which has strong implications but it is seldom justified (Saltelli 2007, Stiglitz et al. 2009), or because they are accused of not being based on sound (economic) theory (Ravallion 2010); thirdly, sometimes they are just examples of the abuse of metrics (Muller 2018; Saisana et al. 2011; Wilsdon 2016). The computation and use of composite indices was also discussed in the Bes Scientific Committee for the measurement of well-being (Section 2).

From a general perspective, composite indices for well-being should provide, in a consistent way, both spatial and temporal comparisons. Cardinal measures are usually more appropriate than counting measures for the measurement of well-being (Mauro et al. 2018), but there is not a well-established methodology to produce composite indices, and researchers have to deal with potentially difficult and problematic issues, such as standardization of variables, implicit weighting, management of substitutability rates. Actually composite indices could “differ in the dimensions and indicators selected, the transformations applied to the indicators, the assumed substitutability between indicators and the relative weights given to them” (Decancq and Lugo 2013, 3).

Furthermore, at the time of the Scientific Committee’s work, one of the most important composite index was the *Human Development Index* (HDI), developed by the UN (UNDP 1990, 2016). It currently uses a min-max normalization (with fixed goalposts) and an aggregation based on a simple geometric mean. It is a summary measure of average achievement in three key dimensions of human development: a long and healthy life, educational attainment and having a decent financial standard of living. Even if it has attracted (and it is still attracting) some criticism (Ravallion et al. 2012; Klugman et al. 2011; Kovacevic et al. 2010), the HDI proved to be enormously useful in shifting attention to other development outcomes beyond income and in setting up a healthy competition between countries on their HDI rank. When it is published it usually leads to national and international press coverage comparing different countries, which in turn can be used by civil society as a lever to pressure their governments (Stiglitz et al. 2018b). As remarked by Nobel Laureate A.K. Sen (initially skeptical about aggregations) this media and public attention would have not been received by a simple set of indicators and the success of HDI in fostering debate on human development would not have occurred if the exercise had stopped before a composite index was created (Sharpe 2004). The HDI has been modified and improved over the years. First, HDI releases were based on the arithmetic mean, but in

2010, to account for inequalities, the arithmetic mean was replaced with the geometric mean because the geometric mean reduces the level of substitutability between different dimensions of well-being, and at the same time it ensures that a given percentage decline in each one of the individual indicators has exactly the same impact on the HDI. “Poor performance in any dimension is now directly reflected in the HDI, and there is no longer perfect substitutability across dimensions. This method captures how well rounded a country’s performance is across the three dimensions. As a basis for comparisons of achievement, this method is also more respectful of the intrinsic differences in the dimensions than a simple average is. It recognizes that health, education and income are all important, but also that it is hard to compare these different dimensions of well-being and that we should not let changes in any of them go unnoticed” (cf. [UNDP 2010](#), 15).

As the HDI experience shows that the success of a composite index is due not only to its statistical rigor, but also to its simplicity and communicability. Thus, the Scientific Committee for the measurement of well-being suggested that Istat, given its role as a producer of official statistics, should adopt a simple, transparent, easy to interpret, but not completely compensatory, aggregation method. After some analyses and experimentations, and following the ten-steps procedure proposed by the OECD ([OECD and JRC 2008](#)), Istat has decided to adopt an aggregation method developed by Mazziotta and Pareto ([Mazziotta and Pareto 2016](#); [Istat 2015](#)) that provides the composite AMPI (*Adjusted Mazziotta–Pareto Index*). AMPI normalizes each individual indicator between 70 and 130, but to make it easier to interpret the results, this interval is shifted so that Italy is assigned 100 for the base year 2010. AMPI aggregates normalized indicators by computing their arithmetic mean and then penalizing the result with respect to the variability between them.

Istat introduced composite indices for the first time in its 2015 report ([Istat 2015](#)), one or two for each domain, and extended their use at the regional level in the 2017 report ([Istat 2017](#)).

In the following years, some issues emerged in the use of AMPI, in particular when commenting on the evolution of well-being over time ([Bacchini et al. 2020](#)). In fact, in the normalization step the search for the minimum and maximum is performed along all the time series in each domain. Then, a constraint on the base year is introduced. These two ways of considering the time dimension might conflict with each other. AMPI, by construction, defines equilibrium as the situation in Italy in 2010. Therefore, if we aggregate two Italian indicators, one already at its best in 2010, and stable over time, and another one that steadily improves from 2010 onwards, then AMPI would unduly impose more and more burdensome penalties at the composite index over burdensome penalties at the composite index over time. As a further and more specific example consider the raw indicators POL3 – Trust in judicial system and POL12 – Prison density for Italy from the domain Politics and institutions ([Istat 2017](#)). The Italian average of POL3, along the time span 2010–2016, is 4.3. The Italian average of POL12, in the same time span, is 127.2. From 2010 to 2015, POL3 falls from 4.6 to 4; in the same period POL12 – that is negatively polarized (the lower, the better) – improves from 151 to 105.2. Considering POL3 and POL12, with respect to their average values, the situation for Italy is almost exactly symmetric in these two years. Then again, in 2010 AMPI does not impose any penalization, while in 2016 AMPI imposes a penalization of 4.6 points, as if the first indicator (that fell) is much more important than the second (that rose).

Moreover, the growth rate of the composite index cannot be easily decomposed in the temporal dynamics of the individual indicators, and fails to provide a measure of social progress that can be juxtaposed with the classical GDP index (as opposed, among others, to the *Well-being Index* by the Portuguese Institute of Statistics, (INE Portugal 2017), or *The Canadian Index of Well-being* by the University of Waterloo; University of Waterloo 2016).

For example, consider the composite index *Health*, elaborated for the period 2009-2016 (Istat 2017). The index is made up of five individual indicators: *SAL1 Life expectancy at birth*, *SAL2 Healthy life expectancy at birth*, *SAL3 Physical status index*, *SAL4 Psychological status index*, *SAL9 Life expectancy without activity limitations at 65 years of age*. In Figure 5 the composite index for Italy is presented in two variants that combine different choices for normalization and aggregation: AMPI (with its own normalization) and index numbers + geometric mean. The main deviations of the two trends are actually due to the normalization process rather than the aggregation method. Between 2012 and 2013, AMPI improved by 0.4%, while index numbers + geometric mean worsened by 0.8%. In fact, the raw indicator SAL9 decreased by 4.2%, while the other raw indicators changed very little. However, even if the raw indicator SAL1 increases by just 0.4%, AMPI normalization leads to much more variability and the normalized indicator improves by 4.4%, overcompensating the change in SAL9 and dragging the composite up. In fact, index numbers completely respect percentage changes, but do not control for variability. A similar case can be made for 2014–2015.

Given the risk that AMPI composite indices could be poorly interpreted and could consequently send misleading messages, Istat currently only publishes composite indices in the regional factsheets at the end of the Bes report to help in reading about how well-being at the regional level has evolved. However because of renewed pressure for clarity and simplicity in communication (see also Section 4), we think that the time is now ripe to reconsider the aggregation methodology and the use of composite indices as an important aid for commenting on the evolution of well-being and on regional inequalities. Indeed,

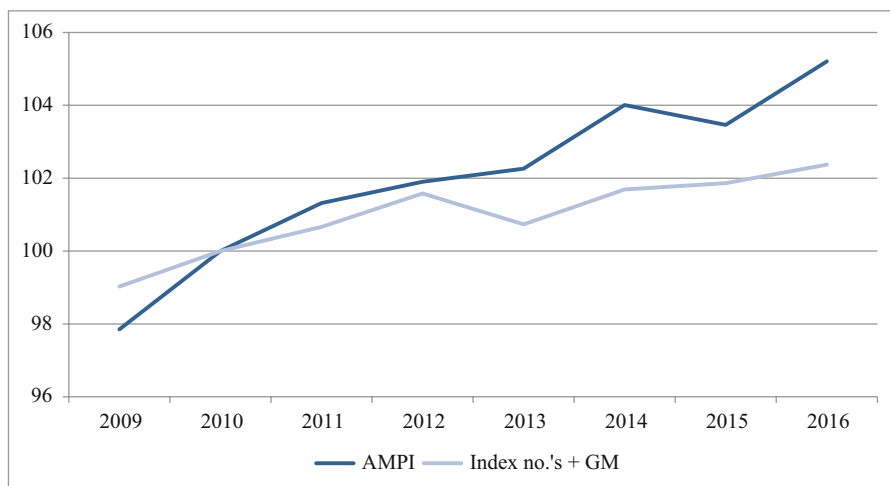


Fig. 5. Composite index health computed with two different normalization and aggregation techniques (AMPI; index numbers and geometric mean), Italy. Years 2009-2016. Data from Istat (2017).

the use of a multidimensional framework demands a comprehensive metric that can compute the progress/decline in well-being over time. However, the identification of such a metric, similar to the integrated system currently adopted to produce GDP measures, is a difficult task (Durand and Exton 2019) and requires more work and deliberation.

4. Well-Being and Policy Making

The opportunity to shift policy makers' focus from the exclusive pursuit of economic growth to a broader consideration of people's well-being and sustainability has had considerable impetus in recent years from the 2030 Agenda on Sustainable Development Goals (SDGs) and through the OECD promotional activity that, starting in the 21st century, recognized that measuring well-being was very important both for the credibility and accountability of public policies. Starting with the OECD's initiative, Stiglitz et al. (2018a, 103) pointed out that "well-being indicators could be used in the different stages of the policy cycle, from identifying priorities for action, to assessing the pros and cons of different strategies to achieve policy goals, to allocate the resources (budgetary, human, political) needed to implement the selected strategy, to monitor interventions in real time as they are implemented, and to assess the results achieved and take decisions on how to change policies in the future".

Along the same line of thought, one of the initial objectives of the Italian Bes was to provide sound quantitative support to policy makers, possibly covering all the phases of the policy cycle as suggested also by Stiglitz et al. (2018a, 103) (Figure 6). Since 2017, the Italian experience had represented an example of the implementation of this approach since well-being indicators are deemed, by law, to be a target in the budget plan. Nevertheless, some work still needs to be done to fully consider them in all phases, especially *Evaluation*.

4.1. The New Budget Law and Well-Being Indicators

In 2016, Italy adopted a law that introduced the indicators of well-being in the budget documents (L. 163/2016). This act was among the most important achievements of the

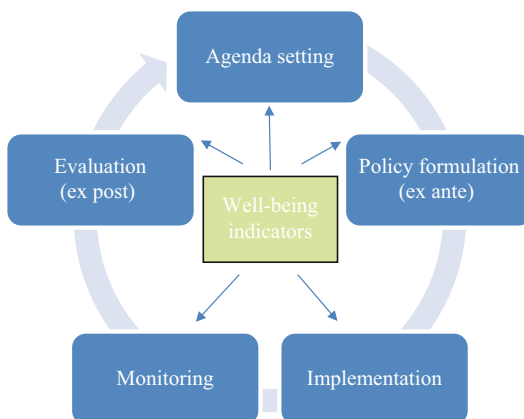


Fig. 6. The policy cycle.

Source: elaboration on Stiglitz et al. 2018a.

efforts addressing the importance of well-being for public policy. Under this act, the effects of fiscal measures must be considered against a selection of well-being indicators. This process includes two new annual reports provided by the Ministry of Treasury. The first one, published in April, is an annex to the Planning Document on Economic and Financial Policy (DEF–Documento di Economia e Finanza) in which the Government outlines the policy actions to be undertaken in the next three years. The report includes an analysis for each well-being indicator on recent developments, as well as two sets of projections of expected developments; one concerning past trends, and the other on the expected impact of measures included in the DEF (policy scenario). In February, a second report presented to Parliament includes an update of the analysis on the well-being indicators, based on the Budget law for the current year approved before the end of the year by the Parliament. This report takes into account changes in the macroeconomic scenario and specific measures set out in the current Budget law.

These two new reports add new perspectives to the policy debate and open the discussion to new directions, some of which were highlighted by Istat (Istat 2016) and the Parliamentary Budget Office (UPB 2016) during the hearings in preparation of the final draft of law 163/2016.

The timing of the two reports puts more pressure on current statistical processes since Istat's updating of the indicators is not always aligned with the timing indicated in the law, due to the organization of complex production processes. In addition, updating the selected well-being indicators means new econometric models are defined and estimated and are able to consistently address the relationship between macroeconomic variables and well-being indicators. Hence, it is important to assign adequate resources to this task, and to limit the number of indicators (recalling that the full Bes framework includes 130 indicators).

These considerations illustrate how the process set up by the new law required strong investment, a testing period and some scope for adjustment before demonstrating the full capacity of the well-being indicators in policy debates. Parts of these issues have been addressed and suitable solutions have been put in place, as described below; further remarks about usability and interpretation of results will be addressed based on the initial implementation.

4.2. *The Selection of Indicators*

How a system of indicators can be set up to inform policy making has been discussed in several publications (Martinuzzi et al. 2013; Swiss Federal Statistical Office 2012; UN 2014; Eurostat 2014a, 2014b; EU DGINS 2015), especially related to the issue of sustainability and well-being measurements.

In addition, in the Italian case selecting indicators from the whole set of the Bes framework was considered to be extremely sensitive. To address this, the new Budget law required the establishment of a high level Commission. The appointed members were: the Minister of Economy and Finance; the Istat President; the Governor of the Bank of Italy; two recognized experts Professor Enrico Giovannini and Professor Luigi Guiso. The Commission was to carry out the selection and to propose a suitable list of well-being indicators to be included in the policy process. Afterwards, the Commission proposal was to be discussed and approved by the relevant parliamentary committees.

This two-step process aimed to ensure both technical soundness and democratic legitimacy: the members of the Commission were entrusted with the scientific, methodological and operational expertise, while Parliament — as the representative body — was responsible for the final decision (Figure 7). Nevertheless, critical remarks were raised by some experts, who maintained that the lack of direct involvement of civil society representatives could weaken the credibility of the whole process (Gawronski 2017; Olini 2017).

Recognizing this issue, the Commission considered it crucial to define the criteria informing the selection process. With a starting point in the 130 Bes indicators, general considerations and specific criteria were discussed and published in a report (Comitato per gli indicatori di benessere equo e sostenibile 2017) to ensure that the whole process was transparent (see also Tinto et al. 2018).

The final list of indicators, unanimously approved by the parliamentary committees, includes 12 indicators:

1. Mean adjusted income (per capita),
2. Income inequality (quintile ratio),
3. Incidence of absolute poverty,
4. Life expectancy in good health at birth,
5. Overweight and obesity,
6. Early school leavers,
7. Non-participation in employment,
8. Employment rate of women aged 25–49 with preschool children versus women without children,
9. Victims of predatory crime,
10. Mean length of civil justice trials,
11. CO₂ and other greenhouse gas emissions (tons per capita), and
12. Illegal building,

Subjective indicators, in particular the indicator on life satisfaction, were not included in the final list, even if it has been argued that increasing the subjective well-being (as expressed by the individual perception of the level of satisfaction for his/her own life) should be considered the ultimate goal of policy (Layard 2011). However, the Commission followed a pragmatic approach: as subjective well-being cannot be easily linked to single

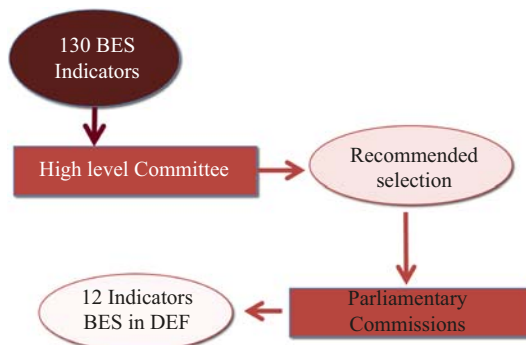


Fig. 7. The selection process.

policy measures, and it depends on a number of different factors that are out of the sphere of the Budget law, they opted for its exclusion.

4.3. Lessons from the First Round of Implementation

In 2017, the Ministry of Economy and Finance (MEF) (Ministry of Economy and Finance 2017) published the DEF including, for the first time, analysis related to a subset of well-being indicators. Only four indicators were considered: mean adjusted income; non-participation in employment; income inequality index; CO₂ and other greenhouse gas emissions. The subjective side of well-being – considered in the full Bes framework – was not included in the final selection of indicators. The debate about this choice is still open: on the one hand the need to fully into account the general level of well-being, on the other the intrinsic difficulty in linking a subjective indicator to economic measures, such as those in the Budget law, in a forecasting model. Moreover, this first selection cannot be considered to represent the different aspects of well-being and sustainability, yet it was used as a benchmark for the new procedure. The four chosen indicators were updated by Istat, whereas the MEF projected their development in the next three years both using the actual scenario (trend) and the measures included in the DEF (policy). See Figure 8.

The policy scenario appears most of the time as a *better world*: the introduction of new policies is expected to produce better effects in terms of well-being. The following February, the same procedure was repeated in the report presented to Parliament, taking into account the updated macroeconomic scenario and the recently approved Budget law. This report led to a revision of the expected target for the four well-being indicators. Following editions of the reports, up to the last one in the 2019 DEF, extended the analysis of recent trends to the 12 indicators, although projections are limited to the four indicators.

In order to examine results and possible weaknesses, one of the four indicators, *Non-participation in the labour market* is used as an example. Figure 9 shows the indicator and the four different scenarios presented by the Government in the three reports produced between April 2017 and April 2018. Some conclusions can be drawn from a comparison of

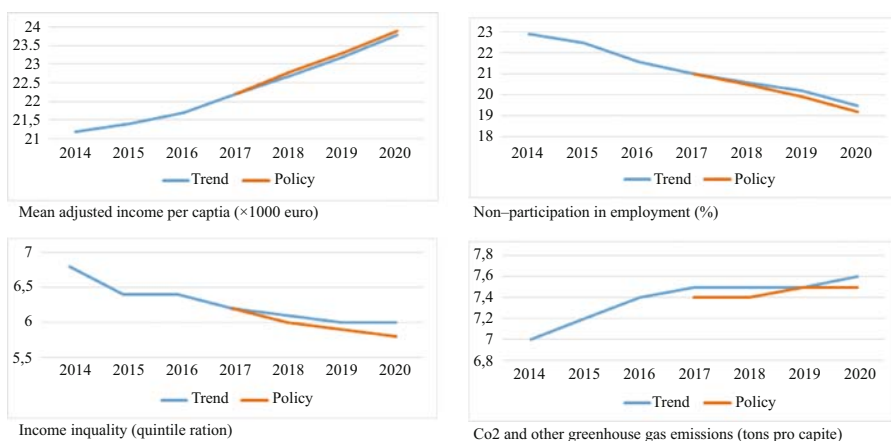


Fig. 8. MEF scenarios in the 2017 DEF. Data from Istat (observed data: 2014–2016) and MEF (forecasts: 2017–2020) from MEF (2017).

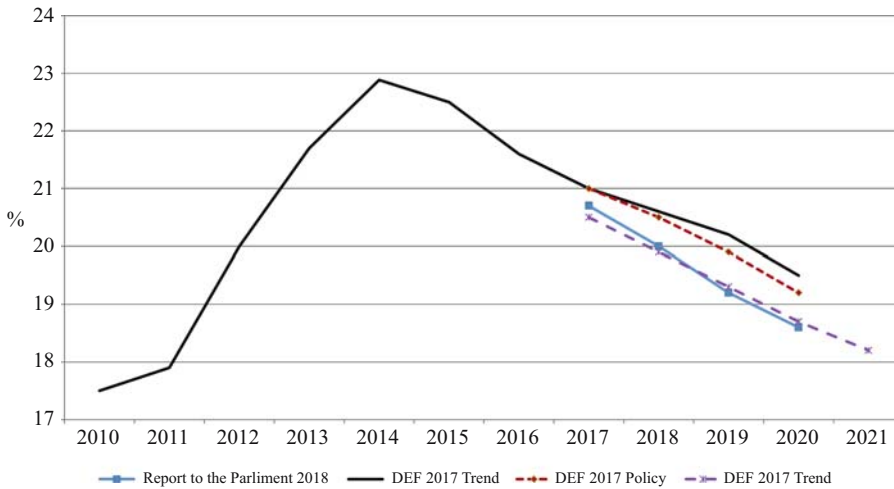


Fig. 9. Non-participation in the labour market –observed and forecasted rates. Italy. Years 2010–2021. Data from Istat (observed data: 2010–2016) and MEF (forecasts: 2017–2021).

indicator forecasts over time. The issue of the starting year for simulations is sensitive: the 2017 point forecasted in the first DEF was 0.5 percentage points higher than the observed one (included in the 2018 DEF), thereby implying an overall overestimation of the whole forecasting period. The estimate used for the Report to Parliament (RP) was not the final one but was surely more accurate, as it could already benefit from data published by Istat. This demonstrates that the preliminary estimation by the Government was more pessimistic compared to more recent estimates and that the economy performed better compared to the forecasts included in the first release.

This brings us to a key issue. Revisions of the indicators must to be correctly analyzed in order to disentangle the movements of the indicators in the different components: updating of the data, forecasting errors, different effects of the policy. Clearly, only the last one needs to be considered for the policy analysis.

With regard to the weakness of the framework highlighted in Subsection 4.1, a key point is timeliness, that is, aligning the time of data dissemination to the provision of the law. Istat, in cooperation with the National Statistical System, speeded up the dissemination process by enhancing some steps, whenever possible, and even producing preliminary pictures for some indicators.

Looking at the 12 indicators, three different approaches were adopted to fill the gap: for seven indicators only the dissemination phase was accelerated; for three indicators provisional data were produced (thanks to an improvement of the production process); two indicators were based on forecasting models (Table 1).

Table 1. Number of indicators updated with 3-months time lag, by method

No.	Method	Source
7	Currently available	Istat, Ministry of Justice, Cresme
3	Ad hoc estimates on provisional data	Istat, Ministry of Interiors
2	Models for ash estimates	Istat and Istat based on Ispra data

Once the well-being indicators were available for the last year, forecasting models developed by the MEF Department of Economic-Financial Analysis and Research were run. This activity was based only partially on past experiences, like for example for the indicator on income-inequality (Di Nicola et al. 2015) that integrates survey data, tax records and an estimate of tax evasion and erosion.

For the other indicators, a specific model has been developed that, like for the indicator for CO₂, runs on a different forecasting model, estimated to fill the data gap until the most recent year. When the complete policy cycle will be related to all well-being indicators, as suggested in Figure 6, the problem of the deviation that might be introduced by estimated data (for some indicators) will become more urgent with respect to the ex post evaluation of forecasted levels.

The full introduction of well-being indicators linked to all phases of the policy cycle is the main issue hindering process fulfillment. At present, the steps of “Policy formulation” and “Monitoring” are explicitly considered in the framework, the “Agenda setting” could be better defined, while “Implementation” and the “Evaluation” phases are not being directly addressed. All these issues need to be put to the attention of public debate to reach the goal of the well-being *revolution*: design and discuss public policies in term of their impact on well-being.

5. Conclusions

The international debate on the development and use of well-being indicators has grown in years, driven both by notable suggestions from the political and economic field (Stiglitz et al. 2009, Stiglitz et al. 2018a), and by the increasing number of measurement frameworks carried out by national and international statistical institutions. This huge effort to “go beyond GDP” has stimulated the policy discussion on enforcing the adoption of well-being indicators to assess the impact of budget plans on citizens’ well-being.

The Italian experience can be considered as a forerunner. The Italian national institute of statistics (Istat) has developed a comprehensive measurement system of 130 indicators. The development of the debate on well-being has been fostered by the annual report on Italian well-being (in 2019 Istat published the seventh edition) and by Istat’s role in national and international projects on the subject (MAKSWELL and GROWINPRO). These experiences were recognized in 2016 by the introduction of the law that made it possible to introduce 12 well-being indicators in the budget documents. In addition, the Italian Government is required to present the impact of its budget plan both on traditional macroeconomic main variables as well as their impact on the 12 well-being indicators.

This process entails new challenges and generates some drawbacks that will need to overcome in the near future: investment to improve the quality and the timeliness of the well-being indicators; identification of a metric to connect the different domains and the evaluation of composite indicators, for which the use of administrative sources such as tax registers to be associated to existing surveys (58% in the case of Bes), looks very promising; finally, the implementation of big data sources could be another important step (see, for example deliverable 2.2 and 2.3 of the project MAKSWELL (Van den Brakel 2019, 2020).

A better understanding of the determinants of well-being is another important issue. In the last two editions of the Bes report, we investigated inequalities such as those between

different generations and opposite social groups. These results are expected to be useful for the design of specific policies aiming to decrease inequalities.

Important progresses are also expected on the econometric side. The current forecasting models and the way in which they aim to address the relationship between well-being indicators and macroeconomics aggregate should be investigated in depth.

Finally, the use of well-being indicators in the budget plan requires fine-tuning of the evaluation phase, in order to identify the robustness of the relationship between policy and indicators. At the same time, this phase could be included in a specific task to extend the use of the sustainable and equitable well-being indicators to evaluate the public administration's performance (see Papi et al. 2020).

Addressing these challenges within the international paradigm of well-being appears ambitious, and reveals a set of different pathways, each of which opens up different and amplified perspectives, none of which is without a degree of coarseness. The Italian experience performed by Istat is an important reference for the international debate in promoting a well-being approach to the definition of public policy.

6. References

- Bacchini, F., B. Baldazzi, and L. Di Biagio. 2020. "The evolution of composite indices of well-being: an application to Italy." *Ecological Indicators*, 117:106603. DOI: <https://doi.org/10.1016/j.ecolind.2020.106603>.
- Becker, W., M. Saisana, P. Paruolo, and I. Vandecasteele. 2017. "Weights and importance in composite indicators: Closing the gap." *Ecological indicators*, 80: 12–22. DOI: <https://doi.org/10.1016/j.ecolind.2017.03.056>.
- Bleys, B. 2012. "Beyond GDP: Classifying alternative measures for progress." *Social Indicators Research*, 109(3): 355–376. DOI: <https://doi.org/10.1007/s11205-011-9906-6>.
- Ciommi, M., C. Gigliarano, A. Emili, S. Taralli, and F.M. Chelli. 2017. "A new class of composite indicators for measuring well-being at the local level: An application to the Equitable and Sustainable Well-being (BES) of the Italian Provinces." *Ecological indicators*, 76: 281–296. DOI: <https://doi.org/10.1016/j.ecolind.2016.12.050>.
- Comitato per gli indicatori di benessere equo e sostenibile. 2017. *Relazione finale*. Technical report. Available at: http://www.istat.it/it/files/2017/12/relazione_comitato_fin.pdf (accessed May 2021).
- Decancq, K., and M.A. Lugo. 2013. "Weights in multidimensional indices of wellbeing: An overview." *Econometric Reviews*, 32(1): 7–34. DOI: <https://doi.org/10.1080/07474938.2012.690641>.
- Di Nicola, F., G. Mongelli, and S. Pellegrino. 2015. "The static microsimulation model of the Italian department of finance: Structure and first results regarding income and housing taxation." *Economia pubblica*, 2: 125–157. DOI: [10.3280/EP2015-002004](https://doi.org/10.3280/EP2015-002004).
- Durand, M., and C. Exton. 2019. *Adopting a well-being approach in central government: Policy mechanisms and practical tools*. Global Happiness and Wellbeing. Policy Report 2019: Chap. 8. Available at: <https://www.happinesscouncil.org/report/2019/> (accessed May 2021).
- EU DGINS. 2015. *Lisbon memorandum indicators for decision making and monitoring*. Technical report. Available at: <https://ec.europa.eu/eurostat/documents/7330775/>

- 7339482/Lisbon+memorandum+28092015/18b9ea8d-bf26-47b7-81da-416dbb465467 (accessed May 2021).
- Eurostat. 2014a. *Getting messages across using indicators*. Technical report, Eurostat. Available at: <https://ec.europa.eu/eurostat/documents/3859598/5936409/KS-GQ-12-001-EN.PDF.pdf/c47039bd-c026-4d99-a819-135b5e4c1da4> (accessed May 2021).
- Eurostat. 2014b. *Towards harmonised methodology for statistical indicators*. Technical report, Eurostat. DOI: <https://doi.org/10.2785/56118>.
- Exton, C., and M. Shinwell. 2018. "Policy use of well-being metrics: Describing countries' experiences." OECD Statistics Working Papers, 2018/ 07. Available at: <http://www.oecd-ilibrary.org/content/paper/d98eb8ed-en> (accessed May 2021).
- Gawronski, P.G. 2017. "Oltre il PIL: col BES nel DEF del MEF?" *Economia e Politica*. <http://www.economiaepolitica.it/politiche-economiche/oltre-il-pil-col-bes-nel-def-del-mef/> (accessed May 2021).
- Greco, S., A. Ishizaka, M. Tasiou, and G. Torrìsi. 2019. "On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness." *Social Indicators Research*, 141(1): 61–94. DOI: <https://doi.org/10.1007/s11205-017-1832-9>.
- Harvey, L. 2020. "Social research glossary." *Quality research international*. <http://www.qualityresearchinternational.com/glossary/> (accessed May 2021).
- Il Sole 24 ore. 2020. *Qualità della vita 2020*. Available at: <https://lab24.ilssole24ore.com/qualita-della-vita/> (accessed May 2021).
- INE–Portugal. 2017. *Índice de bem estar*. INE. Available at: https://www.ine.pt/ngt_servey/attachfileu.jsp?look_parentBoui = 313012429&att_display = n&att_download = y.
- Istat. 2013. *Rapporto Bes: Il benessere equo e sostenibile in Italia*. Available at: <https://www.istat.it/it/archivio/84348>.
- Istat. 2015. *Rapporto Bes: Il benessere equo e sostenibile in Italia*. Istat. <http://www.istat.it/it/archivio/175169> (accessed May 2021).
- Istat. 2016. *Indagine conoscitiva nalizzata all'acquisizione di elementi di valutazione sul contenuto della nuova legge di bilancio*. Technical report, Istat. Available at: <http://www.istat.it/it/archivio/186476> (accessed May 2021).
- Istat. 2017. *Rapporto Bes: Il benessere equo e sostenibile in Italia*. Istat. Available at: <http://www.istat.it/it/archivio/207259> (accessed May 2021).
- Istat. 2018. *Rapporto Bes: Il benessere equo e sostenibile in Italia*. Istat. Available at: <http://www.istat.it/it/archivio/224669> (accessed May 2021).
- Istat. 2019. *Rapporto Bes: Il benessere equo e sostenibile in Italia*. Istat. Available at: <http://www.istat.it/it/archivio/236714> (accessed May 2021).
- Klugman, J., F. Rodriguez, and H.-J. Choi. 2011. "The HDI 2010: new controversies, old critiques." *The Journal of Economic Inequality*, 9(2): 249–288. DOI: <https://doi.org/10.1007/s10888-011-9178-z>.
- Kovacevic, M. 2010. "Review of HDI critiques and potential improvements." *Human development research paper*, 33: 1–44. Available at: <https://core.ac.uk/download/pdf/6224375.pdf> (accessed May 2021).
- Kuc-Czarnecka, M., S. Lo Piano, and A. Saltelli. 2020. "Quantitative Storytelling in the Making of a Composite Indicator." *Social Indicators Research*, 149(3): 775–802. DOI: <https://doi.org/10.1007/s11205-020-02276-0>.

- Layard, R. 2011. *Happiness: Lessons from a new science*. Penguin UK.
- Martinuzzi, A., L. Pintér, P. Hardi, J. Hall, and J. Bakkes. 2013. “BellagioSTAMP.” In *Encyclopedia of corporate social responsibility*,: 154–159. Springer.
- Mauro, V., M. Biggeri, and F. Maggino. 2018. “Measuring and Monitoring Poverty and Well-Being: A New Approach for the Synthesis of Multidimensionality.” *Social Indicators Research*, 135(1): 75–89. DOI: <https://doi.org/10.1007/s11205-016-1484-1>.
- Mazziotta, M., and A. Pareto. 2016. “On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phenomena.” *Social indicators research*, 127(3): 983–1003. DOI: <https://doi.org/10.1007/s11205-015-0998-2>.
- MEF (Ministry of Economy and Finance). 2017. *Documento di Economia e Finanza 2017 – Allegato: Il Benessere equo e sostenibile nel processo decisionale*. Technical report, MEF. Available at: http://www.dt.tesoro.it/modules/documenti_it/analisi_programmazione/documenti_programmatici/def_2017/Allegato_6_AL_DEF_2017.pdf.
- Muller, J.Z. 2018. *The tyranny of metrics*. Princeton University Press.
- Nussbaum, M., and A. Sen. 1993, *The quality of life. WIDER Studies in Development Economics*. Clarendon Press.
- OECD. 2007. Istanbul Declaration. Available at: www.oecd.org/newsroom/38883774.pdf (accessed June 2020).
- OECD. 2011. *How's life?: measuring well-being*. OECD Paris. DOI: <https://doi.org/10.1787/9789264121164-en>.
- OECD. 2013. *Guidelines on Measuring Subjective Well-being*. Technical report, OECD. DOI: <https://doi.org/10.1787/9789264191655-en>.
- OECD. and JRC. 2008. *Handbook on Constructing Composite Indicators: Methodology and user guide*. OECD. Available at: <http://www.oecd.org/sdd/42495745.pdf> (accessed May 2021).
- Olini, G. 2017. “Governance europea, al primo posto il benessere.” *Nuovi Lavori*. Available at: <http://www.nuovi-lavori.it/index.php/sezioni/1099-governance-europea-al-primo-posto-il-benessere> (accessed May 2021).
- Papi, L., G. Gobbo, R. Ievoli, E. Deidda-Gagliardo, and F. Bacchini. 2020. Il Valore Pubblico quale volano per finalizzare le performance di filiera dei Ministeri verso il Benessere Equo e Sostenibile. *Azienda Pubblica*, 33(4): 339–362.
- Ravallion, M. 2010. *Mashup indices of development*. The World Bank.
- Ravallion, M. 2011. *On multidimensional indices of poverty*. The World Bank.
- Ravallion, M. 2012. “Troubling tradeoffs in the Human Development Index.” *Journal of Development Economics*, 99(2): 201–209. DOI: <https://doi.org/10.1016/j.jdeveco.2012.01.003>.
- Sabbadini, L.L., and F. Maggino. 2018. “Quality of Life in Italian Official Surveys.” *Social Indicators Research*, 135(3): 1043–1055. DOI: <https://doi.org/10.1007/s11205-017-1766-2>.
- Saisana, M., B. d’Hombres, and A. Saltelli. 2011. “Rickety numbers: Volatility of university rankings and policy implications.” *Research policy*, 40(1): 165–177. DOI: <https://doi.org/10.1016/j.respol.2010.09.003>.
- Saisana, M. and S. Tarantola. 2002. *State-of-the-art report on current methodologies and practices for composite indicator development*. European Commission – JRC.

- Available at: <https://op.europa.eu/en/publication-detail/-/publication/9253d939-b47b-4428-b792-619e6b6c8645> (accessed May 2021).
- Saltelli, A. 2007. "Composite Indicators between Analysis and Advocacy." *Social indicators research*, 81(1): 65–77. DOI: <https://doi.org/10.1007/s11205-006-0024-9>.
- Sen, A. 1989. "Development as capability expansion." *Journal of Development Planning*, 19: 41–58.
- Sharpe, A. 2004. *Literature review of frameworks for macro-indicators*. Technical report, Centre for the study of Living Standards Ottawa. Available at: <https://core.ac.uk/download/pdf/7033316.pdf> (accessed May 2021).
- Stiglitz, J., J.-P. Fitoussi, and M. Durand. 2018a. *Beyond GDP: Measuring What Counts for Economic and Social Performance*. OECD Publishing. DOI: <https://doi.org/10.1787/9789264307292-en>.
- Stiglitz, J., J.-P. Fitoussi, and M. Durand. 2018b. *For Good Measure – Advancing Research on Well-being Metrics Beyond GDP*. OECD Publishing.
- Stiglitz, J., A. Sen, J.-P. Fitoussi, et al. 2009. *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Available at: <https://ec.europa.eu/eurostat/documents/8131721/8131772/Stiglitz-Sen-Fitoussi-Commission-report.pdf> (accessed May 2021).
- Swiss Federal Statistical Office. 2012. *Revision of the indicator system for the Federal Council and Parliament– Concept, methods and processes*. Technical report Swiss Federal Statistical Office. Available at: <https://www.bfs.admin.ch/bfs/en/home/statistics/catalogues-databases/publications.assetdetail.349191.html> (accessed May 2021).
- Tinto, A., F. Bacchini, B. Baldazzi, A. Ferruzza, T. Rondinella, J. A. van den Brakel, R.M.A. Willems; N. Rosenski, T. Zimmermann; Z. András, M. Farkas, Z. Fábrián; R. Münnich, F. Ertz. 2018. "International and national experiences and main insights for policy use of well-being and sustainability frameworks." Paper for the 16th Conference of the International Association of Official Statisticians (IAOS) OECD Headquarters, Paris, France, 19–21 September 2018. Available at: https://www.makswell.eu/attached_documents/news-events/iaos-oecd2018_item_1-b-4-tinto_et_al.pdf (accessed May 2021).
- UN. 2013. Conference of European Statisticians: *Recommendations on measuring sustainable development*. Technical report. Available at: <https://unece.org/statistics/publications/conference-european-statisticians-recommendations-measuring-sustainable> (accessed May 2021).
- UN–IAEG-MDG. 2013. *Lessons learned from MDG monitoring from a statistical perspective: Report of the Task Team on lessons learned from MDG monitoring of the IAEG-MDG*. Technical report, UN–IAEG-MDG. Available at: [https://unstats.un.org/unsd/broaderprogress/pdf/LessonLearnedfromMDGMonitoring_2013-03-22\(IAEG\).pdf](https://unstats.un.org/unsd/broaderprogress/pdf/LessonLearnedfromMDGMonitoring_2013-03-22(IAEG).pdf) (accessed May 2021).
- UN-DESA. 2016. *Transforming our world: The 2030 agenda for sustainable development*. Available at: <https://sdgs.un.org/2030agenda> (accessed May 2021).
- UNDP. 1990. *The Human Development Report 1990*. UNDP. Available at: http://hdr.undp.org/sites/default/files/reports/219/hdr_1990_en_complete_nostats.pdf (accessed May 2021).

- UNDP. 2010. *The Human Development Report 2010 'The Real Wealth of Nations: Pathways to Human Development'*. UNDP. Available at: http://hdr.undp.org/sites/default/files/reports/270/hdr_2010_en_complete_reprint.pdf (accessed May 2021).
- UNDP. 2016. *The Human Development Report 2016 'Human Development for Everyone'*. UNDP. Available at: <http://hdr.undp.org/en/2016-report/download> (accessed May 2021).
- University of Waterloo. 2016. *Canadian Index of Wellbeing National Report*. University of Waterloo. Available at: https://uwaterloo.ca/canadian-index-wellbeing/sites/ca.canadian-index-wellbeing/files/uploads/files/c011676-nationalreport-ciw_final-s.pdf (accessed May 2021).
- UPB. 2016. *Indagine conoscitiva in materia di contenuto della nuova legge del bilancio dello Stato*. Technical report. Available at: https://www.upbilancio.it/wp-content/uploads/2016/05/Audizione-26_52 (accessed May 2021).
- Van den Brakel, J.A., P.A. Smith, N. Tzavidis, R. Iannaccone, D. Zurlo, F. Bacchini, L. Di Consiglio, T. Tuoto, M. Pratesi, C. Giusti, S. Marchetti, S. Bastianoni, G. Betti, A. Lemmi, F.M. Pulselli, and L. Neri. 2019. *Methodological aspects of measuring SDG indicators with traditional and nontraditional data sources*. Deliverable 2.2 Methodological aspects of using Big data. Available at: https://www.makswell.eu/attached_documents/output_deliverables/deliverable_2.2.pdf (accessed May 2021).
- Van den Brakel, J.A., T. de Jong, P. Smith, N. Tzavidis, F. Bacchini, L. Di Consiglio, A. Ferruzza, A. L. Palma, G. Tagliacozzo, T. Tuoto, M. Köhlmann, N. Rosenski, C. Schartner, C. Caratiola, F. Ertz, L. Güdemann, R. Münnich, C. Giusti, and M. Pratesi. 2020. *Methodological aspects of measuring SDG indicators with traditional and nontraditional data sources*. Deliverable 2.3 Research needs in terms of statistical methodologies and new data. Available at: https://www.makswell.eu/attached_documents/output_deliverables/deliverable_2.3.pdf (accessed May 2021).
- Wilsdon, J. 2016. *The metric tide: Independent review of the role of metrics in research assessment and management*. Sage.
- Wolf M. 2019. "The case for making wellbeing the goal of public policy." *Financial Times*, 30 May 2019. Available at: <https://www.ft.com/content/d4bb3e42-823b-11e9-9935-ad75bb96c849> (accessed May 2021).
- World Bank. 1990. World development report 1990: *Poverty*. Available at: <https://openknowledge.worldbank.org/handle/10986/5973> (accessed May 2021).

Received July 2019

Revised June 2020

Accepted August 2020

Improving Time Use Measurement with Personal Big Data Collection – The Experience of the European Big Data Hackathon 2019

*Mattia Zeni*¹, *Ivano Bison*¹, *Fernando Reis*², *Britta Gauckler*², and *Fausto Giunchiglia*¹

This article assesses the experience with i-Log at the European Big Data Hackathon 2019, a satellite event of the New Techniques and Technologies for Statistics (NTTS) conference, organised by Eurostat. i-Log is a system that enables capturing personal big data from smartphones' internal sensors to be used for time use measurement. It allows the collection of heterogeneous types of data, enabling new possibilities for sociological urban field studies. Sensor data such as those related to the location or the movements of the user can be used to investigate and gain insights into the time diaries' answers and assess their overall quality. The key idea is that the users' answers are used to train machine-learning algorithms, allowing the system to learn from the user's habits and to generate new time diaries' answers. In turn, these new labels can be used to assess the quality of existing ones, or to fill the gaps when the user does not provide an answer. The aim of this paper is to introduce the pilot study, the i-Log system and the methodological evidence that emerged during the survey.

Key words: time use survey; big data; ubiquitous computing; smartphones; smart surveys.

1. Introduction

In October 2018 the official statistics offices of the European Statistical System (ESS) agreed on the “Bucharest Memorandum on Official Statistics in a datafied society (Trusted Smart Statistics)” (ESSC 2018). This memorandum was a major evolution with respect to the Scheveningen Memorandum (ESSC 2013). The ESS committed itself to a set of actions towards the implementation of changes in the way official statistics are produced, with the goal of continuing to fulfil its role in a society where data, smart technologies and artificial intelligence are a reality. As part of this effort, big data is an important potential additional source for the production of official statistics.

Trusted smart surveys make up one of the pillars of smart statistics. Smart surveys are enabled by personal devices, equipped with sensors and mobile applications that combine two data collection modes: one based on active inputs from the subjects (e.g., responses to queries, shared images), and another based on the data collected passively by the device sensors (e.g. accelerometer, GPS). Trusted smart surveys are an extension of smart surveys, administered thanks to a set of technologies which, together, increase privacy

¹ University of Trento, via Calepina, 14 – 38122 Trento, Italy. Emails: mattia.zeni.1@unitn.it, ivano.bison@unitn.it, and fausto.giunchiglia@unitn.it

² European Commission – DG EUROSTAT, 5 rue Alphonse Weicker, Luxembourg. Emails: fernando.reis@ec.europa.eu, and britta.gauckler@ec.europa.eu

Acknowledgments: This research has received funding from the European Union's Horizon 2020 FET Proactive project “WeNet – The Internet of us”, grant agreement No 823783.

preservation and data security, enhancing their degree of trustworthiness and therefore acceptance by citizens (Eurostat 2019b).

The Harmonised European Time Use Survey (HETUS) is one of the European official statistics tools that could take advantage of an implementation as a trusted smart survey. Given the novelty of the use in official statistics of data from sensors available in smartphones, it was important to kick-start with an exploration of the possibilities of use in the context of time use measurement. The tool chosen for this exploration was a hackathon, at which a significant number of competing teams attempted, in a short period of time, to find solutions to a statistical challenge.

This article reports on the experience acquired during the European Big Data Hackathon 2019, as the basis for further future development towards more evolved trusted smart surveys. The remainder of this article is organized as follows: Section 2 presents the state of the art, Section 3 presents the i-Log system (Zeni et al. 2014) that has been used to perform the pilot study. Section 4 details the data collection and preparation, Section 5 describes how i-Log pilot studies are organized and carried out. Section 6 lays out the specific use case of the European Big Data Hackathon 2019. Section 7 draws the lessons learned from the pilot study, and finally Section 8 presents the conclusions and summarises the main findings of this article.

2. Time Use Surveys During the Internet Era

The aim of Time Use Surveys (TUS) is the measurement of time use by individuals and households. In more detail, TUS measure the frequency and duration of human activities, offering a detailed view of the social behaviour of members of society. Finally, they allow us to understand how certain variables influence use of time (Dumazedier 1975) by households and household members.

TUS are more than just frequencies of individual and group time use, since they can be used as “a unique tool for exploring a wide range of policy concerns including social change; division of labour; allocation of time for household work; the estimation of the value of household production; transportation; leisure and recreation; pension plans; and health-care programmes, among others” (United Nations 2010). They can help to answer different questions of social and economic relevance, such as revealing the living conditions of a society and identifying societal changes, as well as allowing us to measure living standards within a population and between countries. Furthermore, they provide information about citizens’ demand for public and private services that are of high relevance in the decision-making process and social planning. They also enable upgrading economic accounts, improving labour force analysis, supporting the evaluation of social change, the study of gender issues, the progress on the improvement of quality of life, and a systematic analysis of leisure time (Robinson 1999).

TUS collect two types of information. The first is diachronic; that is, underlying activity sequences in time episodes (e.g. of ten minutes) over a period of one day up to a week. Such type of data is usually collected by a self-completed time-diary that allows registering, at fixed time intervals, the sequence of an individual’s activities. For each main activity in each interval, additional information is usually recorded, such as a secondary activity and information about “where” and “with whom” this activity was

done. The second type of information collected within a TUS is synchronic, which includes a paper-based or computer-assisted set of personal interviews (CAPI) about socio-economic individual and household background variables and often regarding different aspects of the household and people's wellbeing. Usually, specific information is included in the questionnaire about less frequent activities for a period longer than a day and/or item-specific questions like a seven-day work schedule. The seven-day work schedule proposed by HETUS (Eurostat 2009; Merz 2009), was removed from the guidelines of Eurostat 2019a because it is used neither by a large majority of the participating countries, nor by Eurostat. However, as stated in the guideline "The weekly schedule of working time can be re-introduced into HETUS when . . . new technical solutions will be available for the survey" (Eurostat 2019b,14).

Recently, two main aspects posed new challenges to TUS (Juster and Stafford 1991). Firstly, changes in people's living conditions and resulting use of time require adaptations to TUS. In this regard, the balance between in-home and out-of-home time has remained mainly unchanged (Gershuny and Sullivan 2019) in recent decades. However, the time people spend online has increased considerably, while the offline time spent on social activities, activities with other people, reading books and newspapers and offline hobbies has declined (Vilhelmson et al. 2018; Juster et al. 2004). Secondly, the increasing interest of the academic research community places new demands on TUS (Juster and Stafford 1991).

The three main challenges TUS face today are (1) the ability to capture the complexity of social life completely, (2) the granularity of the information, and (3) the cost to run such a study, both in terms of money and time. In the last three decades, the research community's interest in TUS has grown. There is new interest in investigating the sequence of activities and the time of the day at which they occur. Simultaneous activities can be properly investigated, and if multiple diaries are collected within a single household, researchers can use them to investigate patterns of co-presence, interdependence and cooperation (Gershuny 2015). Research results obtained through TUS pertain to three main thematic areas: (1) debates on the leisure civilisation and the end of work; (2) work and life rhythms; and (3) intra-familial synchronising of social time (Chenu and Lesnard 2006; Bison and Scalcon 2018).

At the same time, computers and modern technologies have completely changed the types of activities that should be recorded. Related to the growing availability of new technologies, the question has been raised as to whether it is more important to measure the time spent on the computer or to capture the activities performed at the computer (e-mailing, researching, reading, chatting, etc.) or on digital media (Kramarczyk 2015). For example, on average in 2018, U.S. adults spent over 11 hours a day connected to linear and digital media, performing different activities like watching, reading, listening to or simply interacting with media, according to the Q1 2018 study by market-research group Nielsen (The Nielsen Company 2018). This increase in the time spent on new technologies is not only due to the younger generations. For instance, in Q1 2018, younger adults (18–34 years old) spent less than nine hours a day, as compared to older adults (50 +) who spend over 12 hours a day with content available across platforms, with a maximum for adult 50–64 of 12.50 hours a day. On the other side, young adults 18–34 spend 57% of their time-consuming media on digital devices (App/Web on a Tablet/Smartphone, Internet on a Computer and TV-Connected Devices).

Internet is changing the individual and societal perception of time and space (Castells 2000; Kramarczyk and Osowiecka 2014). In this perspective: (a) geographic distances are losing their importance and abolishing the distinction between leisure and work, making the division between family, friends and work transparent. (b) The amount of time dedicated to each activity is reduced due to the time compression (Barney 2004); that is, the ability to perform multiple tasks at the same time. For example, while traveling by train, it is now possible to connect to the internet, make a commercial transaction, send an e-mail, eat a sandwich, watch a movie, meet on/offline friends. Paradoxically, on the one hand, new technologies lead to saving time and, at the same time, the increasing importance in our everyday life of such activities make them time-consuming (Kramarczyk and Osowiecka 2014). Nowadays, spatial mobility requires new and more in-depth information.

It is no longer suffices to capture the travel event and the reason for that. It is also crucial to understand travel behaviours. It is not enough to know only the origin and destination of the trip, but it is also important to know, for example, the route and the time taken. Moreover, in a multiplicity of tasks carried out at the same time, it is increasingly important to have more detailed information. For example, computers are a means of carrying out an activity (e.g. office work) but they can also replace an activity, while the activity that has been replaced is also essential (e.g. work during a train journey). Therefore, it is important to collect both information so that the researcher has full flexibility depending on the research question.

Finally, an important challenge concerns the frequencies of the observations and the time at which they are carried out. In fact, increasingly often, there is higher demand for faster provision of data that are of high relevance in decision-making processes and social-economic planning, or measuring well-being (United Nations 2010). However, on the other hand, due to the high cost and the complexity, especially for work required to process the collected data, e.g., the correct coding of open answers by dedicated coders (Hellgren 2014), most of TUS take time intervals of around ten years. Notice, how typical intra-personal issues such as *social well-being, work-life balance, use of information and communication technologies, mobility and travel, physical activity, social environment, geographical context, regularity and frequency of individual activities* cannot be studied if they are observed only for one day, but need to be observed for more, consecutive days, for a typical period and at household level.

In a nutshell, the future challenge is to introduce new methods and technologies to conduct a TUS that allows, for instance, *new ways of sampling time use, to record information that combines automatic and continuous data collection, with/without human intervention, that is more accurate in data collection by leveraging the new opportunities that technology offers*. Through their introduction, more context sensitive data could be collected, the burden of filling out a traditional diary could be reduced, and overall expenses lowered.

To answer most of these challenges, one opportunity comes from new human mobile technologies such as smartphones and the applications they run, or any other type of wearable device (e.g., smartwatches). The smartphone has become an integral part of the life of large parts of the population, both in economically advanced countries and in developing countries. Over time, more and more people are using smartphones all the time and they are using them, for instance, to send text messages, to be active on social media,

to check the news, to find places on a map and (even) to call other people. The Mobile Economy report Europe 2018 (GSM Association 2018) forecasts that by 2025 the penetration rate of connected devices on the European population will be 88%, with individual subscriptions (SIM cards) at 128% and that smartphone adoption, as a percentage of the total connections, will be 83%.

These technologies are a valuable alternative to traditional paper diary instruments used for surveying and they allow time use research to be carried out in a completely different way (Fernee and Sonck 2014). Smartphones not only allow respondents to report their activities at a finer grade pace per day and over multiple days, but also enable the collection of complementary information, such as the person's mood or how people feel at random moments during the day (e.g. experience sampling), what short-term activities they do throughout the day, and so on. Moreover, smartphones are a perfect tool for collecting multiple types of 'passive' data, such as geospatial or inertial sensor streams (from GPS and inertial sensors), and for collecting the interactions or communications with others (by monitoring social media apps, calling, voice, text, SMS, email, video-chat but also using Bluetooth-enabled measurements). Finally, they allow us to collect data about how people use smartphones (by use of specialized applications supporting, for example, visual data collection, audio recording, scanning, taking pictures, listening to music, visiting social network sites). Even more interesting is how these types of data can be combined with data collected with other modalities (e.g., personal and household questionnaire by Computer Assisted Web Interviews (CAWI), smartphone beeper/notifications that collect information at regular points in time including the time diary information (Robinson 2002), and continuous data from sensors). The result is a much more comprehensive overview of the respondent's time use, behaviour and well-being (Fernee et al. 2013; Fernee and Sonck 2014). In this way, smartphones are not simply a replacement for the traditional paper and pencil time use diaries, but a 'multifunctional tool' that allows us to combine the traditional methods with new data sources which would not be possible without smartphones (Link et al. 2014).

Research on surveys administered via smartphone is rather new, with only few early examples of applications that administer time use surveys (Fernee et al. 2013; Giunchiglia et al. 2017, 2018). As pointed out by Link (2018), with such new approaches, new methodological and technical issues arise. The former ranging from sample definition and penetration rate, to ethics and privacy. The latter instead comprehending usability due to poor design of the layout of the applications, connectivity related issues and battery life. Due to these considerations, the "who, what, when, where and why" varies drastically.

Conversely, only now we can start to imagine the potential and the opportunity presented by this new way of data collection for the scientific community and other stakeholders to increase knowledge about human behaviour and social rhythms. Paraphrasing and reversing the suggestion provided by Groves (2011), with a smartphone, we now collect "Organic Data" supplemented by "Designed Data": a fruitful combination of behavioural data from sensors and self-report data from human respondents. In the last two decades, the smartphone has opened the door to a new generation of measurement tools for those who study public opinion, attitudes and behaviours as well as other sociological phenomena (Link et al. 2014). They enable researchers to collect information that was previously unobservable or difficult to measure, expanding the realm of empirical

investigation (Sugie 2018). With the new functionality of a smartphone, we can capture information on people's attitudes, surroundings, interactions, and behaviours to gain a rich gratitude for the different lifestyles and personalities that characterize a particular population. In this scenario, "the use of multimode data-collection apps is not simply the next stage in evolution of CAI, but rather a species unto itself, with elements of CAI interacting with a new set of user expectations." (Link et al. 2014).

3. The i-Log System

The i-Log (Zeni et al. 2014) is a system used to carry out data collection campaigns with the ultimate goal of studying different aspects of human behaviour related to the use of time. The system consists of (1) a backend infrastructure deployed in the cloud, designed to handle a huge number of users and workloads. It is responsible for collecting, processing, storing and making the collected data available for further analysis. The second component is (2) a mobile application that runs on the users' personal mobile devices.

The mobile application has been created for Android mobile devices (an iOS version is currently under development) and enables continuous data collection about the user. In more detail, two different types of data are generated: streams of personal big data from the smartphone's internal sensors and time diaries in the form of answers to specific questions. This duality of data types makes i-Log unique with respect to other tools currently available (Runyan et al. 2013; Wang et al. 2004; Hatuka and Toch 2017) and enables addressing new challenges that concern the sociological and urban fields, in three ways. First of all, it enables investigating the real world through data recorded by phone sensors, for example, geolocation. Secondly, it enables improving existing time diaries (Sorokin and Berger 1939; Zeni 2017), especially for structured ones (Hellgren 2014). Generally, the problem of time diaries is that they are expensive and time consuming, both for the participants to fill them in and for the researcher to analyse the data. i-Log can help in this regard thanks to its ability to produce highly standardized and comparable survey results. Each answer to the survey is mapped automatically to a hierarchy of concepts collected in an ontology based on WordNet (Miller 1998). In this way, even if the survey is provided in different languages, the output is always composed by a set of standardized concepts that do not need manual processing. Finally, the third advantage enabled by i-Log is that it can help the respondent in providing the answer, that is, by reducing her cognitive load. In fact, it can compensate for gaps in the data due to the subject's attention and memory deficits that appear in traditional measurement tools. This is solved through the combination of data that i-Log collects, which essentially enable the training of machine learning models using the time-diaries answers as annotation labels. As a result, the trained models can be used to generate labels when these are not available, filling the gaps in the data.

i-Log is operational for a data collection experiment in an uncontrolled field environment, outside laboratory settings and with inexperienced users. Its main characteristics in this regard are:

3.1. Optimized Battery Usage

Today's smartphones are powerful devices with hardware characteristics comparable to high-end personal computers. Additionally, they are empowered by an operating system

that is designed precisely to run applications that allow it to be used in almost any circumstance. However, this did not come without costs; energy consumption increased significantly. The fact that in the past decade there is no major breakthrough in battery technology highlights that the battery is currently the main limitation of today's smartphones. The main solution that smartphone and application producers found is to limit the execution time of the applications on the devices as much as possible, every year with a more aggressive solution. Therefore, creating an application that runs continuously and in an efficient way is particularly challenging. i-Log runs continuously in the background to collect sensor data from the device, without creating a major impact on the battery life.

3.2. Generation of Truthful Data

In order to collect truthful data from users, we decided to install i-Log on their personal devices instead of providing dedicated ones. This choice presents different advantages, starting from cost reduction, to speeding up the set-up time of an experiment among others, but also multiple challenges. A considerable challenge was to reduce the impact of the usage of the tool on the users' smartphones as much as possible. In fact, a user who realises her/his device has slowed down, or been impacted by our application, would have quitted the experiment immediately, or have altered normal use, consequently altering the collected data. For this reason, we put a lot of effort into the simplification and improvement of i-Log performances. We completely removed the user interface typical of most applications on the market, and instead used an approach based on *notifications* (Figure 1). In fact, the user was still informed about the data collection process (Figure 1 left panel), but we decided to do so through a non-invasive notification present in the notification area of the device. The user can understand that i-Log is running and collecting data and perform some basic actions like stopping the data collection (another crucial aspect related to privacy and ethics is that the user should always be in control), or opening the settings and questions menu (Figure 1 right panel). The only situation in which the user is required to interact with a user interface element in i-Log is when time diaries are filled in. These questions are downloaded from a remote server at specific time intervals as per the experiment characteristics and have specific formats.

At the moment, i-Log enables replying to different combinations of types of questions/answers:

- Text question, multiple choice answer (Figure 2 left panel)
- Text question, single choice answer
- Text question, open text answer
- Text question, map (component)
- Map (component) question, multiple-choice answer, that is, what were you doing in the location selected on the map below? (Figure 2 right panel)
- Image question, multiple-choice answer, that is, what do you see in this picture?

3.3. Low Resource Usage

The current version of i-Log has been designed to run on Android, but an iOS version is currently in development. Android is the most adopted operating system worldwide and

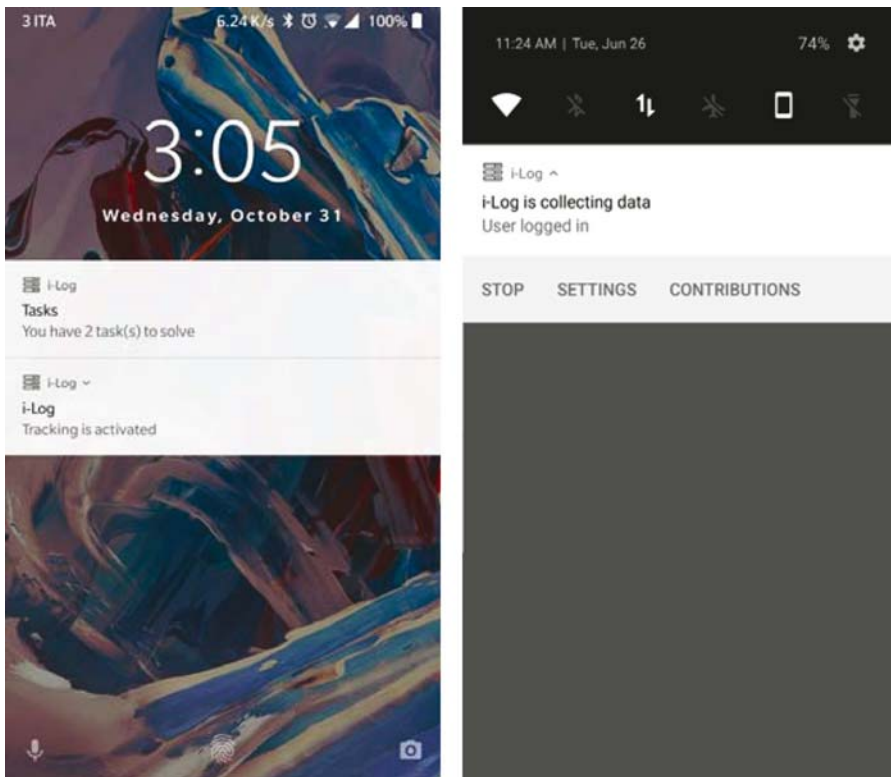


Fig. 1. The i-Log notification system. The first notification is always on and is used to inform the user about the data collection (“Tracking is activated”). The second one instead is present when time-diaries are available to be filled (“You have 2 task(s) to solve”).

runs on thousands of different devices. To accommodate the requirements of most of them we had to reduce the resources (CPU and RAM) i-Log uses as much as possible by optimizing its code and delegating part of the intensive tasks to the backend. This as an obvious advantage also in terms of energy use.

3.4. Ability to Work Offline

i-Log has been designed to work offline. It can perform most of its tasks even if the phone is temporarily not connected to the internet. This is indeed what happens in reality, a smartphone is always connected except for some specific situations, i.e., the user is in a tunnel, in the basement, in the metro. The data are collected locally and temporarily stored in a secure location in form of log compressed files. Periodically, these files are opportunistically sent over the network to the backend system that processes them.

4. Data Collection and Data Preparation

Each smartphone is different: different brand, model, hardware components and different software versions. For such reasons, available sensors differ and have different

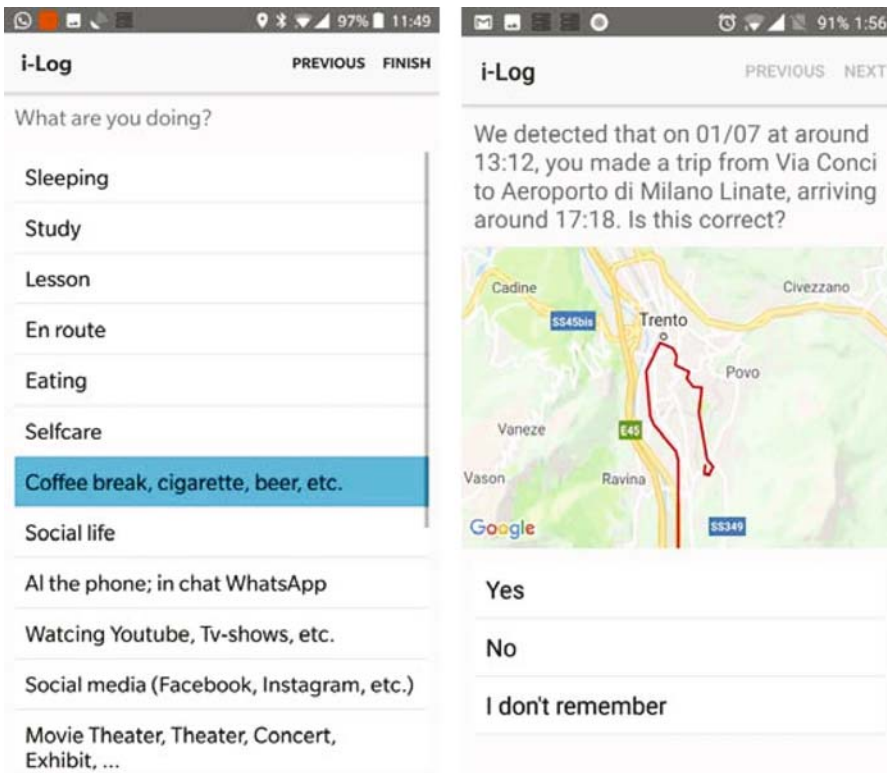


Fig. 2. The i-Log user interface about text questions (left) and map questions (right).

characteristics (collection frequency, accuracy, reliability, etc.). This makes data collection on a smartphone a challenging task.

In i-Log, we can distinguish between two types of sensors, *hardware* and *software*. The former refers to those physically embedded in the device, like the accelerometer, the gyroscope and the GPS, among others. The latter instead refers to software components that generate measurable features, such as an event when a new notification pops up, or when the device connects to a Wi-Fi network. In i-Log, the sensor data collection process can be configured based on the needs of each pilot study. Configurable aspects are the sensors from which to collect data, and the frequency at which data is collected from them. For example, it can be decided to collect data from the accelerometer at a pace of 60 values per second, while not collecting from the gyroscope at all. The following table shows the complete list of sensors available at the time of writing, together with their default collection frequency: [Table 1](#).

On a technical level, all data are generated as time-series, consisting of a tuple composed of a timestamp and one or more values. As briefly mentioned above, the smartphone generates and stores data locally before synchronizing it with the backend server for permanent storage. The device stores time-series tuples in a buffer in memory and as soon as the buffer is full, it is unloaded in a compressed and encrypted file on the device local storage, inside the application sandbox that prevents other applications from accessing them. Upon receiving the logs, the backend processes the data thanks to

Table 1. List of sensors available in the i-Log application with used collection frequency.

Sensor	Frequency	Sensor	Frequency	Sensor	Frequency
Acceleration	20HZ	Screen Status	On change	Proximity	On change
Linear Acceleration	20HZ	Flight Mode	On change	Incoming Calls	On change
Gyroscope	20HZ	Audio Mode	On change	Outgoing Calls	On change
Gravity	20HZ	Battery Charge	On change	Incoming Sms	On change
Rotation Vector	20HZ	Battery Level	On change	Outgoing Sms	On change
Magnetic Fields	20HZ	Doze	On change	Notifications	On change
Orientation	20HZ	MOdality	On change	Bluetooth Device Available	Once every minute
Temperture	20HZ	Headset plugged in	On change	Bluetooth Device Available (Low Energy)	Once every minute
Atmospheric Pressure	20HZ	Music Playback	On change	Running Application	Once every 5 seconds
Humidity	20HZ	WIFI Networks Available	Once every minute	Location	Once every minute
		WIFI Networks Connected to	On change		

pipelines that make use of Apache Spark (Zaharia et al. 2016). The processed data is then stored in a distributed Cassandra database (Hewitt 2010). On average, we expect a modern smartphone to generate 500 Mega Bytes (MB) per day of uncompressed data. A data collection with 500 participants would generate around 7.5 Tera Bytes (TB), without redundancies and backups, in one month.

Once the data is stored, it is immediately available for analysis. The main way to access these data is to read them directly from the Cassandra database which, due to its distribution and scalability, enables responding to queries in a linear amount of time even with huge amounts of data (in the order of TB). To make the data available to a broader audience, not only limited to computer scientists, pipelines were created to export the data, making use of Apache Spark (Zaharia et al. 2016), a distributed computation tool that reads the data directly from Apache Cassandra and writes them in files on a file system according to the Apache Parquet (Vohra 2016) format.

For the European Big Data Hackathon 2019, the participants were provided with an environment with a big data cluster, where they could use a distributed computing infrastructure powered by Apache Spark that naturally integrates with the Apache Parquet file format.

5. Pilot Studies

A pilot study is composed of six steps that the participants are asked to perform. These steps are:

1. Once the subject decides to participate, he is invited to fill out a personal questionnaire in which the sociodemographic characteristics, the psycho-social information, together with their personal data (phone number, address, smartphone characteristics, etc.) are collected. While initially this was run separately, in the latest iterations of the studies, this process was included directly in i-Log, without the need to use an external solution,
2. The subject is provided with a code (the same for every participant) to enable initializing the i-Log application,
3. The subject is allowed to download the i-Log application from the Google Play Store (i-Log 2019) and install it on his personal Android smartphone (at the time of writing, the iOS version was not yet available and is currently under development),
4. The subject is required to insert the code that identifies the study into i-Log to start using it. Without this code, i-Log does not perform any operation. We decided to add this additional security layer to be sure about the participants of each study, forbidding external people to participate,
5. At the first execution, i-Log presents the user with an installation procedure (Figure 3). The objective of this procedure is to explain the purpose of the study, to formally ask the subject to read and to give consent to the privacy statement and to

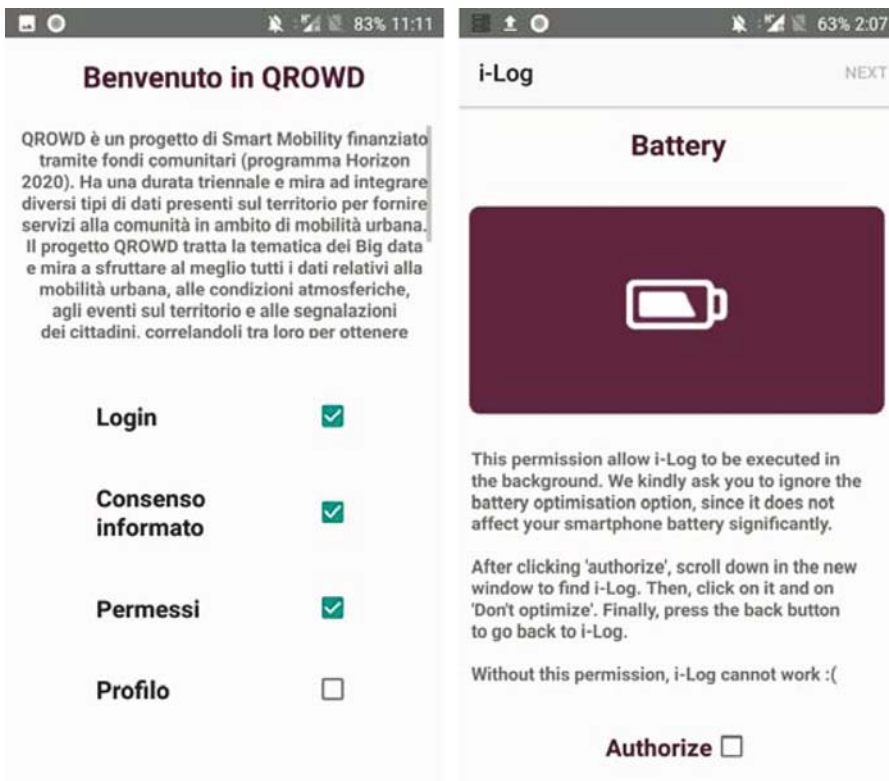


Fig. 3. The i-Log initialization procedure with informed consent, login, permissions and profile (left) and example of permission screen prompted to the user (right).

grant permissions to collect data from the personal device (both from a technical and a legal point of view), and

6. Once the installation of i-Log is complete, the pilot study enters its active stage. i-Log will now collect data from the smartphone's internal sensors and administer the Time Use Surveys (TUS). During this stage, the participants are asked to use the mobile application for a specified period of time (from days to months). During this period, a helpdesk is available for (technical) issues via email or phone, in case the participants encounter any issues with the application that cannot be solved by reading the provided written manuals.

HETUS uses a time resolution of ten minutes for the recording of the time of the activities. However, in i-Log this value can be changed for every data collection campaign. Each set of questions is pushed to the application from a cloud server, to guarantee synchronization among the participants, and once received by i-Log it is shown as a notification. Even if the respondent is instructed to reply as soon as the notification is received, in a real-life study, this is not always possible. For this reason, the respondent can be given a limited or unlimited amount of time to reply to a time episode once the notification is received. An important feature of the application is the possibility to monitor different aspects of the behaviour of the respondent in terms of answering behaviour to each question. In fact, it records the time elapsed between the time of the notification and when the subject begins to fill in the diary, and the time taken to complete the time-episode diary. This information is useful for testing the reliability of the respondent answers (Bison et al. 2018) and is an innovative aspect introduced by i-Log and not present in paper-based TUS.

During a study, the smartphone of the user is required to be online at times to receive the questionnaire initially or to synchronize the collected data with the cloud server. This is not a problem for modern smartphones since they are connected for most of the time to a network, either Wi-Fi or 3/4G. However, the application can work even if the connection is not available for long periods of time. If for example, the user is in a building without network coverage, the individual questions are not received when supposed to, but instead they are all delivered only once the device is back online. Each time episode is composed of the same questions as the HETUS study (what are you doing? where are you? who are you with?) plus a fourth question that is a seven-point scale about the person's mood. For each question, the user is presented with a list of possible pre-coded activities, places and peoples. Additionally, for the activities, it does not collect secondary activities. These choices have been made to reduce the respondents' cognitive load and time necessary to reply. If people doubted whether their activity matched one of the predetermined categories, they could find additional explanations and examples in the user manual. In addition to the time use diary, the respondent was required to reply to two additional experience questions per day, one in the morning and one in the evening about their mood, with a seven-point scale.

6. The European Big Data Hackathon 2019

Between the 8th and 12th of March 2019, the second European Union (EU) Big Data Hackathon (Eurostat 2019b) took place alongside the 10th New Techniques and Technologies for Statistics (NTTS) conference in Brussels, Belgium. Seventeen teams nominated by European National Statistical Institutes competed to develop a data

analytics tool to address the annual challenge: “How can innovative solutions for data collection reduce response burden and enrich or replace the statistical information/data provided by the time use survey?”

The European Big Data Hackathon had three main objectives:

1. to solve statistical problems by leveraging algorithms and available data, by engaging with developers and data scientists across Europe, giving them the opportunity to work with relevant data sets in order to generate new ideas and potentially contrive novel algorithms,
2. to produce innovative products, including visualisation tools, developing prototypes that official statistics will be able to integrate at European and national level, and
3. to promote partnerships with the research community and the private sector, by raising awareness about big data initiatives in Official Statistics in Europe.

6.1. Data Sources

Given the focus of the Hackathon, the teams were suggested to use big data, either the data sets provided for the event and/or some acquired by themselves. The organizers decided to provide personal big data about individuals, collected from their smartphones. The members of the teams and additional volunteers collected data before the hackathon using two frameworks, i-Log and myBigO. Additionally, the teams had at their disposal traditional time diary data. Part of the diary data was collected on purpose for the Hackathon, via i-Log, and part was collected previously for other purposes, from the Modular Online Time Use Survey (MOTUS) and the HETUS.

i-Log was used to provide the main data set for the Hackathon. The time diary user input and the sensors data specifications were adapted for the Hackathon (see Subsection 6.2). i-Log provided the only data set where sensors big data and time diary data referred to the same sample.

The second source of big data was the one provided by myBigO, a framework developed in the context of BigO, an International European research project to fight against obesity (Diou et al. 2018). Through the myBigO mobile application, activity data together with information about mood, and pictures of meal and food advertisement was collected from volunteers. It contained raw data from sensors, pre-processed data and self-reported data. Sensor data included geolocation data and recorded signals from accelerometer, barometer, light, proximity sensor, relative humidity sensor and thermometer. The processed data contained the recognised (i.e. predicted) physical activity (steps, walking, jogging, biking, . . .), recognized visited points-of-interest (POI) and recognized transportation mode for trips between detected POIs (foot, bike, car, bus, train). The prediction models used were the ones trained in the BigO project. The self-reported data contained the pictures of meals and the mood. myBigO did not collect time diary data.

A first source of time diary data was MOTUS, which is an online time use survey administered via a website and a mobile application. The data set from MOTUS and used in the Hackathon was collected from a sample of teachers in primary and secondary schools in Flanders during one week in 2018 (Minnen et al. 2018). Participants encoded their activities with reference to a pre-specified classification of 81 work-related activities. 21 activities related to personal and free time, together with the exact start and finish time.

For each activity, participants registered where they were and with whom. In the case of a travelling activity, they registered the mode of transportation used. In the case of a work-related activity, they registered which technical tools they were using, type of teaching platform and if they were satisfied with the working activity itself (scale one to seven). The data set included individual validated data for 8,571 teachers.

A second source of time diary data was HETUS, which is a traditional paper-based time use survey (Eurostat 2019a). This data set contained anonymised micro-data from HETUS wave 2010. HETUS wave 2010 consisted of 18 countries that had collected TUS data between 2008 and 2015 based on harmonised guidelines. From the 18 participating countries, five of them were included in the Hackathon data sets: Austria (AT, 8,234 observations), Belgium (BE, 11,118 observations); France (FR, 27,903 observations); Hungary (HU, 8,391 observations); and Norway (NO, 7,882 observations). The data contained the background information of the individuals and their households, and a diary where every ten minutes of the day the following information is recorded: main and secondary activity, where the activity took place, if the individual was alone or with someone and if ICT was used. Each data record (per diary day) contained a total of 1,656 variables.

6.2. *i-Log Data Collection*

To optimise *i-Log* for the Hackathon, both the time diary input questions and the sensor data collection were adapted. The purpose of the adaptation was to provide the teams with a wide range of possibilities for their analytic choices. One option was to allow the comparison of data collected by *i-Log* with data collected by the HETUS survey to some degree. The two surveys are very different, and a direct comparison is not possible; however, it offers insights into what is feasible with innovative data collection such as smart surveys, versus traditional data collection like the HETUS.

The app collected data through three modes: a one-time user input of personal background characteristics at the start of the collection phase, a regular user input and an automatic collection of sensor data throughout the whole data collection period (for privacy preservation and data protection see Subsection 6.3).

Once, at registration time, the participants filled out personal background characteristics, namely gender, occupation, their main activity status, the employer, and the place of employment.

The regular user input was triggered once per hour. The choice of the frequency was driven by personal experience during an initial pilot phase, where a frequency of twice per hour was tested and considered too burdensome. Each hour the participants received a notification on their smartphone with four questions, and were prompted to fill out information about:

- their activity “What are you doing?” with 19 answer categories such as sleeping, eating, working, and so on,
- the current location “Where are you?” with 13 categories such as home, workplace, restaurant, and so on,
- the mode of transport (if travelling with a selection of 8 categories such as car, bus, and so on.),

- the persons with the participants at the time of the question “Who is with you?” with seven categories such as nobody, partner, friends, and so on, and
- their mood “What is your mood?”.

Each question included one open-ended category. If the participants did not fill out each user input, this created a backlog of questions that could be answered at a later stage.

The time diary input questions and its categories were adjusted as closely as possible to the HETUS survey questions, to allow maximum analytical possibilities for the Hackathon. It is evident that a different mode requires a revised design. When rewriting the questions to fit the screen of the smartphone, we shortened the questions and categories. Considering that some studies suggest that the quality of responses increase with the switch to app and online modes with respect to paper-based surveys (Stella et al. 2018), it can be assumed that this redesign reduces the response burden and improved the quality of the answers. The exact wording considered possible response burden and survey mode effects. Due to time restrictions, the user input questions could not be pretested as extensively as they could have been.

The possibility of the collection of the sensor data in i-Log is manifold. For the Hackathon, the decision on which sensor data to collect took account of privacy preservation rationale and collection needs. Automatically, the app collected the following sensor data: acceleration/ gyroscope/ gravity/ rotation vector/ magnetic field/ orientation/ temperature/ atmospheric pressure/ humidity/ proximity/ position/ Wi-Fi network connections/ running applications/ screen status, flight mode, battery status, doze modality/ headset, audio mode, music playback (no track info)/ notifications received, touch event/ cellular network info.

Before the actual data collection, the developed i-Log app for the Hackathon underwent a small and brief experimental pre-test. This helped to improve the actual collection phase. Small initial communication problems that the volunteer participants in the experimental test encountered, like how to switch on permissions to receive the input questions, were solved instantly without any disturbance for the data collection – thanks to the instant feedback from the backend system.

An important result of this small test showed that the user was prompted too frequently, to a point where the risk of dropouts was too high. Therefore, the frequency of user input was reduced to an hourly intervention. The backlog of questions created when the participant did not fill out the user input or was offline, created some irritation and was reduced to eight times. This is justifiable as in most cases it can be assumed that the participant has long stretches of the same activity, such as work or sleep, where he or she should not be asked to fill out the same activity too often. Those decisions maintained the balance between the data collection needs on the one hand, and the volunteer data collection of the target volunteers on the other.

The target-volunteering participants for the data collection were the hackathon participants themselves, as well as other volunteers recruited by Eurostat and the participants. The target group was only persons using Android phones, as i-Log was only available for this operating system at the time of the hackathon. Eurostat colleagues received an article via the intranet describing the project in a convincing way and referring to the privacy statement. The registration for i-Log consisted of downloading the app from the Google play store, installing it and entering a four-digit access code. The data collection period was from 28 January until 10 February 2019. In total, 95 persons registered for the participation in the i-Log data collection experiment for the Hackathon.

At the end of the data collection period, 66 participants registered around 190,000 hours of sensor data (between all the sensors and all volunteers), as reported in Figure 4. Besides the 29 volunteers who did not register any data, some of them did not report data every day. The number of volunteers reporting data throughout the 14 days of data collection varied between 39 and 52, with a clear declining trend in time (Figure 5).

Besides the data collected automatically by the application, each volunteer has registered on average around 15 diary hourly entries per day (Figure 6). In total, between all the volunteers 8,548 entries have been registered.

6.3. Privacy Preservation and Data Protection

Personal big data poses particular challenges for the preservation of participants’ privacy. In particular, geolocation data makes the re-identification of individuals in anonymised data set relatively easy (De Montjoye et al. 2013). For this reason, privacy preservation and data protection placed particular importance in this experiment.

Privacy preservation is about the non-invasion of the private sphere of the data subject, that is, the volunteer, meaning the non-disclosure of information he/she is not willing to share about him/herself. This was accomplished at three levels.

At the first level, this was done by allowing the volunteer to not share the information. In the case of the data actively inputted (i.e. the activities), the volunteer had the possibility of not answering. In the case of the data collected automatically (i.e. sensors) there were two

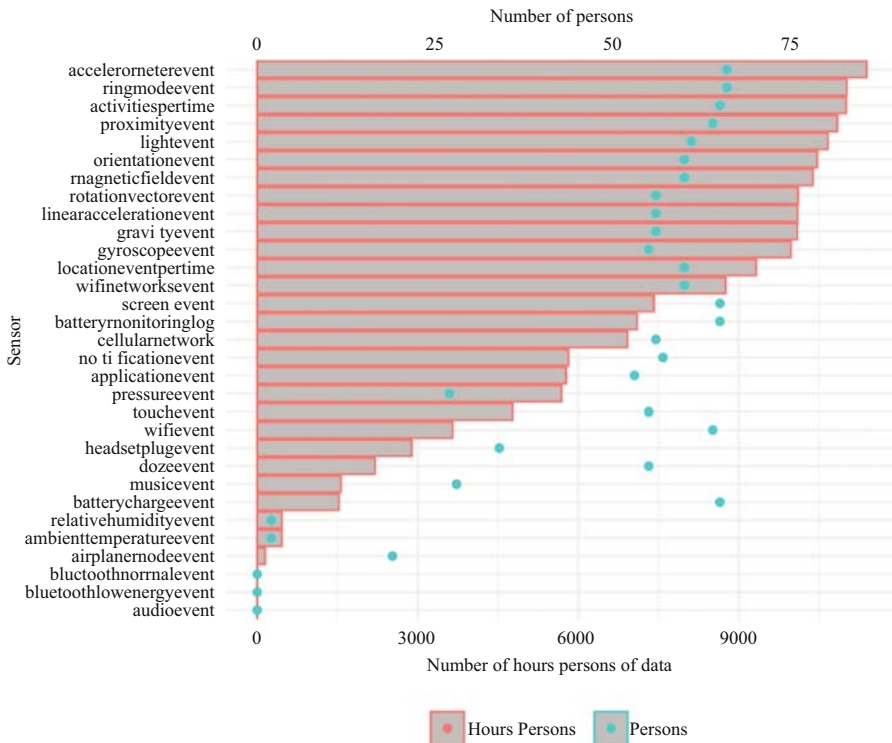


Fig. 4. Amount of data collected (in hours and number of persons) from each sensor.



Fig. 5. Amount of data collected (in hours and number of persons) per day.

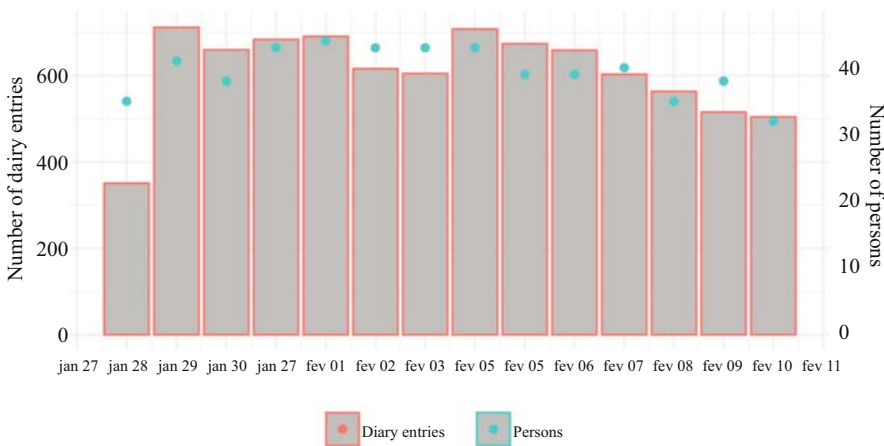


Fig. 6. Amount of time diary entries collected per day.

mechanisms. The first one was by not giving permissions to the mobile application to access particular types of data (i.e. location data). This could be done in the Android settings and via the i-Log mobile application. The second mechanism was by giving the volunteer the possibility at any time to stop the collection of data (done via the mobile application).

At the second level, given that the volunteer chose to share his or her data, his or her privacy is preserved by minimising the risk that his or her identity is associated with the data, which was collected about him or her. His or her identity may be revealed by pieces of information which are public (or easy to obtain) and unique to her. In our case, this was mainly the email address obtained when the volunteer registered for the data collection. We minimised the risk by separating this identifying information from the data collected by the mobile application and by minimising the number of people and the cases where access was given to both types of data. The linking of both types of data needs to be possible, to comply with the legal obligation of giving the data subject the possibility to review, change and delete her data.

At the third level, given that even with only the data collected by the mobile application, it may be possible to indirectly identify the volunteer if additional information about her is obtained via other means, her privacy is preserved by minimizing the number of people who have access to the data and by restricting the time during which that access is granted. The people who had access to the data were the members of the 17 teams participating in the Hackathon, two persons in Eurostat (the data controller) and two persons responsible for the system administration of the backend. The access was granted until one month after the end of the Hackathon. On the 12 April 2019, all the data held by Eurostat was deleted and all individuals who had access to the data were legally obliged to delete any data still held (by the terms of use agreed by the participants in the Hackathon).

Data protection is about assuring that only rightful people can be granted access to the data, and not anyone else. This was done technically via the use of encryption during any data transmission and access control to any stored data. Besides those technical measures, this was done by imposing legal obligations to all individuals who had access to the data. The members of the teams participating in the Hackathon had to agree to the terms of use of the data during registration in the Hackathon, and these required them to “preserve the confidentiality of information”. The staff of the App providers are bound legally by the contractual relationship between their organisations and the Commission and signed non-disclosure agreements. The staff of Eurostat are bound by the staff regulation of the European institutions.

6.4. Results

After the announcement of the challenge in the evening of the 8th of March 2019, the teams had two days to work on the development of a data product addressing the challenge. On the 11th of March 2019 in the morning, each team had ten minutes to present their data product prototype.

The advantage of a hackathon is that it allows the generation of a relatively large number of ideas in a short time period. In total there were 17 teams participating in the Hackathon. Out of those, six teams used, or partly used, data collected by i-Log.

The Swedish team identified places of interest (“zones”) and visualised trajectories in 3D, with the objective to improve the response burden for time use surveys by predicting the type of location (e.g. work, home). Based on i-Log data they created zones of mobility patterns over the day by using geo-positioning data. Every time of movement out of the defined zone triggers a question to the respondent to name their location. Over time this creates a multi-selection of activities and the system is able to make suggestions about which type of activities are being performed by the respondent based on sensor information.

The presented visualisation of trajectories can be used to reward and motivate participants. The tool can be used for machine learning to train the model before the user has to start filling out the questions in order to reduce the response burden and add to the transparency of the data collection process.

The Romanian team had the objective of improving the quality of the time use survey by using smart survey data and analysing work-life balance variables based on TUS data. The team explored all data available to assess if there are variables that could be removed to reduce the response burden for respondents. An imputation method for the transportation

mode (on foot, public transport and by bicycle) was developed by using an R imputation package to visualise the variables and results in real time. Furthermore, an analysis of work-life balance variables was performed using HETUS and MOTUS data.

The Greek team explored how smartphone data can reduce the response burden of time use surveys by predicting the ‘sleep activity’ events occurrence by using sensor data concerning the ‘doze’ mode of the device. By inferring the sleep activity, the corresponding time diaries annotations could be automatically filled by the system with the correct label, rather than requiring the user to manually annotate them. Doze mode is the sensor that allows to detect when the phone is in sleep mode, not used for a prolonged period of time. The team was able to correlate such modality with the user’s sleep patterns. A model was developed and evaluated with the respondents’ answers to check if it was indeed sleep. As future work, other sensors could be integrated to refine precision and recall and even to extend the model to recognize other activities.

The Dutch team chose to focus on physical activities for time use surveys. Like many other teams, they have identified the time diary as the main source of response burden and found that respondents tend to either reduce the amount of questions, or the time interval between the answers gets larger over time. To tackle this problem, the team proposed to label the activities automatically. After some exploration, the team found the i-Log data as the most promising to use. After some data cleaning, the team built a model for predicting activities from sensor data. Furthermore, the team started to train the model and developed a convolutional neural network. However, some issues prevented further training. The labelling should have been more frequent, or closer to the activity, to allow an efficient search for the matching activity. Missing labels (nonresponse) in the accelerometer and the small sample size made it difficult to find the matching activity.

The Croatian team developed an app to visualise the relationship between HETUS data with i-Log data for the variables on activity and location. The objective was to compare sensor-based data information with traditional survey questionnaire information of the respondents. Two modules are possible to visualise in the app: activities and places of location. The team used open source software for their development, and the major difficulty the team faced was to standardise both sources in a way that made it usable for the app.

The Latvian team developed an app to compare data from i-Log sensor- and user input data on questions of activity and location. The idea was to find out if no significant movement of location might correspond to an inactive lifestyle, and if significant movement might indicate the mode of transportation. In sum, to predict activities from training using user input information. The team reported problems setting up the infrastructure, which led to insufficient time for the development in the Hackathon.

After the jury evaluated all proposals, they announced the six winners of the Hackathon. The six winners were: 1st Statistics Poland, 2nd Istat (Italy), 3rd ONS (UK), 4th Statistics Estonia, 5th Destatis (Germany), and 6th Statistics Netherlands. At the award ceremony as part of the conference of New Techniques and Technologies in Statistics (NTTS 2019), the first three winners received their prize and gave the large audience a laureate lecture of their work;

- Statistics Poland received the 1st prize for the creation of an open source prototype, delivering a dashboard for the data analysis of the population time use,

- Istat from Italy received the 2nd prize for the creation of “SMUTIS”, an integrated open source environment for data analytics, visualisation and food classification.
- The 3rd prize was given to the ONS, UK, for their development of a system to enrich the data collected via traditional questionnaire-based surveys with an automatic processing of photos of meals taken by respondents.

The outcomes of the Hackathon and the event itself were a tremendous success. Some teams are now in contact with HETUS and HBS production domains for further development and/or integration of their prototypes.

To conclude, the use of sensor data to predict the location and to pre-fill the questionnaire would reduce the response burden immensely. Furthermore, the time use survey can be enriched by visualisations, not only to increase the motivation for respondents to fill out the sometimes lengthy questionnaires (time diaries), but also to make the collected data more accessible to a wider audience, and promote the richness of information collected by the surveys.

7. Lesson Learned

The pilot study presented in this article for the European Big Data Hackathon 2019 is only one example of the kind of study that can be performed with the i-Log system. Other previous studies ([Zeni et al. 2019](#)), also as part of European H2020 projects ([Maddalena et al. 2019](#)), as well as planned studies in different countries around the world for 2019 and 2020 prove the feasibility of using i-Log in the field. Each study allows i-Log to improve the methodology and the system, as well as to introduce new functionalities.

Smartphones and related technologies are creating new opportunities and at the same time presenting new challenges for TUS. They create new ways of sampling and recording information, which combine automatic and continuous data collection with limited or no human intervention. This is more accurate and reduces the burden of manually filling in a traditional diary. It reduces the cost of performing a study and potentially increases the number of participants. At the same time, such hybrid solutions present many new methodological and technical problems. Mainly, challenges are the selection of a sampling mode and the penetration rate, but it also raises issues on the ethical and privacy side, as well as technical challenges such as usability, connectivity, design, app layout and battery life of devices.

From a technical point of view, each study results in a vast amount of data together with feedback collected from the participants that can be used to improve the i-Log application and its usability. In general, we can distinguish between two main categories of elements that can be improved: the time use survey part and the big data collection part. Regarding the former, the main elements that the users reported as possible points to be improved concern extending the current functionalities of the application while replying. Examples are the possibility of replying “Same as the previous one”, having a mechanism to automatically reply when is not possible to do so (for example, while sleeping, at the cinema, and so on) and defining standard routines to avoid replying to all the three sub-questions. With respect to answering behaviour, some users highlighted that they needed a larger testing phase to understand the question wording, sequence, etc., as well as the categories, and for the technical handling/user interface of the app. In terms of the big data

collection part, we learned that an interesting feature to be introduced concerns collecting data also from wearable devices in addition to smartphones. In fact, many users now have a smartwatch or a smartband connected via bluetooth, and the data generated from them can provide additional insights into their activities and could also help to fill in the gaps that occur when users are not using their smartphones. Bluetooth could also be used to detect nearby devices and physical networks of people in the real world. This functionality was originally present in *i-Log* but temporarily removed due to high battery usage, but insights show that it should be restored because it enables collecting richer data, which is important for time use analysis.

Concerning the backend part of the system, we learned some valuable lessons from the feedback received and from our observations. First, by moving to the whole system to the cloud, we could reduce and optimize the resources needed to run a pilot study (and consequently the cost). In fact, in the cloud, everything is on demand and the system can scale up linearly depending on the load, while with standard servers they have to be bought in advance since the purchasing process is long and complicated from a bureaucratic angle. Also, when the pilot study is finished, the resources can be released, and the costs are reduced to zero. The migration to a cloud infrastructure also helped in improving the deployment phase of the whole architecture. With the Big Data Hackathon use case, we were able to move to a one-click-deployment pattern, whereby all the components of the backend were deployed instantly with a single user operation. This increases the reusability of our approach and reduces the time needed to run a new pilot study in a different site.

An additional element that improved data collection was the helpdesk towards the final users, granting different levels of assistance. Level 1 consisted of an exhaustive FAQ guide available online, level 2 an email address where a dedicated person could answer, and finally, level 3 email support where requests not satisfied by level 1 or level 2 could be answered directly by the engineers who built the system. Additionally, the helpdesk and the different levels showed the need for a dedicated role in a pilot study called Field Supervisor. The responsibilities of this person are to monitor the pilot study through a dedicated backend interface that leverages on insights generated by the collected data. This data-driven approach to support the field study helps to prevent possible unwanted situations that prompt a user to request assistance from the help desk. One example is a specific user who does not send data to the backend server. In such a case, the field supervisor can be notified by the system of such behaviour, be proactive, contact the user and ask focused questions to better understand the problem, or trigger a specific functionality. For example, enabling the synchronization of the files over the Wi-Fi network.

8. Conclusions

In this article, we have described the experience with a pilot study of a smart survey in the context of the European Big Data Hackathon 2019, a satellite event of the NTTS conference, organized by Eurostat. The main tool used for this pilot was *i-Log*, which uses the smartphones of a pre-selected sample of respondents and combines two data collection modes based on active input from the subjects together with data collected passively from sensors inside the smartphone.

The results of this study look promising. i-Log proved capable of carrying out a real smart survey that combines multiple data sources, which is not simply an extension of computer-assisted interviewing (CAI). It is a very new type of data collection, with elements of CAI interacting with a new set of user expectations. The challenges of these novel tools are still new and partially unknown. Among them, we can mention the validation and completeness of the data due to malfunctions in the automatic systems.

We are at the beginning of a long and challenging journey. There are many issues to be addressed, from both a technical and a methodological point of view, such as the exploitation of the data, and above all the protection of the respondents' privacy. Nevertheless, this new data collection tools offers tremendous and unimaginable new opportunities. The important aspect is to get started.

9. References

- Barney, D. 2004. *The Network Society*. Cambridge: Polity Press.
- Bison, I., and A. Scalcon. 2018. *From 07.00 to 22.00: A Dual-Earner Couple's Typical Day in Italy: Old Questions and New Evidence from Social Sequence Analysis*. Edited by G. Ritschard, and M. Studer. *Sequence Analysis and Related Approaches*.
- Bison, I., M. Zeni, M. Busso, E. Bignotti, F. Giunchiglia, and G. Veltri. 2018. "More Than Meets the Eyes: Complementing Surveys with Mobile Phone Digital Data Trail." ESRA BigSurv18 Conference. Available at: <https://www.bigsurv18.org/program2018?sess=52#222> (accessed May 2021).
- Castells, M. 2000. *The Rise of the Network Society. The Information Age: Economy, Society and Culture*. Volume I (2nd ed.). Oxford: Wiley-Blackwell.
- Chenu, A., and L. Lesnard. 2006. "Time Use Surveys: a Review of their Aims, Methods, and Results." *Archives Européennes de Sociologie / European Journal of Sociology*, Cambridge University Press (CUP), 47(3): 335–359. Available at: <http://www.jstor.org/stable/23998949> (accessed May 2021).
- De Montjoye, Y.A., C.A. Hidalgo, M. Verleysen, and V.D. Blondel. 2013. "Unique in the crowd: The privacy bounds of human mobility." *Scientific Reports* 3: 1376. DOI: <https://doi.org/10.1038/srep01376>.
- Diou, C., I. Ioakeimidis, E. Charmandari, P. Kassari, I. Lekka, M. Mars, C. Bergh, T. Kechadi, G. Doyle, G.K. O'Malley G, Heimeier R, Lindroos AK, S. Sotiriou, E. Koukoula, S. Guillén, G. Lymperopoulos, N. Maglaveras, and A. Delopoulos. 2018. "BigO: Big Data Against Childhood Obesity." *European Society for Paediatric Endocrinology* Vol. 89. Presented at: 57th Annual European Society of Pediatric Endocrinology; 27–29 September 2018; Athens, Greece. Available at: <https://abstracts.eurospe.org/hrp/0089/hrp0089p3-p127>.
- Dumazedier, J. 1975. "The Use of Time. Daily activities of urban and suburban population in twelve countries." Edited by A. Szalai. *Revue française de sociologie* 16 no. 1: 125–129. DOI: <https://doi.org/10.1080/00222216.1974.11970168>.
- ESSC (European Statistical System Committee). 2013. "Scheveningen Memorandum-Big Data and Official Statistics." Adopted 27 September 2013. Luxembourg. Available at: <http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningenmemorandum-27-09-13> (accessed March 2016).

- ESSC (European Statistical System Committee). 2018. “Bucharest memorandum on Official Statistics in a datafied society (Trusted Smart Statistics).” DGINS Conference. Available at: <https://ec.europa.eu/eurostat/web/ess/-/dgins2018-bucharest-memorandum-adopted> (accessed May 2121).
- Eurostat. 2009. *Harmonized European time use surveys, Guidelines 2008. Methodologies and Working Papers*. Luxembourg: Office for Official Publications of the European Communities. Available at: <https://ec.europa.eu/eurostat/documents/3859598/5909673/KS-RA-08-014-EN.PDF.pdf/a745ca2e-7dc6-48a9-a36c-000ad120380e?t=1414781526000> (accessed May 2121).
- Eurostat. 2019a. *Harmonised European Time Use Surveys (HETUS) – 2018 Guidelines*. Eurostat Manuals and Guidelines. Available at: <https://ec.europa.eu/eurostat/documents/3859598/9710775/KSGQ-19-003-EN-N.pdf/ee48c0bd-7287-411a-86b6-fb0f6d5068cc?t=1554468617000> (accessed 20 May 2121).
- Eurostat. 2019b. *European Big Data Hackathon*. Available at: https://ec.europa.eu/eurostat/cros/system/files/european_big_data_hackathon_2019_-_description_20181119.pdf (accessed April 2020).
- Ferneer, H., N. Sonck, and A. Scherpenzeel. 2013. “Data Collection with Smartphones: Experiences in a Time Use Survey.” NTTTS-Conferences on New Techniques and Technologies for Statistics. Brussels, 5–7 March 2013;868-875 Available at: https://ec.europa.eu/eurostat/cros/system/files/NTTS2013%20Proceedings_0.pdf.
- Ferneer, H., and N. Sonck. 2014. “Measuring Smarter – Time-Use Data Collected by Smartphones.” *International Journal of Time Use Research* 11(1): 94–111. DOI: <https://dx.doi.org/10.13085/eIJTUR.11.1.94-111>.
- Gershuny, J. 2015. *Time Use Research Methods*. International Encyclopedia of the Social and Behavioral Sciences (Second Edition) 24.
- Gershuny J., and O. Sullivan. 2019. *What We Really Do All Day: Insights from the Centre for Time Use Research*. Penguin Books Ltd.
- Giunchiglia, F., M. Zeni, E. Gobbi, E. Bignotti, and I. Bison. 2017. “Mobile Social Media and Academic Performance.” The 9th International Conference on Social Informatics (SocInfo 2017), September, 2017. Oxford, UK. DOI: https://doi.org/10.1007/978-3-319-67256-4_1.
- Giunchiglia, F., M. Zeni, E. Gobbi, E. Bignotti, and I. Bison. 2018. “Mobile social media usage and academic performance.” *Computers in Human Behavior* 82: 177–185. DOI: <https://doi.org/10.1016/j.chb.2017.12.041>.
- Groves, R.M. 2011. “Three Eras of Survey Research.” *Public Opinion Quarterly* 75(5): 861–71. DOI: <http://doi.org/10.1093/poq/nfr057>.
- GSM Association. 2018. *The Mobile Economy Europe 2018*. Available at: <https://www.gsmainelligence.com/research/?file=884c77f3bc0a405b2d5fd356689be340&download> (accessed April 2020).
- Hatuka, T., and E. Toch. 2017. “Being visible in public space: The normalisation of asymmetrical visibility.” *Urban Studies* 54(4): 984–998. DOI: <https://doi.dox.-org/10.1177/0042098015624384>.
- Hellgren, M. 2014. “Extracting More Knowledge from Time Diaries?” *Social Indicators Research* 119(3): 1517–1534. DOI: <https://doi.org/10.1007/s11205-013-0558-6>.
- Hewitt, Eben. 2010 *Cassandra: the definitive guide*. O’Reilly Media, Inc.

- i-Log. 2019. i-Log on the Google Play Store. Available at: <https://play.google.com/store/apps/details?id=it.unitn.disi.witmee.sensorlog> (accessed April 2020).
- Juster, F., and F. Stafford. 1991. "The Allocation of Time: Empirical Findings, Behavioral Models, and Problems of Measurement." *Journal of Economic Literature* 29(2): 471–522. Available at: <http://www.jstor.org/stable/2727521> (accessed May 2021).
- Juster, F., H. Ono, and F. Stafford. 2004. *Changing Times of American Youth: 1981–2003*. Institute for Social Research. University of Michigan. Available at: http://ns.umich.edu/Releases/2004/Nov04/teen_time_report.pdf (accessed May 2021).
- Kramarczyk, J. 2015. "Spending Time on Media – Results of Using Multitasking Frequency Questionnaire In Poland." *International Journal of Time Use Research* 12(1): 153–190. DOI: <https://doi.org/10.13085/eIJTUR.10.1.153-190>.
- Kramarczyk, J., and M. Osowiecka. 2014. "Time is Running Differently on the Internet." *International Journal of Time Use Research* 11(1): 94–111. DOI: <http://doi.org/10.13085/eIJTUR.11.1.94-111>.
- Link M.W., J. Murphy, M.F. Schober, T.D. Buskirk, J. Hunter Childs, T. Casey Langer. 2014. "Mobile Technologies for Conducting, Augmenting and Potentially Replacing Surveys: Executive Summary of the AAPOR Task Force on Emerging Technologies in Public Opinion Research." *Public Opinion Quarterly* 78: 779–87.
- Link, M. 2018. "New data strategies: nonprobability sampling, mobile, big data." *Quality Assurance in Education* 26(2): 303–314. DOI: <https://doi.org/10.1108/QAE-06-2017-0029>.
- Maddalena, E., L.-D. Ibáñez, E. Simperl, R. Gomer, M. Zeni, D. Song, and F. Giunchiglia. 2019. "Hybrid Human Machine workflows for mobility management." In Companion Proceedings of The 2019 World Wide Web Conference (WWW '19). May 2019, San Francisco, USA: 102–109. DOI: <https://doi.org/10.1145/3308560.3317056>.
- Merz J. 2009. *Time use and time budgets: Improvements, future challenges and recommendations*. Society for the Study of Economic Inequality ECINEQ 125.
- Miller, G.A. 1998. *WordNet: An electronic lexical database*. MIT press.
- Minnen, J., J. Verbeylen, and I. Glorieux. 2018. *Onderzoek naar de tijdsbesteding van leraren in het basis- en secundair onderwijs. Deel 1: Algemeen. (Time allocation of teachers in the primary and secondary school. Part 1: General)*. Vlaamse Overheid, Brussel: Vakgroep Sociologie, Onderzoeksgroep TOR 57 blz.
- Robinson, J.P. 1999. *The Time-Diary Method: Structure and Uses*. In *Time Use Research in the Social Sciences*. New York: Academic/Plenum Publishers.
- Robinson, J.P. 2002. "The time-diary method." *Time use research in the social sciences*: 47–89. DOI: https://doi.org/10.1007/0-306-47155-8_3.
- Runyan, J.D., T.A. Steenbergh, C. Bainbridge, D.A. Daugherty, L. Oke, and B.N. Fry. 2013. "A smartphone ecological momentary assessment/intervention 'app' for collecting real-time data and promoting self-awareness." *PLoS One* 8(8). DOI: <https://doi.org/10.1371/journal.pone.0071325>.
- Sorokin, P.A., and C.Q. Berger. 1939. *Time-budgets of human behavior*. Harvard University Press
- Stella C., K. Fisher, E. Gilbert, L. Calderwood, T. Huskinson, A. Cleary, and J. Gershuny. 2018. "Using new technologies for time diary data collection: Instrument design and

- data quality findings from a mixed-mode pilot survey.” *Social Indicators Research* 137(1): 379–390. DOI: <https://doi.org/10.1007/s11205-017-1569-5>.
- Sugie, N.F. 2018. “Utilizing Smartphones to Study Disadvantaged and Hard-to-Reach Groups.” *Sociological Methods and Research* 47(3): 458–491. DOI: <https://doi.org/10.1177/0049124115626176>.
- The Nielsen Company. 2018. *The Nielsen Total Audience Report Q1 2018*. Available at: <https://www.nielsen.com/content/dam/corporate/us/en/reports-downloads/2018-reports/q1-2018-total-audience-report.pdf> (accessed April 2020).
- United Nations. 2010. “In-depth review on time-use surveys, Economic Commission for Europe.” Conference of European Statisticians, Note by the German Federal Statistical Office. ECE/CES/2010/25. Paris, France, 2018. Available at: <http://unstats.un.org/unsd/demographic/sconcerns/tuse/> (accessed April 2020).
- Vilhelmson, B., E. Eildèr, and E. Thulin. 2018. “What did we do when the Internet wasn’t around? Variation in free-time activities among three young-adult cohorts from 1990/1991, 2000/2001, and 2010/2011.” *New Media and Society* 20(8): 2898–2916. DOI: <https://doi.org/10.1177/1461444817737296>.
- Vohra D. 2016. “Practical Hadoop Ecosystem.” Chapter in *Apache Parquet*: chap. 8. Apress. DOI: https://doi.org/10.1007/978-1-4842-2199-0_8.
- Wang, X.H., D.Q. Zhang, T. Gu, and H.K. Pung. 2004. “Ontology based context modeling and reasoning using OWL.” In IEEE annual conference on pervasive computing and communications workshops, March, 2004. 18–22. Orlando, FL, USA. DOI: <https://doi.org/10.1109/PERCOMW.2004.1276898>.
- Zaharia, M., R.S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, and M.J. Franklin. 2016. “Apache spark: a unified engine for big data processing.” *Communications of the ACM* 59(11): 56–65. DOI: <https://doi.org/10.1145/2934664>.
- Zeni, M., I. Zaihrayeu, and F. Giunchiglia. 2014. “Multi-device activity logging.” ACM International Joint Conference on Pervasive and Ubiquitous Computing. September 13–17, 2014. 299–302. Seattle, WA, USA. DOI: <http://dx.doi.org/10.1145/2638728.2638756/>.
- Zeni, M. 2017. *Bridging Sensor Data Streams and Human Knowledge*. Trento: University of Trento. Available at: <http://eprints-phd.biblio.unitn.it/2724/> (accessed April 2020).
- Zeni, M., W. Zhang, E. Bignotti, A. Passerini, and F. Giunchiglia. 2019. “Fixing Mislabeling by Human Annotators Leveraging Conflict Resolution and Prior Knowledge.” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3(1): 1–23. DOI: <https://doi.org/10.1145/3314419>.

Received June 2019

Revised December 2019

Accepted July 2020

A Diagnostic for Seasonality Based Upon Polynomial Roots of ARMA Models

*Tucker McElroy*¹

Methodology for seasonality diagnostics is extremely important for statistical agencies, because such tools are necessary for making decisions whether to seasonally adjust a given series, and whether such an adjustment is adequate. This methodology must be statistical, in order to furnish quantification of Type I and II errors, and also to provide understanding about the requisite assumptions. We connect the concept of seasonality to a mathematical definition regarding the oscillatory character of the moving average (MA) representation coefficients, and define a new seasonality diagnostic based on autoregressive (AR) roots. The diagnostic is able to assess different forms of seasonality: dynamic versus stable, of arbitrary seasonal periods, for both raw data and seasonally adjusted data. An extension of the AR diagnostic to an MA diagnostic allows for the detection of over-adjustment. Joint asymptotic results are provided for the diagnostics as they are applied to multiple seasonal frequencies, allowing for a global test of seasonality. We illustrate the method through simulation studies and several empirical examples.

Key words: Autoregressive estimator; seasonal adjustment; spectral peaks; visual significance.

1. Introduction

The problem of identifying seasonality in published time series is of enduring importance. Many official time series – such as gross domestic product (GDP) data – have an enormous impact on public policy, and are heavily scrutinized by economists and journalists. Obscuring the issue is the lack of universally agreed-upon criteria for detecting seasonality. Furthermore, the tools that critics use to assess seasonality (e.g., seasonal averages of growth rates, as in [Rudebusch et al. 2015](#)) sometimes differ from the diagnostics actually employed at statistical agencies, such as Visual Significance ([Soukup and Findley 1999](#); [McElroy and Roy 2017](#)), the Q_s diagnostic of Maravall ([Findley et al. 2017](#)), and the model-based F test ([Lytras et al. 2007](#)).

An overview of seasonality diagnostics that are currently available in popular statistical software is given in [Findley et al. \(2017\)](#); also see discussions in [Fase et al. \(1973\)](#) and [Den Butter and Fase \(1991\)](#). Desiderata for seasonality diagnostics include:

1. a rigorous statistical theory,
2. a precise correspondence between actual seasonal dynamics and diagnostic values,

¹ Research and Methodology Directorate, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100, U.S.A. Email: tucker.s.mcelroy@census.gov

Acknowledgments: This article is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the author and not those of the U.S. Census Bureau. All time series analyzed in this article are from public data sources.

3. applicability to diverse sampling frequencies,
4. applicability to multiple frequencies of latent seasonality, which is potentially of non-integer period, and
5. ability to address over- and under-adjustment.

Currently available diagnostics satisfy some of these criteria, but none address all facets.

With regard to the first point, a rigorous quantification of Type I and II errors is needed so that in a production environment with thousands of time series, an analyst can set tolerance levels for quality control. The distribution theory for diagnostic statistics should be developed under a broad set of time series data generating processes, so that critical values are approximately germane for typical sample sizes. As to the second point, we intend that a diagnostic takes a low value if and only if seasonality is present in a stochastic process; if low values could be obtained by non-seasonal processes, or high values could be generated by seasonal processes, then the diagnostic will be worse than useless – because it will generate false (spurious) conclusions. Thirdly, we observe that diagnostics developed for X-12-ARIMA (Findley et al. 1998) were intended for monthly data, with some extensions possible for quarterly series. However, there is an increasing public demand for the publication of weekly and daily time series (see McElroy et al. 2018 for an overview of the applications of daily time series to understanding retail holiday patterns), which presents new challenges for official statistics; moreover, multiple forms of seasonality (point number four) can be present. For example, daily time series have a weekly effect and an annual effect, which are really just two different types of seasonal effects. Note that for daily data, the annual period is 365.25, and monthly effects have an average period of 30.25 – these non-integer seasonal periods create new challenges for diagnostics based upon seasonal autocorrelations.

The fifth point is concerned with over- and under-adjustment of time series, for which there is a long literature: Granger (1978) noted with concern the introduction of negative seasonal autocorrelation into a time series by application of the Wiener-Kolmogorov (Bell 1984) seasonal adjustment filters, although the phenomenon had been already described in Nerlove (1964). Also see Sims (1978), Tukey (1978), and Bell and Hillmer (1984). If the extraction of seasonality involves using overly stable seasonal filters, then seasonality will remain – we refer to this as *under-adjustment*. On the other hand, using overly dynamic seasonal filters produces negative seasonal correlation – we refer to this as *over-adjustment*. Ansley and Wecker (1984) and McElroy (2012) discuss a method that reduces over-adjustment, while model-based diagnostics of under- and over-adjustment are described in Maravall (2003), McElroy (2008), and Blakely and McElroy (2017). Whereas under-adjustment is clearly a problem – since measurable seasonality remains (possibly marring interpretations of growth rates) – over-adjustment may also be undesirable, because it indicates that non-seasonal dynamics (such as the business cycle) may have been removed from the data and erroneously allocated to the seasonal component. This is akin to the problem of trend extraction: under-smoothing means the extracted trend will have too many oscillations, whereas over-smoothing will force long-term trend movements into the business cycle.

This article focuses on proposing a test for over- or under-adjustment, while allowing for non-integer periods of seasonality. First, we must clearly parse the phenomenon of seasonality. We propose – based on ideas developed in Lin et al. (2019) – the following

definition of seasonality: persistency in a time series over seasonal periods that is not explainable by intervening time periods. For a monthly series with a seasonal period equal to twelve, seasonality is indicated by persistency from year to year that is not explained by month-to-month changes. Note that both parts of this definition are crucial: without seasonal persistency from year to year, no seasonal pattern will be apparent, so this facet is clearly necessary; however, any trending time series also has persistency from year to year, which comes through the intervening months – we need to screen out such cases.

If the seasonality is non-stationary, there are diagnostic tools available in the econometric literature. Seasonal unit root tests (Hylleberg 1986; Hylleberg et al. 1990; Canova and Hansen 1995; Buseti and Harvey 2003) adopt either as null or alternative hypothesis that the form of persistence resembles a random walk year to year, for each season. Another framework involves periodically integrated processes (Franses 1994). Although such tests satisfy our first three criteria, these methods are not easily adapted to non-integer periods, and cannot address the over-adjustment problem; neither are they effective for diagnosing milder, dynamic seasonality that can be present in stationary processes. This latter application is vital for the detection of residual seasonality in seasonally adjusted data – seasonality in such series will not manifest unit roots, but rather a highly evolutive pattern that is consistent with a stationary formulation.

If a time series is covariance stationary, it is natural to parse persistency in terms of autocorrelation (cf. Proietti (1996), which measures the strength of autocorrelation at seasonal lags). As we show in this article, we can adapt persistency to non-integer lags of the autocovariance function (acvf) via its decomposition in terms of autoregressive (AR) roots, and examine seasonality of arbitrary frequency through the modulus and phase of the root. Whereas under-adjustment would be indicated by the presence of AR roots of near-unit magnitude and seasonal phase, over-adjustment corresponds to a negative form of persistency (i.e., negative seasonal autocorrelations) termed anti-persistency, and can be measured through moving average (MA) roots computed from the inverse autocovariances (McElroy and Roy 2018), that is, the autocovariances of the spectral density's reciprocal.

This framework of using AR roots as a diagnostic of under-adjustment (and MA roots for over-adjustment) satisfies the five criteria listed above. In Section 2 we develop the asymptotic theory and hypothesis testing framework, also demonstrating that small values of the diagnostic occur if and only if seasonality is present. (Appendix A of the Online Supplement offers a foundation for understanding oscillations and seasonality.) Section 3 applies the methodology, providing the implementation details; because seasonality becomes associated with the phase of AR roots, we can address arbitrary (regular) sampling frequencies and multiple non-integer period seasonalities. Simulation studies are given in Section 4, and data illustrations in Section 5. Section 6 concludes, with proofs and an additional illustration in the Supplement.

2. Methodology

2.1. Framework

Consider a weakly stationary process $\{X_t\}$ that is mean zero and purely unpredictable; then by the Wold Decomposition (Theorem 7.6.4 of McElroy and Politis 2020) $X_t =$

$\sum_{j=0}^{\infty} \psi_j Z_{t-j}$ for a white noise sequence $\{Z_t\}$, and the $\{\psi_j\}$ are called the Wold coefficients. Oscillatory behavior in the Wold coefficients corresponds to seasonality, as discussed in online Appendix A. That background discussion also shows that the oscillations of a sequence are governed by the magnitude and phase of the roots of the z -transform, that is, $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$. In particular, large values of $|\psi(\rho^{-1} e^{i\omega})|$ for $\rho \in (0, 1)$ correspond to an oscillatory pattern in the ψ_j , where the frequency of the oscillation is ω and the pattern is “damped” (i.e., it decays as $j \rightarrow \infty$) by the value ρ . The higher the value of ρ , the slower the decay in the coefficients, resulting in a more persistent oscillation.

Consider a seasonal pattern of period s (i.e., the number of seasons per year), which has frequency $\omega = 2\pi/s$. Writing $\pi(z) = 1/\psi(z)$ (which converges outside the unit circle if the process is invertible), we say there is ρ -persistent seasonality of frequency ω if $\pi(\rho^{-1} e^{i\omega}) = 0$. In the special case that $\{X_t\}$ is an ARMA process with MA polynomial $\theta(z)$ and AR polynomial $\phi(z)$, such a seasonal pattern exists if $\phi(\rho^{-1} e^{i\omega}) = 0$, because $\pi(z) = \phi(z)/\theta(z)$ (see Theorem 5.5.3 of McElroy and Politis 2020). Similarly, oscillatory patterns in the coefficients $\{\pi_j\}$ corresponding to $\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j$ correspond to anti-seasonality (see discussion in online Appendix A). We say there is ρ -persistent anti-seasonality of frequency ω if $\psi(\rho^{-1} e^{i\omega}) = 0$; in the case of an ARMA process, this occurs if $\theta(\rho^{-1} e^{i\omega}) = 0$, because $\psi(z) = \theta(z)/\phi(z)$ (see Theorem 5.4.3 of McElroy and Politis 2020).

For the remainder of the article we suppose that an invertible ARIMA model has been identified and fitted to a sample X_1, \dots, X_T of size T from the data process $\{X_t\}$. If the process is stationary, then no differencing is needed and we can fit an ARMA model; the AR(∞) representation of the data process is $\pi(z) = \phi(z)/\theta(z)$, where ϕ and θ are stable AR and MA polynomials (i.e., all of their roots are outside the unit circle) that are relatively prime. If the process requires differencing, then we allow the AR polynomial $\phi(z)$ to have roots on the unit circle. Let $\delta(z) = 1 - \sum_{j=1}^d \delta_j z^j$ be the unit root portion of the polynomial $\phi(z)$, and let $\varphi(z)$ (of degree p) correspond to the non-unit roots. Then the pseudo-autoregressive polynomial $\phi(z)$ is defined as $\delta(z) \varphi(z)$. Hence, if we difference the data with $\delta(B)$ (where B is the backward shift operator), the resulting process is a stationary ARMA with AR polynomial $\varphi(z)$ and MA polynomial $\theta(z)$.

2.2. Testing for Seasonality

Whether or not there are unit roots in $\phi(z)$, seasonality can be tested in terms of the polynomial $\phi(z)$, because $\pi(z) = \phi(z)/\theta(z)$. In particular, for any given ω , the null hypothesis – that ρ_0 -persistent seasonality of frequency ω is present – can be formulated as

$$H_0(\rho_0) : \pi(\rho_0^{-1} e^{i\omega}) = 0 \text{ has solution } r = \rho_0. \quad (1)$$

Note that $H_0(\rho_0)$ holds if and only if $\phi(\rho_0^{-1} e^{i\omega}) = 0$ for some $r = \rho_0$. We can measure departures from ρ_0 -persistent seasonality by computing $|\pi(\rho_0^{-1} e^{i\omega})|^2$, or its estimate based upon maximum likelihood estimation (MLE) of the ARMA parameters. Alternatively, in the case of an AR model we can compute ordinary least square estimates (OLS) of the parameters; the asymptotic theory is the same. Let $g(r) = |\pi(r^{-1} e^{i\omega})|^2$, and set

$$\hat{g}(r) = |\hat{\pi}(r^{-1} e^{i\omega})|^2, \quad (2)$$

where $\hat{\pi}(z) = \hat{\phi}(z)/\hat{\theta}(z)$ and the polynomials are estimated by replacing the coefficients with MLEs. Our test statistic of $H_0(\rho_0)$ is $T\hat{g}(\rho_0)$, where the rate T is justified by the subsequent asymptotic theory. Next, we present theory for this test statistic in the case that there are no unit roots in $\phi(z)$.

THEOREM 1. *Let $\{X_t\}$ be a causal invertible ARMA(p, q) process with AR polynomial $\phi(z)$ and MA polynomial $\theta(z)$. Let $\pi(z) = \phi(z)/\theta(z)$ and $\hat{g}(r)$ defined via (2), where the estimates for the ARMA parameters are obtained from a sample of size T via either MLE or OLS (in the pure AR case). It follows that when $g(r) = 0$*

$$T\hat{g}(r) \xrightarrow{L} \frac{|\underline{Z}'\underline{\zeta}|^2}{|\theta(r^{-1}e^{i\omega})|^2},$$

where $\underline{\zeta}_j = (re^{i\omega})^{-j}$ for $1 \leq j \leq p$ and $\underline{Z} \sim \mathcal{N}(0, \Gamma_p^{-1})$ such that Γ_p is the $p \times p$ Toeplitz covariance matrix corresponding to spectral density $|\phi(e^{-i\lambda})|^{-2}$. When $g(r) > 0$, instead

$$\sqrt{T}(\hat{g}(r) - g(r)) \xrightarrow{L} \mathcal{N}(0, V),$$

where $V = \underline{\eta}'F^{-1}\underline{\eta}$, F is the Fisher information matrix for the ARMA process (described in the proof), and

$$\underline{\eta} = \begin{bmatrix} -(\phi(r^{-1}e^{i\omega})\underline{\zeta} + \phi(r^{-1}e^{-i\omega})\bar{\underline{\zeta}})|\theta(r^{-1}e^{i\omega})|^{-2} \\ (\theta(r^{-1}e^{i\omega})\underline{\xi} + \theta(r^{-1}e^{-i\omega})\bar{\underline{\xi}})|\theta(r^{-1}e^{i\omega})|^{-4}|\phi(r^{-1}e^{i\omega})|^2 \end{bmatrix},$$

where $\underline{\xi}_j = (re^{i\omega})^{-j}$ for $1 \leq j \leq q$.

REMARK 1. The alternative hypothesis indicates that $g(\rho_0) > 0$, and Theorem 1 indicates that the test statistic is $O_p(T^{1/2})$ plus $Tg(\rho_0)$ in that case, yielding a consistent test.

In cases where the data may have unit roots, a different theory is needed. If a pure AR model is fitted, one can use OLS, as this allows for the parameters corresponding to unit or explosive roots. (The Yule-Walker method (McElroy and Politis 2020), which enforces stability, should be avoided because of substantial bias when the process has roots close to the unit circle.) Alternatively, one can apply $\delta(B)$ to difference the data, and then fit a stationary AR model. If fitting an ARIMA model, this latter strategy is used: apply $\delta(B)$, and then fit via MLE an ARMA model. The following result allows us to test a null hypothesis of stationary seasonality ($\rho_0 < 1$) when unit roots are known to be present in the process – testing $\rho_0 = 1$ requires a more complicated limit theory that is only mentioned in the proof.

THEOREM 2. *Let $\{X_t\}$ be an invertible ARIMA($p + d, q$) process with differencing polynomial $\delta(B)$, stable AR polynomial $\varphi(z)$, and MA polynomial $\theta(z)$. Set $\phi(z) = \delta(z)\varphi(z)$, the pseudo- autoregressive polynomial. Let $\pi(z) = \phi(z)/\theta(z)$ and $\hat{g}(r)$ defined via (2), where the estimates for the ARMA parameters are obtained from a sample of size T via*

either MLE or OLS (in the pure AR case). It follows that when $g(r) = 0$ and $r < 1$ that

$$T\hat{g}(r) \xrightarrow{\mathcal{L}} \frac{|\underline{Z}'\underline{\xi}|^2}{|\theta(r^{-1}e^{i\omega})|^2},$$

where $\underline{\xi}_j = (re^{i\omega})^{-j}$ for $1 \leq j \leq p + d$ and $\underline{Z} \sim \mathcal{N}(0, P'\Gamma_p^{-1}P)$ such that Γ_p is the $p \times p$ Toeplitz covariance matrix corresponding to spectral density $|\varphi(e^{-i\lambda})|^{-2}$ and P is a $p \times p + d$ -dimensional matrix given by

$$P = \begin{bmatrix} 1 & -\delta_1 & \dots & -\delta_d & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 1 & -\delta_1 & \dots & -\delta_d \end{bmatrix}. \tag{3}$$

When $g(r) > 0$, instead

$$\sqrt{T}(\hat{g}(r) - g(r)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V),$$

where $V = \underline{\eta}'R'F^{-1}R\underline{\eta}$, $R = \text{diag}\{P, I_q\}$ (here I_q is a q -dimensional identity matrix), F is the Fisher information matrix for the stationary ARMA process (described in the proof), and

$$\underline{\eta} = \begin{bmatrix} -(\phi(r^{-1}e^{i\omega})\underline{\xi} + \phi(r^{-1}e^{-i\omega})\underline{\bar{\xi}})|\theta(r^{-1}e^{i\omega})|^{-2} \\ (\theta(r^{-1}e^{i\omega})\underline{\xi} + \theta(r^{-1}e^{-i\omega})\underline{\bar{\xi}})|\theta(r^{-1}e^{i\omega})|^{-4}|\phi(r^{-1}e^{i\omega})|^2 \end{bmatrix},$$

where $\underline{\xi}_j = (re^{i\omega})^{-j}$ for $1 \leq j \leq q$.

2.3. Testing for Anti-Seasonality

Suppose now that we wish to test for the presence of anti-seasonality. Now we wish to examine $\psi(\rho^{-1}e^{i\omega})$, and it is important that the process be stationary. Therefore suppose that $\{X_t\}$ is an ARMA process, where any non-stationary effects have been previously removed by a differencing polynomial $\delta(B)$. Then, for any given ω , the null hypothesis of ρ_0 -persistent anti-seasonality is written

$$H_0(\rho_0) : \pi(r^{-1}e^{i\omega}) = 0 \text{ has solution } r = \rho_0, \tag{4}$$

where $\psi(z) = \theta(z)/\phi(z)$. Here $H_0(\rho_0)$ holds if and only if $\theta(r^{-1}e^{i\omega}) = 0$ for some $r = \rho_0$. We can measure departures from ρ_0 -persistent anti-seasonality by computing $|\psi(\rho_0^{-1}e^{i\omega})|^2$; set $h(r) = |\psi(r^{-1}e^{i\omega})|^2$, and let

$$\hat{h}(r) = |\hat{\psi}(r^{-1}e^{i\omega})|^2. \tag{5}$$

Here $\hat{\psi}(z) = \hat{\theta}(z)/\hat{\phi}(z)$, and the polynomials are estimated by replacing the coefficients with MLEs. Theorem 1 can likewise be adapted by swapping the polynomials appropriately, as stated below. (The proof follows the same techniques, and is therefore omitted.) We emphasize that this theory requires an invertible moving average polynomial, and thus cannot be used to test an over-adjustment hypothesis where $\rho_0 = 1$.

COROLLARY 1. Let $\{X_t\}$ be an invertible ARMA(p, q) process with AR polynomial $\phi(z)$ and MA polynomial $\theta(z)$. Let $\psi(z) = \theta(z)/\phi(z)$ and $h(r)$ defined via (5), where the estimates for the ARMA parameters are obtained from a sample of size T via either MLE or OLS (in the pure AR case). It follows that when $h(r) = 0$

$$T\hat{h}(r) \xrightarrow{L} \frac{|\underline{Z}'\underline{\xi}|^2}{|\phi(r^{-1}e^{i\omega})|^2},$$

where $\underline{\xi}_j = (re^{i\omega})^{-j}$ for $1 \leq j \leq q$ and $\underline{Z} \sim \mathcal{N}(0, \Gamma_q^{-1})$ such that Γ_q is the $q \times q$ Toeplitz covariance matrix corresponding to spectral density $|\theta(e^{-i\lambda})|^{-2}$. When $h(r) > 0$, instead

$$\sqrt{T}(\hat{h}(r) - h(r)) \xrightarrow{L} \mathcal{N}(0, V),$$

where $V = \underline{\eta}'F^{-1}\underline{\eta}$, F is the Fisher information matrix for the ARMA process, and

$$\underline{\eta} = \begin{bmatrix} -(\phi(r^{-1}e^{i\omega})\underline{\zeta} + \phi(r^{-1}e^{-i\omega})\underline{\bar{\zeta}})|\phi(r^{-1}e^{i\omega})|^{-4}|\theta(r^{-1}e^{i\omega})|^2 \\ (\theta(r^{-1}e^{i\omega})\underline{\xi} + \theta(r^{-1}e^{-i\omega})\underline{\bar{\xi}})|\phi(r^{-1}e^{i\omega})|^{-2} \end{bmatrix},$$

where $\underline{\zeta}_j = (re^{i\omega})^{-j}$ for $1 \leq j \leq p$.

A limitation of the methodology behind Corollary 1 is that it cannot be applied to non-invertible processes; to rectify this, we proceed by considering instead of $h(r)$ the related quantity

$$f(z) = \sum_{|h| \leq q} \gamma_h z^h = \psi(z)\psi(z^{-1})\sigma^2 \tag{6}$$

for $z \in \mathbb{C}$, where $\gamma_h = \text{Cov}[X_{t+h}, X_t]$ is the autocovariance function of the stationary process $\{X_t\}$. We evaluate Equation (6) at $z = r^{-1}e^{i\omega}$. Although $f(r^{-1}e^{i\omega})$ is not equal to $h(r)$ (in fact, it is complex-valued when $r \neq 1$), we still have $f(r^{-1}e^{i\omega}) = 0$ if and only if $\psi(r^{-1}e^{i\omega}) = 0$. This suggests basing a test statistic on

$$\hat{f}(z) = \sum_{|h| \leq q} \hat{\gamma}_h z^h, \tag{7}$$

where $\hat{\gamma}_h$ is the sample autocovariance based on a sample of size T . Although $\hat{f}(z)$ is not real-valued (and need not be positive-definite when $z = e^{-i\lambda}$, because the choice of q implicitly generates a truncation taper), a distribution theory is easily developed based upon the sample autocovariances, and a central limit theorem can be established. Hence, we propose to test the null $H_0(\rho_0)$ of anti-seasonality with $T|\hat{f}(\rho_0^{-1}e^{i\omega})|^2$. The following theory describes the asymptotic distribution, and allows for $r = 1$ without any qualitative change to the results.

THEOREM 3. Let $\{X_t\}$ be a possibly non-invertible MA(q) process with independent and identically distributed inputs and moving average polynomial $\theta(z)$. With $f(z)$ and $\hat{f}(z)$ defined via Equations (6) and (7) for any $z \in \mathbb{C}$,

$$T|\hat{f}(z) - f(z)|^2 \xrightarrow{L} |\underline{Z}'\underline{v}|^2,$$

where $\underline{v}' = [1, z + z^{-1}, \dots, z^q + z^{-q}]$ and $\underline{Z} \sim \mathcal{N}(0, V)$ such that V is the $q + 1 \times q + 1$ asymptotic covariance matrix of the sample autocovariances at lags 0 through q , that is, the jk th entry (for $0 \leq j, k \leq q$) is given by

$$V_{jk} = \frac{2}{2\pi} \int_{-\pi}^{\pi} \cos(\lambda j) \cos(\lambda k) f(e^{-i\lambda})^2 d\lambda \\ + \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \cos(\lambda j) \cos(\omega k) \Omega(e^{i\lambda}, e^{-i\omega}, e^{i\omega}) d\lambda d\omega,$$

where Ω is the tri-spectral density given in (B.2).

REMARK 2. Under the null hypothesis, with $z = \rho_0^{-1} e^{i\omega}$, the test statistic converges to $|\underline{Z}' \underline{v}|^2$, whose critical values can be simulated. If the inputs have zero kurtosis then the tri-spectral density is zero, and we can estimate V by utilizing $\hat{f}(e^{-i\lambda})$ as a plug-in estimator of $f(e^{-i\lambda})$, the spectral density; alternatively, we can plug in the periodogram and divide by 2, approximating the integral by a Riemann sum over Fourier frequencies – see Chiu (1988) and Deo and Chen (2000). We find that in simulations, the size of this latter approach is far superior. The alternative hypothesis indicates that $f(\rho_0^{-1} e^{i\omega})$ is non-zero; asymptotically the test statistic equals $T |f(\rho_0^{-1} e^{i\omega})|^2$ plus lower-order terms, and this quantity will generate power (because it is $O(T)$).

3. Applications and Implementation

In the practice of seasonal adjustment one of the key tasks is to decide whether a given time series should be seasonally adjusted. Secondly, series that have been seasonally adjusted should be assessed for adequacy, and among the potential problems it is important to determine whether the series have been under-adjusted (or over-adjusted). As discussed in the Introduction (with further exposition in online Appendix A), under-adjustment is characterized by the presence of dynamic seasonality, whereas over-adjustment is characterized by the presence of anti-seasonality. Therefore, there are three potential applications of the testing methodology of Section 2:

1. a test for residual seasonality (i.e., a test of under-adjustment), used upon stationary data that has already been adjusted (or clearly has no unit roots) and trend-differenced, if needed,
2. a test for raw seasonality, used upon potentially non-stationary data with unit roots, and
3. a test of over-adjustment, used upon seasonally adjusted data where there is concern about residual anti-seasonality.

For these three cases, we propose using the test statistics discussed in Theorem 1, Theorem 2, and Theorem 3 respectively, using the distribution theory for the null hypothesis to obtain critical values. In particular, cases one and two utilize the test statistic $T \hat{g}(\rho_0)$ given by (2) to test $H_0(\rho_0)$ given by (1). The third case utilizes the test statistic $T |\hat{f}(\rho_0^{-1} e^{i\omega})|^2$ given by (7) to test $H_0(\rho_0)$ given by (4). Note that these null hypotheses can be calibrated according to the concerns and priorities of the seasonal adjuster, through the determination of ρ_0 .

All of these tests are upper one-sided, with large values of the test statistic indicating rejection of $H_0(\rho_0)$. However, we do not have a way of knowing whether the true seasonal

persistence (or anti-persistence) – if it exists, because possibly $\pi(r^{-1}e^{i\omega}) \neq 0$ for all r – is greater than or less than ρ_0 . To address this question, we can consider computing both the test statistic and p -values for a range of values of $\rho_0 \in (0,1)$. Because both the test statistic and the critical values are continuous functions of ρ_0 , we thereby obtain p -values as a continuous function of ρ_0 , say $p(\rho_0)$. If desired, we can invert this function, for example obtaining $p^{-1}(\alpha,1)$ as the open set of ρ_0 such that we fail to reject $H_0(\rho_0)$ at confidence level α . Of course, it may happen that the p -values are low for all values of ρ_0 , and such intervals are empty; then there is no seasonality of any degree of persistence present. For other applications, it is useful to plot $p(\rho_0)$ for $\rho_0 \in (0,1)$, or for ρ in some sub-interval of $(0,1)$ that corresponds to the degrees of seasonal persistency that are deemed to be of interest. If $p(\rho) < \alpha$ for all ρ_0 in the given interval, then we can reject seasonality at all those persistencies.

If it is desired to obtain a joint test over J different frequencies $\omega_1, \dots, \omega_J$, then the following approach can be used. We may have a null hypothesis for each frequency ω_j of persistency $\rho_0^{(j)}$, and we say that the null hypothesis for seasonality across all frequencies holds if and only if all the individual null hypotheses are valid:

$$H_0(\rho_0^{(1)}, \dots, \rho_0^{(J)}) : \pi(\rho_0^{(j)} e^{i\omega_j}) = 0 \quad \text{for all } 1 \leq j \leq J. \tag{8}$$

(In the case of testing for anti-seasonality, we examine $\psi(\rho_0^{(j)} e^{i\omega_j})$ instead of $\pi(\rho_0^{(j)} e^{i\omega_j})$.) By taking the minimum of the various test statistics, a rejection (of seasonality) occurs if and only if all of the individual test statistics are significant; we fail to reject if at least one test statistic is small, that is, there is at least one j for which $\rho_0^{(j)}$ -persistent seasonality exists at frequency ω_j . This indicates that the minimum is an appropriate statistic; for testing seasonality we have the joint test statistic

$$\min_{1 \leq j \leq J} T |\hat{\pi}(e^{i\omega_j} / \rho_0^{(j)})|^2, \tag{9}$$

whereas for anti-seasonality testing we use

$$\min_{1 \leq j \leq J} T |\hat{\psi}(e^{i\omega_j} / \rho_0^{(j)})|^2. \tag{10}$$

The critical values are easily obtained by simulation, as the results of Theorems 1, 2, and 3 are clearly joint across various ω_j . If we set each $\rho_0^{(j)}$ equal to a common ρ_0 for each of the component null hypotheses, then we obtain a test statistic and critical value as a function of a single number ρ_0 , and hence we can obtain $p(\rho_0)$ in the manner described for a single frequency.

Each application will dictate the frequencies ω that should be considered. For a time series with s seasons per year (or other appropriate unit of time), one should examine $\omega = 2\pi j/s$ for $1 \leq j \leq \lfloor s/2 \rfloor$. For monthly data there are six frequencies to test; for quarterly data there are two frequencies to test. (Though commonly, the last frequencies – which are equal to π – are ignored, so that either five or one frequencies are actually considered for monthly and quarterly data.) A daily series has both a weekly and annual periodicity, and hence $\omega = 2\pi j/7$ for $1 \leq j \leq 3$ and $\omega = 2\pi j/365.25$ for $1 \leq j \leq 182$ are potential frequencies of interest. Given the large number of test statistics with a joint dependence structure – which, however, test different null hypotheses – some discretion

is needed when combining into a single test of seasonality, as the significance levels may become modified.

In practice, when testing for seasonality we must specify an ARMA model and fit it via MLE. Our own implementation instead fits a high order AR(p) model via OLS, with p selected via some empirical means – such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) discussed in [McElroy and Politis \(2020\)](#). When the order p is over-specified, but fixed, the test statistics are still asymptotically correctly sized – though there may be a loss of power in finite samples, due to the inefficiency of estimating some AR coefficients that are truly zero. However, if p is allowed to grow with sample size (as with the use of AIC and BIC), then there is some probability of getting a high order model, which unduly adds to the statistic's variability, thereby lowering power. On the other hand, a value of p that is too small implies a mis-specification, causing bias in the test statistics. Whereas BIC is consistent for model order, AIC is upward-biased, tending to over-specify models. Because in this particular case over-specification is less hazardous to the asymptotic theory, the case can be made that AIC is preferable – and this surmise is borne out by our simulation studies.

When testing for anti-seasonality we need to determine the order q of the MA representation of the differenced seasonally adjusted data. Again, an information criterion could be used in conjunction with maximum likelihood fitting of the MA model – however, in the case that $\rho_0 = 1$ the moving average process is non-invertible and the usual asymptotic theory upon which AIC and BIC are founded cannot be applied ([Tanaka 1996](#)). Alternatively, we can let \hat{q} be the largest q such that the sample autocovariances at all lags higher than q fail to reject the null hypothesis that the corresponding autocovariance is zero; we can use the variance estimate given in Paradigm 10.1.2 of [McElroy and Politis \(2020\)](#) to obtain a studentized test statistic. In our implementation we use normal critical values corresponding to a significance level of $1/\sqrt{T/3}$, which allow for the test's size to dwindle as sample size increases – thereby ensuring that q increases with T in an empirical fashion.

A final consideration is that components – such as the seasonal adjustment or the irregular – that are the output of a seasonal adjustment procedure will typically have nonlinear distortions in the beginning and final portions of the sample, due to forecast extension and/or adaptive filters; because the impact of forecast extension is localized to the edges of the sample, one can trim the component of the first and last 2 to 3 years – see [Findley et al. \(1998\)](#). In summary, we propose the following sequential procedure for de-seasonalizing a time series:

1. Test the raw series for seasonality by fitting an ARIMA model without seasonal differencing, testing the null (8) for $\rho_0 \in (.97, 1)$ with test statistic (9) for desired seasonal frequencies,
2. If dynamic seasonality is present (indicated by failure to reject in step 1 above, for at least one $\rho_0 \in (.97, 1)$), seasonally adjust the time series and proceed to steps 3 and 4 (otherwise, the procedure is complete),
3. Test the seasonally adjusted series for seasonality by trimming the first and last three years of data (to reduce the impact of non-linearity at the sample boundaries) and

- fitting an ARIMA model, testing the null (8) for $\rho_0 \in (.9, 1)$ with test statistic (9) for desired seasonal frequencies, and
4. Test the seasonally adjusted series for anti-seasonality by trimming the first and last three years of data (to reduce the impact of non-linearity at the sample boundaries) and testing the joint null – given by swapping ψ for π in (8) – for $\rho_0 \in (.5, 1)$ with test statistic (10) for desired seasonal frequencies.

We remark that critical values for tests are generated through Monte Carlo simulation, using the asymptotic distributions given in Section 2; R code is available from the author. The recommended seasonal frequencies are $\pi/2$ for quarterly data, and $\omega_j = \pi j/6$ (for $1 \leq j \leq 5$) for monthly data. The range of ρ_0 considered in the seasonality tests for raw data are taken fairly high, the interval $(.97, 1)$ being a suggestion based on the discussion in online Appendix A; the idea here is that seasonal adjustment should not be undertaken unless a fairly substantial degree of dynamic seasonality is present. However, some practitioners have suggested seasonally adjusting time series in which a much milder degree of seasonality is present (i.e., broaden the range of ρ_0), to mitigate the possibility of seasonality manifesting in aggregations of seasonally adjusted series – which has been documented in [Moulton and Cowan \(2016\)](#).

In step 3, we instead consider a broader interval of $(.9, 1)$, so that we are more concerned about the presence of dynamic seasonality in the adjustment, as compared to the raw. Alternatively, one could make this range consistent with step 1, in this example setting the interval to $(.97, 1)$. Similarly, in step 4 we screen out even mild cases of anti-seasonality by setting the interval to $(.5, 1)$. Finally, we note that some practitioners may not be concerned about the presence of anti- seasonality, since it is a necessary outcome of model-based seasonal adjustment; such users could just omit step 4.

After completing these four steps, either the analyst is satisfied with the outcome (i.e., either the series needs no adjustment, or it does, and its seasonal adjustment is deemed to be adequate) or there is some deficiency, e.g., under- or over-adjustment. In this latter case, the analyst may wish to re-examine the modeling of the time series – frequently, a different specification of outlier effects, or a change to the ARIMA model, can result in an improved seasonal adjustment. The case of over-adjustment is harder to address, possibly requiring a different set of seasonal adjustment filters that produce narrower seasonal troughs in the spectral density; this is in contrast to the case of under-adjustment, where a more dynamic filter (with a wider seasonal spectral trough) is needed.

4. Numerical Experiments

To discern the efficacy of the method in finite samples, we consider simulating from a few different processes.

4.1. Atomic Seasonality With Transient Noise

We study monthly Gaussian time series $\{X_t\}$ generated from an AR(3) model with AR polynomial

$$\phi(z) = (1 - \tau z)(1 - 2\rho \cos(\pi/6)z + \rho^2 z^2). \quad (11)$$

This process corresponds to an atomic seasonality (i.e., there is a single seasonal frequency involved) at the first seasonal frequency, muddled by the presence of a transient effect. The acvf and spectrum are plotted in Figure 1, where we have set $\tau = .8$ and $\rho = .9$. The true AR roots have magnitudes of 1.25 (for the real root, corresponding to the transient effect) and 1.11 (for the complex roots corresponding to the atomic seasonality). From the plots, it is apparent that the strong seasonality is somewhat attenuated by the transient effect, so the impact of the atomic seasonality is weaker than it would be if $\tau = 0$. As a second example, we lower the seasonal persistency to $\rho = .8$, and dampen the transient component by setting $\tau = .3$, displayed in Figure 2. Here the weak seasonality is apparent, no longer being obfuscated by a transient effect.

For both of these processes (Figures 1 and 2) we generate 10,000 Gaussian simulations for each sample size $T = 12n$, where n is the number of years and $n = 5, 10, 15, 20$. We apply our procedure with $\omega = \pi/6$ under four different scenarios: first, assuming that the AR order $p = 3$ is known in the calculation of the test statistic (based on fitting an AR(3) to each simulation) – in this case, we use critical values from the true AR(3) process. Second, a more realistic scenario determines the critical values from the AR(3) fitted to each simulation, but still assumes the true order $p = 3$ is known. Third, we use an over-specified order ($p = 24$, which is twice the number of seasons) to compute the test statistic and critical values. Fourth, we use AIC to identify the order p . We also explored the use of BIC and an AR identification rule given in McElroy and Politis (2020, 335), but these yielded much more badly mis-sized results and were not pursued further. In each case the critical

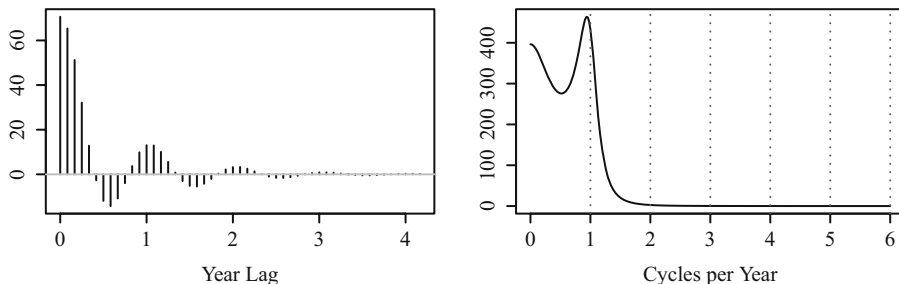


Fig. 1. Autocovariance function (left panel) and spectral density (right panel) for a seasonal AR(3) process ($\rho = .9$) with transient effect ($\tau = .8$). Autocovariance function is plotted as a function of lag divided by 12; spectral density is plotted as a function of cycles per year, or 12 divided by period.

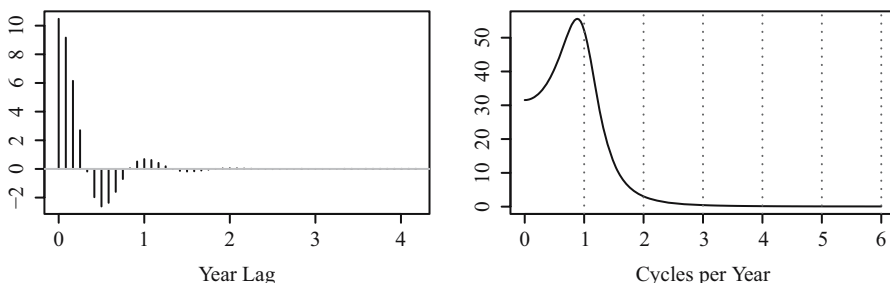


Fig. 2. Autocovariance function (left panel) and spectral density (right panel) for a seasonal AR(3) process ($\rho = .8$) with transient effect ($\tau = .3$). Autocovariance function is plotted as a function of lag divided by 12; spectral density is plotted as a function of cycles per year, or 12 divided by period.

values are generated for $\alpha = .10, .05, .01$; for the second two cases, we report the proportion of p-values that are less than α .

We examine the size of the procedure by taking as null hypothesis that $\rho = .9$ and $\rho = .8$ respectively, for the two processes. We also examine the power in both directions: with the first process, we adopt the null hypothesis that $\rho = .8$, and with the second process we adopt the null hypothesis that $\rho = .9$. The size and power results for the first process are given in [Tables 1 and 3](#), where the alternative entails a greater degree ($\rho = .9$) of

Table 1. Size simulations from an AR(3) DGP (corresponding to [Figure 1](#)) based on a null hypothesis of .9-persistent seasonality at frequency $\pi/6$. Results are for known AR order (first three rows), for unknown parameters (second three rows), over-specified AR order (third three rows), and AIC-determined AR order (last three rows).

α	5 years	10 years	15 years	20 years
.10	.149	.116	.113	.101
.05	.092	.067	.061	.054
.01	.039	.026	.021	.018
<hr/>				
.10	.139	.115	.108	.100
.05	.084	.059	.058	.048
.01	.028	.014	.012	.011
<hr/>				
.10	.381	.206	.152	.140
.05	.274	.133	.089	.079
.01	.135	.050	.026	.020
<hr/>				
.10	.625	.340	.183	.139
.05	.582	.297	.131	.086
.01	.508	.258	.087	.035

Table 2. Size simulations from an AR(3) DGP (corresponding to [Figure 2](#)) based on a null hypothesis of .8-persistent seasonality at frequency $\pi/6$. Results are for known AR order (first three rows), for unknown parameters (second three rows), over-specified AR order (third three rows), and AIC-determined AR order (last three rows).

α	5 years	10 years	15 years	20 years
.10	.135	.115	.109	.104
.05	.077	.058	.057	.053
.01	.021	.012	.013	.011
<hr/>				
.10	.136	.114	.107	.103
.05	.074	.060	.057	.055
.01	.020	.014	.011	.012
<hr/>				
.10	.424	.192	.152	.132
.05	.323	.117	.086	.072
.01	.157	.040	.023	.018
<hr/>				
.10	.413	.483	.458	.377
.05	.297	.375	.374	.308
.01	.135	.182	.217	.192

Table 3. Power simulations from an AR(3) DGP (corresponding to Figure 1) with null hypothesis of .8-persistent seasonality at frequency $\pi/6$. Results are for known AR order (first three rows), for unknown parameters (second three rows), over-specified AR order (third three rows), and AIC- determined AR order (last three rows).

α	5 years	10 years	15 years	20 years
.10	.502	.758	.908	.961
.05	.304	.604	.820	.918
.01	.060	.230	.509	.719
.10	.465	.707	.852	.928
.05	.320	.565	.751	.865
.01	.107	.281	.476	.659
.10	.437	.207	.151	.134
.05	.328	.130	.087	.074
.01	.170	.045	.024	.020
.10	.787	.685	.677	.740
.05	.721	.578	.554	.625
.01	.611	.395	.321	.365

persistence than is hypothesized ($\rho = .8$). As for the second process, the size and power results are given in Tables 2 and 4, where the alternative entails a lesser degree ($\rho = .8$) of persistence than is hypothesized ($\rho = .9$).

In general, the size is over-estimated but may be considered adequate for 10 years of data when p is known; when p is fixed and over-specified, at least 20 years of data is needed. Moreover, the use of AIC results in badly mis-sized test statistics in the case of the second process, but AIC works adequately with 20 years of data in the case of the first

Table 4. Power simulations from an AR(3) DGP (corresponding to Figure 2) with null hypothesis of .9-persistent seasonality at frequency $\pi/6$. Results are for known AR order (first three rows), for unknown parameters (second three rows), over-specified AR order (third three rows), and AIC- determined AR order (last three rows).

α	5 years	10 years	15 years	20 years
.10	.322	.506	.659	.786
.05	.230	.386	.541	.681
.01	.108	.205	.321	.460
.10	.281	.503	.683	.804
.05	.181	.365	.554	.694
.01	.054	.149	.292	.440
.10	.409	.204	.160	.142
.05	.306	.133	.092	.083
.01	.155	.049	.025	.023
.10	.248	.421	.605	.740
.05	.166	.318	.502	.644
.01	.062	.151	.285	.420

process. Power is good for both processes when p is known, but drops considerably when p is over-specified, as expected. This occurs because the additional variability due to over-specification overwhelms the $T g(\rho_0)$ quantity in small samples. In these two cases where H_a holds, $g(\rho_0)$ is given by .0048 and .0059 respectively.

4.2. Multiple Seasonal Peaks

Next, we consider a more nuanced process where there is apparent either a mild or intense degree of seasonality at one or more seasonal frequencies. The process $\{X_t\}$ is a monthly Gaussian AR(10) time series with AR polynomial defined by

$$\phi(z) = \prod_{j=1}^5 \left(1 - 2\rho^{(j)} \cos(\pi j/6)z + \rho^{(j)2} z^2 \right). \tag{12}$$

As a result, $g(r) = 0$ for $r = \rho^{(j)}$, for each $\omega_j = \pi j/6$ ($1 \leq j \leq 5$). In Table 5 we report size results for the process corresponding to $\rho_0^{(j)} = .9$ for $1 \leq j \leq 5$ (again with 10,000 simulations for each sample size), where the joint seasonality test is utilized by taking the minimum of the frequency-specific test statistics. As above, we consider three scenarios regarding the model order p , either using the true $p = 10$ or the over-specified $p = 24$. In each case the critical values are generated for $\alpha = .10, .05, .01$; for the second two cases, we report the proportion of p-values that are less than α .

For power, we first consider the same process but change the null hypothesis to $\rho_0^{(j)} = .97$, with results reported in Table 6. Here $g(.97)$ takes the values .399, .135, .102, .135, and .399 respectively for $\omega_j = \pi j/6$ ($1 \leq j \leq 5$). This situation corresponds to where the actual degree of seasonality is milder than is hypothesized, with rejections indicating no seasonality is present. As a second study, we adopt this same null hypothesis but now the true process is non-stationary with mingled seasonal persistencies:

Table 5. Size simulations from a stationary AR(10) DGP with null hypothesis of .9-persistent seasonality at various frequencies $\pi j/6$, $1 \leq j \leq 5$. Results are for known AR order (first three rows), for unknown parameters (second three rows), over-specified AR order (third three rows), and AIC-determined AR order (last three rows).

α	5 years	10 years	15 years	20 years
.10	.124	.114	.105	.107
.05	.068	.054	.053	.053
.01	.017	.011	.012	.011
.10	.169	.129	.118	.107
.05	.104	.070	.063	.053
.01	.035	.017	.017	.012
.10	.395	.169	.139	.122
.05	.291	.099	.077	.064
.01	.147	.028	.021	.014
.10	.577	.271	.179	.177
.05	.477	.201	.117	.109
.01	.285	.112	.043	.035

Table 6. Power simulations from a stationary AR(10) DGP with null hypothesis of .97-persistent seasonality at various frequencies $\pi j/6$, $1 \leq j \leq 5$. Results are for known AR order (first three rows), for unknown parameters (second three rows), over-specified AR order (third three rows), and AIC-determined AR order (last three rows).

α	5 years	10 years	15 years	20 years
.10	.447	.818	.955	.990
.05	.333	.732	.922	.981
.01	.172	.537	.810	.936
.10	.509	.859	.967	.994
.05	.394	.778	.941	.986
.01	.195	.574	.842	.950
.10	.427	.394	.531	.660
.05	.330	.294	.421	.560
.01	.178	.143	.231	.353
.10	.776	.838	.942	.979
.05	.737	.762	.907	.967
.01	.635	.577	.794	.919

$\rho_0^{(1)} = \rho_0^{(2)} = \rho_0^{(3)} = 1$, $\rho_0^{(4)} = \rho_0^{(5)} = .9$. The results are reported in Table 7; we omit entries for the case of known parameters, since the true process is non-stationary. The values of the functional $g(.97)$ are .121, .040, .031, .247, and .742 respectively for $\omega_j = \pi j/6$ ($1 \leq j \leq 5$). This situation corresponds to where the actual degree of seasonality is greater than is hypothesized for the first three seasonal frequencies, with rejections indicating strong seasonality is present.

The size results are similar to the cases of atomic seasonality considered previously: when the AR order is known, at least 10 years of data is needed, but 20 years may be needed if one uses AIC or an over-specified order. Power against the stationary alternative is above 50% with 10 years of data if the order is known, but there is a loss of power in the

Table 7. Power simulations from a non-stationary AR(10) DGP with null hypothesis of .97-persistent seasonality at various frequencies $\pi j/6$, $1 \leq j \leq 5$. Results are for unknown parameters (first three rows), over-specified AR order (second three rows), and AIC-determined AR order (last three rows).

α	5 years	10 years	15 years	20 years
.10	.939	.988	.996	.999
.05	.918	.982	.996	.998
.01	.866	.970	.991	.997
.10	.738	.863	.939	.965
.05	.660	.803	.913	.951
.01	.478	.661	.841	.915
.10	.912	.979	.991	.995
.05	.883	.971	.986	.993
.01	.805	.948	.975	.989

over-specified case (as expected). Therefore the results for multiple peaks are qualitatively similar to those of a single peak. For the case of a non-stationary alternative, the power is much higher, even in the over-specified case; this indicates it is easier to differentiate between non-stationary seasonality and strong stationary seasonality, versus discriminating between various cases of stationary seasonality.

4.3. Completely Non-Seasonal

Next, consider the case of an AR(1) process of parameter .8, which we suppose is observed as a quarterly series. Because this process is clearly non-seasonal, we should expect our AR diagnostic to have high power when we set the null hypothesis at a moderate degree of seasonality. In this case we set $\rho_0 = .9$ for the null, and consider frequency $\omega = \pi/2$, the quarterly frequency (again generating 10,000 simulations for each sample size). For comparison, we also investigate the Q_s statistic of Maravall (Findley et al. 2017): because the AR(1) process is non-seasonal, we expect the proportion of p-values less than a given α to be approximately α . However, the actual autocorrelations are $\rho_4 = .41$ and $\rho_8 = .17$, which are substantially different from zero – and Q_s is predicated upon $\rho_4 = \rho_8 \leq 0$ as an appropriate metric (i.e., necessary and sufficient) for non-seasonality. Table 8 indicates that Q_s is mis-sized with five years of data, and has size surpassing 50% as the sample size increases. This merely illustrates that Q_s tends to flag such non-seasonal processes as seasonal, because it fails to account for seasonal lag correlation that is explained through the linkages of the other seasons. By way of contrast, the AR diagnostic has 100% power for all the settings given in Table 8, demonstrating that the new test correctly classifies the AR(1) process as non-seasonal.

4.4. Testing for Over-Adjustment

To study the test for over-adjustment, we consider a process with anti-seasonality at frequency $\pi/6$, that is, such that $f(\rho^{-1}e^{i\pi/6}) = 0$. Such a condition is satisfied by an MA(2) model with MA polynomial

$$\theta(z) = 1 - 2\rho \cos(\pi/6)z + \rho^2z^2.$$

The inverse acvf and spectrum are plotted in Figure 3, where we have set $\rho = .9$. The true MA roots have magnitudes of 1.11 (for the complex roots corresponding to the anti-seasonality). From the plots, it is apparent that the strong anti-seasonality exists due to the trough, as well as the oscillations of the inverse acvf.

We generate 10,000 Gaussian simulations of each sample size $T = 12n$, for $n = 5, 10, 15, 20$, and apply the over-adjustment test based upon the sample autocovariances. For

Table 8. Size simulations from an AR(1) DGP for the Q_s diagnostic, based on a null hypothesis of no seasonality.

α	5 years	10 years	15 years	20 years
.10	.012	.286	.509	.660
.05	.000	.225	.436	.590
.01	.000	.125	.300	.456

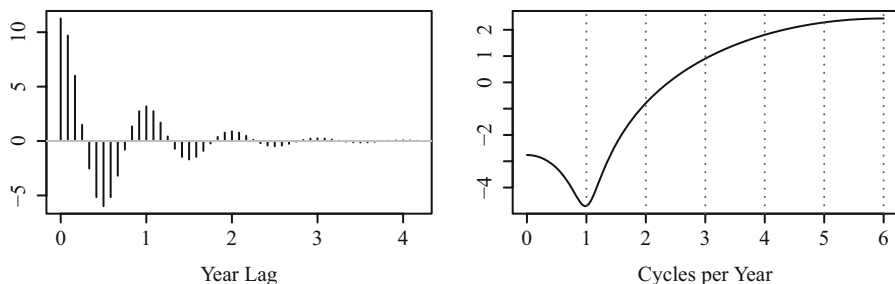


Fig. 3. Inverse autocovariance function (left panel) and log spectral density (right panel) for an anti-seasonal MA(2) process ($\rho = .9$). Inverse autocovariance function is plotted as a function of lag divided by 12; log spectral density is plotted as a function of cycles per year, or 12 divided by period.

scenarios, we suppose that the MA order $p = 2$ is known in the calculation of the test statistic, and we use critical values from the true MA(2) process. Second, we obtain critical values for the specified MA(2) with autocovariances fitted to each simulation. Third, we use an over-specified order ($q = 12$, which equals the number of seasons) to compute the test statistic and critical values. Finally, we determine the order q empirically by the procedure mentioned in Section 3. In each case the critical values are generated for $\alpha = .10, .05, .01$; for the second two cases, we report the proportion of p-values that are less than α .

We examine the size of the procedure by taking as null hypothesis that $\rho = .9$, and then investigate power by changing the process to having less anti-seasonality ($\rho = .5$). The size results are given in Table 9, and the power results are in Table 10. The test is slightly conservative, but for the first two cases (where the MA order is known) the size is adequate if there are 15 years of data. There is some deterioration in size with the third case (where q is over-specified), but results are not much worse when q is determined empirically. As for power, the gap between true and hypothesized ρ is fairly large, and the

Table 9. Size simulations from a MA(2) DGP based on a null hypothesis of .9-persistent anti-seasonality at frequency $\pi/6$. Results are for known MA order (first three rows), for unknown parameters (second three rows), for over-specified MA order (third three rows), and for empirically-determined MA order (final three rows).

α	5 years	10 years	15 years	20 years
.10	.084	.086	.096	.092
.05	.044	.046	.050	.047
.01	.013	.012	.012	.011
.10	.069	.080	.089	.094
.05	.022	.033	.036	.047
.01	.002	.003	.005	.007
.10	.047	.067	.074	.084
.05	.016	.025	.030	.037
.01	.001	.002	.003	.004
.10	.046	.056	.075	.091
.05	.018	.019	.023	.022
.01	.003	.003	.003	.002

Table 10. Power simulations from a MA(2) DGP with $\rho = .5$, based on a null hypothesis of .9- persistent anti-seasonality at frequency $\pi/6$. Results are for known MA order (first three rows), for unknown parameters (second three rows), for over-specified MA order (third three rows), and for empirically-determined MA order (final three rows).

α	5 years	10 years	15 years	20 years
.10	.373	.630	.829	.921
.05	.239	.465	.682	.821
.01	.085	.205	.369	.517
.10	.413	.671	.834	.921
.05	.267	.512	.697	.826
.01	.085	.233	.381	.563
.10	.057	.093	.117	.128
.05	.022	.041	.057	.069
.01	.000	.005	.009	.013
.10	.057	.066	.087	.111
.05	.021	.021	.022	.030
.01	.002	.001	.002	.002

results are adequate when model order is known. However, when q is over-specified or empirically determined, the variability in the estimation of V is excessive and the power is fairly low.

5. Empirical Illustrations

This section contains several empirical examples. First, we study a fairly typical monthly retail series and apply the testing procedure described in Section 3. Second, we test several raw daily time series for seasonality, illustrating the capability of the new diagnostics to handle different seasonal frequencies. Then, we consider the problem of residual seasonality in published US GDP. Finally, we compare two types of seasonal adjustments for a large collection of published monthly time series, and compare the diagnostic results to those obtained using Q_s . An additional illustration is included in Appendix C of the online Supplement.

5.1. Retail 442

We first analyze series 442 (Furniture and Home Furnishings Stores) of Retail Trade and Food Services, U.S. Census Bureau, covering the sample period of January 1992 through August 2019 (see Figure 4). We test the raw series for seasonality, and then seasonally adjust the series using the X11 method of X-13ARIMA-SEATS (U.S. Census Bureau 2015), which involves identification of outliers, holiday effects, and a SARIMA model for forecast extension of the data. Next, we test the seasonal adjustment for both under-adjustment and over-adjustment. Here we focus on producing a confidence interval for ρ , for each of the five seasonal frequencies $\pi j/6$ ($1 \leq j \leq 5$), for the cases considered.

The series is displayed with its seasonal adjustment (left panel) and seasonal factors (right panel) of Figure 4. The left panel seems to indicate a very obvious seasonal pattern, which is confirmed by the seasonal factors displayed in the right panel. To apply the test

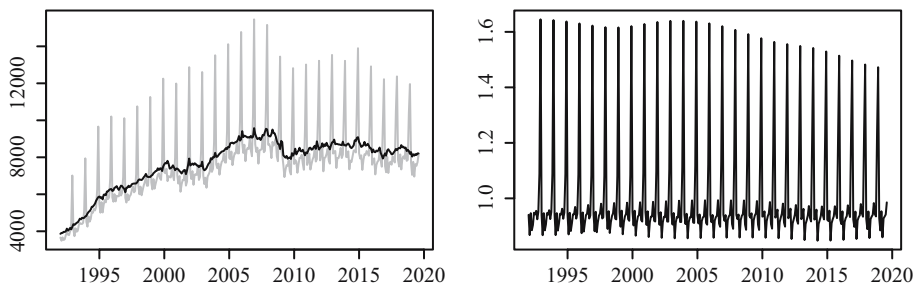


Fig. 4. Retail series 442 (furniture and home furnishings stores): raw series in grey with seasonal adjustment in black (left panel), and seasonal factors in black (right panel).

statistic in practice, we fit the $AR(\hat{p})$ model with \hat{p} selected by AIC, first differencing the data if appropriate to remove deterministic trend effects and possible non-stationary trend unit roots. For the seasonal adjustment and the seasonal factors, we also remove the first and last three years of data. For the differenced raw data AIC yields $\hat{p} = 16$, but $\hat{p} = 2$ for the trimmed (and differenced) seasonal adjustment, and $\hat{p} = 15$ for the seasonal factors. For the over-adjustment test, we set $q = 12$. The intervals for ρ are given in Table 11, where all values of ρ between .5 and 1 (using a grid mesh of size .0001) are listed such that the p-value for the corresponding null hypothesis exceeds .01.

The results indicate that the raw data has strong seasonality present at all five of the seasonal frequencies. Next, testing the seasonal adjustment for residual seasonality indicates adequacy – no seasonality is detected. The seasonal factors appear to have captured all the seasonality that was present in the raw data, because there is failure to reject the null of persistent seasonality with very high values of ρ . Finally, the test for over-adjustment indicates there are some spectral troughs of moderate scope at all the seasonal frequencies, because anti-seasonality cannot be rejected for persistencies up to .838 (for the multiple test). This is consistent with known features of seasonal adjustment filters.

5.2. NZ Immigration

We consider an analysis of daily data described in McElroy and Jach (2019). Figure 5 displays six daily immigration series of New Zealand, covering the period September 1, 1997 through July 31, 2012. The six series are labeled as NZArr, NZDep, VisArr, VisDep, PLTArr, and PLTDep. The plots show trend and seasonal behavior, and there is also a weekly effect. There is some evidence that each of these six series may not be stationary even after trend-

Table 11. Intervals for ρ , such that the corresponding null hypothesis is not rejected at a 1% level. Rows correspond to raw data, seasonally adjusted data tested for under-adjustment, seasonal factors, and seasonally adjusted data tested for over-adjustment.

Component	$\pi/6$	$2\pi/6$	$3\pi/6$	$4\pi/6$	$5\pi/6$	All
Raw	[.994, 1)	[.994, 1)	[.996, 1)	[.994, 1)	[.989, 1)	[.994, 1)
SA (Under)	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
SF	[.997, 1)	[.997, 1)	[.997, 1)	{.999}	[.992, 1)	[.998, 1)
SA (Over)	(0,.907)	(0,.870)	(0,.913)	(0,.899)	(0,.860)	(0,.838)

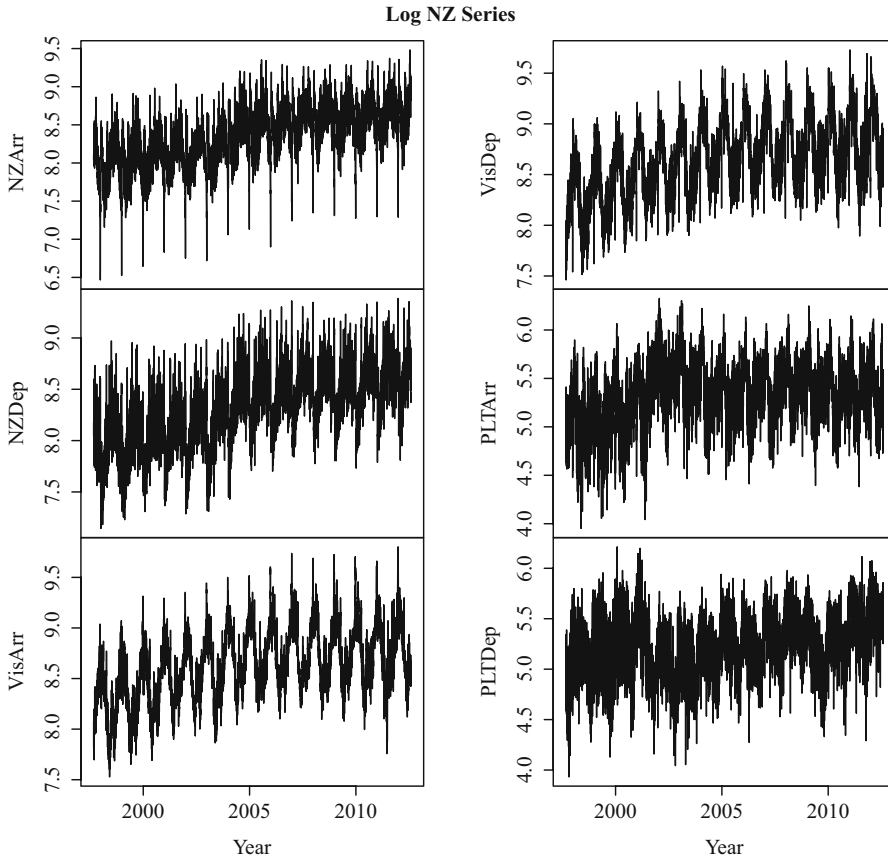


Fig. 5. Log of six New Zealand immigration series (September 1, 1997 through July 31, 2012).

differencing (there are fixed effects present), which tempers our findings accordingly. We apply the seasonality diagnostics to the differenced logged data for some of the frequencies suggested in Section 3, namely $\omega = 2\pi j/7$ for $1 \leq j \leq 3$ and $\omega = 2\pi/365.25$. An $AR(\hat{p})$ model is fitted with \hat{p} selected by AIC, which is recorded in Table 12 along with the identified intervals for ρ (all values between .5 and 1 such that the p-value exceeds .01). We find that high order AR processes are needed (due to the autocorrelation present at an annual period), and that strong seasonality is present at both annual and weekly frequencies for all six series (with

Table 12. Intervals for ρ , such that the corresponding null hypothesis is not rejected at a 1% level. Rows correspond to each of the six component series.

Series	\hat{p}	$2\pi/365.25$	$2\pi/7$	$4\pi/7$	$6\pi/7$
NZArr	447	{.999}	{.999}	[.998, 1)	[.998, 1)
NZDep	405	{.999}	{.999}	[.998, 1)	[.998, 1)
VisArr	391	{.999}	{.999}	[.997, 1)	[.999, 1)
VisDep	404	{.999}	\emptyset	[.999, 1)	[.999, 1)
PLTArr	391	[.999, 1)	{.999}	[.999, 1)	[.999, 1)
PLTDep	398	[.999, 1)	[.999, 1)	[.999, 1)	[.999, 1)

the exception of the first weekly seasonal for the VisDep series) – this confirms the exploratory analysis of these series given in [McElroy and Jach \(2019\)](#).

5.3. Gross Domestic Product (GDP)

There has been an ongoing public debate regarding the presence of residual seasonality in GDP, which is published by the Bureau of Economic Analysis (BEA). In recent years GDP (and some of its major components) has been observed to grow at a lower rate in the first quarter (see [Furman 2015](#); [Gilbert et al. 2015](#); [Stark 2015](#); [Rudebusch et al. 2015](#); [Groen and Russo 2015](#)). These critiques have prompted research into seasonality diagnostics and seasonal adjustment at BEA (see the discussion in [Lengerman et al. 2017](#)). [McCulla and Smith \(2015\)](#) review BEA’s response, and [Phillips and Wang \(2016\)](#), [Lunsford \(2017\)](#), and [Wright \(2018\)](#) delineate continuing difficulties.

We plot the logged quarterly data (left panel of [Figure 6](#)) together with growth rates (differences of logs) plotted by quarter (right panel of [Figure 6](#)), where the vertical dotted lines demarcate the different annual series for each of the four quarters. In more recent years, it appears that the first quarter is lower than the other quarters, and the question is whether this is significant. We applied the seasonality diagnostic to the differences of the logged data at frequency $\pi/2$, having obtained $\hat{p} = 16$ from AIC, and found that seasonality was rejected for all values of $\rho \in [.5, 1)$. Repeating the procedure, but focused on just the last 20 years of data, we found (with $\hat{p} = 2$) the same results. This is confirmed by examination of sample acvf and spectral plots – though when p is allowed to increase (corresponding to a likely over-specification) to 8, 12, or 16, a modest peak appears in the spectrum, somewhat off-shifted from frequency $\pi/2$. From this preliminary analysis, we do not find evidence of residual seasonality in GDP.

5.4. Census Data

We examined a collection of 233 monthly time series published by the U.S. Census Bureau, available from www.census.gov/retail/index.html. In particular, we study 65 time series of Retail Trade and Food Services (MRTS), 22 time series of Wholesale Trade: Sales and Inventories (MWTIS), 4 time series of Manufacturers’ Shipments, Inventories, and Orders (M3), 87 time series of Manufacturing and Trade Inventories and Sales (MTIS), and 55 time series of New Residential Construction (RES). All are monthly with a

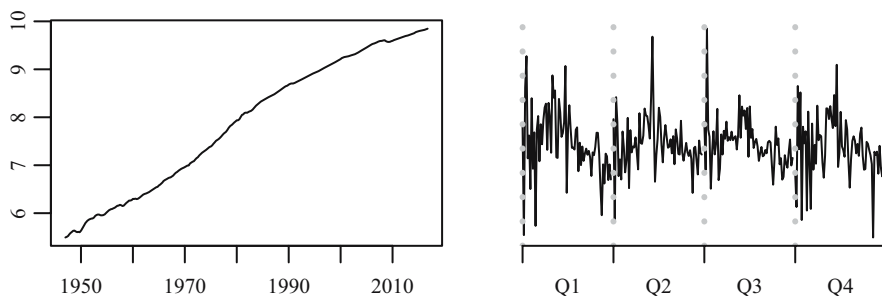


Fig. 6. Log GDP (1947 through 2016) as a time series plot (left panel) and plotted by quarter for growth rates (right panel).

start date of January 1992 or later, and with an end date of September 2019. A variety of features are present in these series: varying degrees of persistence and evolution in seasonal patterns; presence of outliers and calendrical effects; varying degrees of aggregation. In order to assess the performance of the new diagnostics, we extract seasonal adjustments and irregular components by application of the software X-13ARIMA-SEATS (U.S. Census Bureau 2015), using either the X11 option (which uses ad hoc filters in a nonlinear extraction scheme, together with ARIMA model-based forecast extension) or the SEATS option (which uses a fully model-based filtering paradigm).

We use the automatic modeling option of the software, so that calendrical effects and extremes (except for additive outliers) are removed along with the seasonal component. In many cases a log transformation is identified, and the order of differencing needed for the seasonally adjusted component is extracted from the software run; after transformation and differencing (this is not applied to the irregular), we apply the Q_s statistic as well as our proposed diagnostic for under- adjustment and over-adjustment. For the latter two diagnostics, we set the values of ρ between .98 and 1 in order to focus on the more egregious cases of residual seasonality. Setting the threshold $\alpha = .05$, any Q_s p-values above α indicate adequacy of the seasonal adjustment, whereas for the under-adjustment test we require the p-value to be less than α for adequacy. For the over- adjustment test (for which Q_s cannot be used without substantial modifications) p-values less than α indicate that over-adjustment has not occurred. We tally for each of the five batches of series the incidences of adequacy, presented in Tables 13 and 14.

The results based on the seasonal adjustments (SA) and irregular (Irr) are fairly similar, except when examining the case of over-adjustment, where there are some discrepancies (in particular, for MRTS). For both the X11 and SEATS methods, results for Q_s and the proposed test of under-adjustment are broadly similar, although Q_s is more prone to identifying residual seasonality. This is intuitive, because any degree of positive lag 12 autocorrelation will place the process in the alternative space of the Q_s test, whereas only high degrees of such positive autocorrelation will trigger the test for under-adjustment. For the test of over-adjustment most of the SA and Irr are deemed inadequate due to the large spectral troughs induced by filtering – both the X11 and SEATS methods produce this effect. Given the discussion of Bell and Hillmer (1984), these findings are not surprising.

Table 13. Number of series seasonally adjusted by X11 that are deemed to be adequate, according to whether the Q_s test, the test of under-adjustment, or the test for over-adjustment is applied, for either the seasonally adjusted (SA) component or the Irregular (Irr).

X11 Test	MRTS	MWTS	M3	MTIS	RES
SA Q_s	63	22	86	4	53
SA Under	64	22	87	4	55
SA Over	3	2	4	1	1
Irr Q_s	63	22	86	4	55
Irr Under	64	22	87	4	55
Irr Over	4	0	4	0	1
Total	65	22	87	4	55

Table 14. Number of series seasonally adjusted by SEATS that are deemed to be adequate, according to whether the Q_s test, the Root test of under-adjustment, or the Root test for over-adjustment is applied, for either the seasonally adjusted (SA) component or the irregular (Irr).

SEATS Test	MRTS	MWTS	MTIS	M3	RES
SA Q_s	64	22	85	4	52
SA Under	65	22	87	4	55
SA Over	3	1	2	0	2
Irr Q_s	65	22	86	4	55
Irr Under	65	22	87	4	55
Irr Over	6	1	2	0	2
Total	65	22	87	4	55

6. Summary

We have shown an approach to seasonality detection that aims to achieve the five criteria outlined in the Introduction. This approach depends upon a notion of seasonal persistency that relies upon the Wold Decomposition, and seasonality is measured by evaluating the corresponding AR representation at $z = \rho^{-1}e^{i\omega}$ for various $\rho \in (0, 1)$. By examining different values of ω , various periodic effects can be simultaneously investigated. For instance, in daily time series we can examine weekly effects (ω given by $2\pi/7$, $4\pi/7$, or $6\pi/7$) together with annual effects ($\omega = 2\pi/365.25$, among other integer multiples). When a series is down-sampled, or flow-aggregated, one can easily adapt the phase criteria for seasonality, although we do not mathematically derive here how the AR roots are changed by such alterations of sampling frequency.

The AR diagnostic tests are principally useful for detecting under-adjustment in a seasonal adjustment, as well as for detecting whether seasonality exists in a raw series. The former exercise is designated as testing for adequacy of seasonally adjusted data, whereas the latter is referred to as pre-testing, that is, determining whether a given series is a candidate for seasonal adjustment. (Because seasonal adjustment procedures are not idempotent in general, there is a cost associated with seasonally adjusting data that does not warrant such a procedure.) In order to detect over-adjustment, we propose the use of the MA diagnostic tests. Theory and simulation supports the use of these methods on stationary or non-stationary time series data.

The article shows how the diagnostics can be applied as a joint test over multiple frequencies, thereby allowing for a single test of seasonality (or anti-seasonality). Directionality of rejections can be determined through computing p-values as a function of the persistency ρ ; by this means, we can determine whether a rejection of seasonality at a given level ρ favors the presence of more or less persistent seasonality. We stress that testing for raw seasonality avoids the difficult distribution theory of unit root tests by the device of rejecting ever more persistent formulations of stationary seasonality – this is a pragmatic approach to a thorny statistical problem.

A limitation of this article's approach is that the time series must either be stationary, or can be reduced to stationarity by a differencing polynomial. This precludes testing for fixed seasonality (given by seasonal dummies in a regression), and precludes detecting

more exotic forms of seasonality (such as seasonal heteroscedasticity); mainly, the method is useful for detecting dynamic ($\rho < 1$) or unit root ($\rho = 1$) types of seasonality. A mitigating factor is that in practice an AR or ARMA model will be fitted to compute the diagnostic, and the size and power are sensitive to model selection.

7. References

- Ansley, C., and W. Wecker. 1984. "Issues involved with the seasonal adjustment of economic time series, comment: on dips in the spectrum of a seasonally adjusted time series." *Journal of Business and Economic Statistics* 2(4): 323–324. DOI: <https://doi.org/10.1080/07350015.1984.10509400>.
- Bell, W.R. 1984. "Signal extraction for nonstationary time series." *Annals of Statistics* 12(2): 646–664. DOI: <https://doi.org/10.1214/aos/1176346512>.
- Bell, W.R., and S. Hillmer. 1984. "Issues involved with the seasonal adjustment of economic time series." *Journal of Business and Economic Statistics* 2(4): 291–320. DOI: <https://doi.org/10.1080/07350015.1984.10509398>.
- Blakely, C., and T.S. McElroy. 2017. "Signal extraction goodness-of-fit diagnostic tests under model parameter uncertainty: formulations and empirical evaluation." *Econometric Reviews* 36(4): 447–467. DOI: <https://doi.org/10.1080/07474938.2016.1140277>.
- Busetti, F., and A. Harvey. 2003. "Seasonality tests." *Journal of Business and Economic Statistics* 21(3): 420–436. DOI: <https://doi.org/10.1198/073500103288619061>.
- Canova, F., and B.E. Hansen. 1995. "Are seasonal patterns constant over time? A test for seasonal stability." *Journal of Business and Economic Statistics* 13(3): 237–252. DOI: <https://doi.org/10.1080/07350015.1995.10524598>.
- Chiu, S.T. 1988. "Weighted least squares estimators on the frequency domain for the parameters of a time series." *The Annals of Statistics* 16(3): 1315–1326. DOI: <https://doi.org/10.1214/aos/1176350963>.
- Den Butter, F., and M. Fase. 1991. *Seasonal Adjustment as a Practical Problem*, (Vol. 1991. Amsterdam: North Holland.
- Deo, R.S., and W.W. Chen. 2000. "On the integral of the squared periodogram." *Stochastic processes and their applications* 85(1): 159–176. DOI: [https://doi.org/doi.org/10.1016/S0304-4149\(99\)00071-X](https://doi.org/doi.org/10.1016/S0304-4149(99)00071-X).
- Fase, M., J. Koning, and A. Volgenant. 1973. "An experimental look at seasonal adjustment: a comparative analysis of nine adjustment methods." *De Economist* 121: 177–180. DOI: <https://doi.org/doi.org/10.1007/BF01712804>.
- Findley, D.F., D.P. Lytras, and T.S. McElroy. 2017. *Detecting seasonality in seasonally adjusted monthly time series*. Census Bureau research report. Available at: www.census.gov/ts/papers/frs2017-03.pdf (accessed May 2021).
- Findley, D.F., B.C. Monsell, W.R. Bell, M.C. Otto, and B.C. Chen. 1998. "New capabilities and methods of the X-12-ARIMA seasonal adjustment program." *Journal of Business and Economic Statistics* 16: 127–177. DOI: <https://doi.org/doi.org/10.1080/07350015.1998.10524743>.

- Franses, P.H. 1994. "A multivariate approach to modeling univariate seasonal time series." *Journal of Econometrics*: 133–151. DOI: [https://doi.org/doi.org/10.1016/0304-4076\(93\)01563-2](https://doi.org/doi.org/10.1016/0304-4076(93)01563-2).
- Furman, J. 2015. "Second estimate of GDP for the first quarter of 2015." *Council of Economic Advisers Blog*, May 29, 2015. Available at: <https://obamawhitehouse.archives.gov/blog/2015/05/29/second-estimate-gdp-first-quarter-2015> (accessed May 2021).
- Gilbert, C.E., N.J. Morin, A.D. Paciorek, and C.R. Sahm. 2015. "Residual seasonality in GDP." *FEDS Notes*. Washington: Board of Governors of the Federal Reserve System, May 14, 2015. DOI: <https://doi.org/10.17016/2380-7172.1538>.
- Granger, C. 1978. "Seasonality: causation, interpretation, and implications." In *Seasonal Analysis of Economic Time Series*, edited by A. Zellner: 33–46. Cambridge, MA: NBER.
- Groen, J., and Russo. 2015. "The myth of first-quarter residual seasonality." *Liberty Street Economics*. June 8, 2015. Available at: <https://libertystreeteconomics.newyorkfed.org/2015/06/the-myth-of-first-quarter-residual-seasonality.html> (accessed May 2021).
- Hylleberg, S. 1986. *Seasonality in Regression*, Orlando, Florida: Academic Press.
- Hylleberg, S., R.F. Engle, C.W. Granger, and B.S. Yoo. 1990. "Seasonal integration and cointegration." *Journal of econometrics* 44(1–2): 215–238. DOI: [https://doi.org/10.1016/0304-4076\(90\)90080-D](https://doi.org/10.1016/0304-4076(90)90080-D).
- Lengermann, P., N. Morin, A. Paciorek, E. Pinto, and C. Sahm. 2017. "Another Look at Residual Seasonality in GDP." *FEDS Notes*. Washington: Board of Governors of the Federal Reserve System, July 28, 2017. DOI: <https://doi.org/10.17016/2380-7172.2031>.
- Lin, W., J. Huang, and T. McElroy. 2019. "Time series seasonal adjustment methods using regularized singular value decomposition." Published online. *Journal of Business and Economics Statistics*: 1–23. DOI: <https://doi.org/10.1080/07350015.2018.1515081>.
- Lunsford, K.G. 2017. *Lingering residual seasonality in GDP growth*. Economic Commentary, Federal Reserve Bank of Cleveland, March 28, 2017.
- Lytras, D.P., R.M. Feldpausch, and W.R. Bell. 2007. "Determining seasonality: a comparison of diagnostics from X-12-ARIMA." In *Proceedings of the Third International Conference on Establishment Surveys (ICES-III June 18–21, 2007, Montreal, Canada*. Available at: www.census.gov/ts/papers/ices2007dpl.pdf (accessed May 2021).
- Maravall, A. 2003. "A class of diagnostics in the ARIMA-model-based decomposition of a time series." In *Seasonal Adjustment*, edited by M. Manna and R. Peronaci: pp. 23–36. Frankfurt am Main: European Central Bank.
- McCulla, S.H., and S. Smith. 2015. "Preview of the 2015 annual revision of the national income and product accounts." *Survey of Current Business* 95(6): 1–8. Available at: https://apps.bea.gov/scb/pdf/2015/06%20June/0615_preview_of_2015_annual_revision_of_national_income_and_product_accounts.pdf (accessed May 2021).
- McElroy, T.S. 2008. "Statistical properties of model-based signal extraction diagnostic tests." *Communications in Statistics, Theory and Methods* 37: 591–616. DOI: <https://doi.org/10.1080/03610920701669785>.
- McElroy, T.S. 2012. "An alternative model-based seasonal adjustment that reduces over-adjustment." *Taiwan Economic Forecast and Policy* 43: 33–70. Available at: <http://>

- www.econ.sinica.edu.tw/UpFiles/2013092817175327692/Periodicals_Pdf2013093010104847832/EC431-02.pdf (accessed May 2021).
- McElroy, T., and Jach, A. 2019. "Testing collinearity of vector time series." *The Econometrics Journal* 22(2): 97–116. DOI: <https://doi.org/10.1093/ectj/uty002>.
- McElroy, T.S., B.C. Monsell, and R.J. Hutchinson. 2018. *Modeling of holiday effects and seasonality in daily time series*. Census Bureau research report 2018-01. Available at: www.census.gov/srd/papers/pdf/RRS2018-01.pdf (accessed May 2021).
- McElroy, T., and D. Politis. 2020. *Time Series: a First Course with Bootstrap Starter*. New York: Chapman and Hall.
- McElroy, T.S. and A. Roy. 2017. *Detection of seasonality in the frequency domain*. Census Bureau research report, 2017–01. Available at: www.census.gov/ts/papers/rrs2017-01.pdf (accessed May 2021).
- McElroy, T., and A. Roy. 2018. "The inverse Kullback Leibler method for fitting vector moving averages." *Journal of Time Series Analysis* 39: 172–191. DOI: <https://doi.org/10.1111/jtsa.12276>.
- Moulton, B.R., and B.D. Cowan. 2016. "Residual seasonality in GDP and GDI: findings and next steps." *Survey of Current Business* 96(7): 1–6. Available at: https://apps.bea.gov/scb/pdf/2016/07%20July/0716_residual_seasonality_in_gdp_and_gdi.pdf (accessed May 2021).
- Nerlove, M. 1964. "Spectral analysis of seasonal adjustment procedures." *Econometrica* 32: 241–286. DOI: <https://doi.org/10.2307/1913037>.
- Phillips, K., and J. Wang. 2016. "Residual seasonality in U.S. GDP data." FRB of Dallas Working Paper No. 1608. DOI: <https://doi.org/10.24149/wp1608>.
- Proietti, T. 1996. "Persistence of shocks on seasonal processes." *Journal of Applied Econometrics* 11(4): 383–398. DOI: [https://doi.org/10.1002/\(SICI\)1099-1255\(199607\)11:4<383:AID-JAE403>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1099-1255(199607)11:4<383:AID-JAE403>3.0.CO;2-3).
- Rudebusch, G.D., D. Wilson, and T. Mahedy. 2015. "The puzzle of weak first-quarter GDP growth." *FRBSF Economic Letter*, May 18, 2015. Available at: <https://www.frbsf.org/economic-research/publications/economic-letter/2015/may/weak-first-quarter-gdp-residual-seasonality-adjustment/> (accessed May 2021).
- Sims, C. 1978. "Comments on 'Seasonality: causation, interpretation, and implications' by Clive W.J. Granger." In *Seasonal Analysis of Economic Time Series*, edited by A. Zellner: 47–49. Cambridge, MA: NBER.
- Soukup, R., and D. Findley. 1999. "On the spectrum diagnostics used by X-12-ARIMA to indicate the presence of trading day effects after modeling or adjustment." In 1999 Proceedings American Statistical Association: Alexandria. August 8–12, 1999, Baltimore, Maryland, USA. Available at: www.census.gov/ts/papers/tr9903s.pdf (accessed May 2021).
- Stark, T. 2015. "First quarter in the national income and product accounts." *Research Rap*, Federal Reserve Bank of Philadelphia, May 14, 2015. Available at: https://www.philadelphiafed.org/-/media/frbp/assets/economy/reports/research-rap/2015/first_quarters_national_income_product_accounts.pdf (accessed May 2021).
- Tanaka, K. 1996. *Time series analysis: nonstationary and noninvertible distribution theory*. New York: John Wiley and Sons.

- Tukey, J. 1978. "Comments on 'Seasonality: causation, interpretation, and implications' by Clive W.J. Granger." In *Seasonal Analysis of Economic Time Series*, edited by A. Zellner: 50–54. Cambridge, MA: NBER.
- U.S Census Bureau. 2015. *X-13ARIMA-SEATS reference manual*. U.S. Census Bureau, Washington D.C., USA. Available at: www.census.gov/ts/x13as/docX13AS.pdf (accessed date May 2021).
- Wright, J.H. 2018. "Seasonal adjustment of NIPA data." *NBER working paper*, 24895, August 2018. DOI: <https://doi.org/10.3386/w24895>.

Received May 2019

Revised June 2020

Accepted September 2020

Applying Machine Learning for Automatic Product Categorization

*Andrea Roberson*¹

Every five years, the U.S. Census Bureau conducts the Economic Census, the official count of US businesses and the most extensive collection of data related to business activity. Businesses, policymakers, governments and communities use Economic Census data for economic development, business decisions, and strategic planning. The Economic Census provides key inputs for economic measures such as the Gross Domestic Product and the Producer Price Index. The Economic Census requires businesses to fill out a lengthy questionnaire, including an extended section about the goods and services provided by the business.

To address the challenges of high respondent burden and low survey response rates, we devised a strategy to automatically classify goods and services based on product information provided by the business. We asked several businesses to provide a spreadsheet containing Universal Product Codes and associated text descriptions for the products they sell. We then used natural language processing to classify the products according to the North American Product Classification System. This novel strategy classified text with very high accuracy rates - our best algorithms surpassed over 90%.

Key words: Text analytics; artificial intelligence; data collection.

1. Introduction

The North American Product Classification System (NAPCS) is a comprehensive, hierarchical classification system for products (goods and services) that is consistent across the United States, Canada, and Mexico, and promotes improvements in the identification and classification of products across international classification systems, such as the Central Product Classification System of the United Nations.

Beginning with the 2017 Economic Census, NAPCS will be used to produce economy-wide product tabulations. Respondents were asked to report data from a long, pre-specified list of potential products in a given industry, with some lists containing more than 50 potential products. The definitions of these NAPCS codes can be very complex and ambiguous. Businesses have expressed the desire to alternatively supply Universal Product Codes (UPC) to the U. S. Census Bureau, as they already maintain UPCs in their databases ([Thompson and Ellis 2015](#)).

Businesses are generally readily able to report attributes for the business such as total sales, total payroll, and total number of employees. It is much more burdensome for businesses to provide, in a traditional survey instrument, detailed information about their products. Much work has been done around the categorization of products using product

¹ U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C., 20233, U.S.A. Email: andrea.roberson@census.gov

descriptions (Chen and Warren 2013), but no known study has applied these techniques for the calculation of official statistics (statistics published by government agencies) using only the text of UPC product descriptions. The question we address in this article is: *Can we use UPC codes and their associated descriptions to accurately classify products into NAPCS?*

The main contributions of the article are:

- We present a novel method for survey data collection to automate the U.S. Economic Census with supervised machine learning.
- We provide a product categorization strategy for the three North American countries' classification systems.
- This strategy leverages new approaches and technologies to improve collection speeds, reduce costs, and alleviate respondent burden.

2. Classification Methods

The classification of data requires a class membership decision y' of an unidentified data item x' given some data set $D = (x_1, y_1), \dots, (x_n, y_n)$ of data elements x_i that belong to the class y_i . Here, x_i is a UPC description and $y_i \in \{1, \dots, P\}$ is its associated NAPCS code. We use three classification algorithms to categorize products: Support Vector Machine (SVM) with linear kernel, Logistic Regression (LR), and Multinomial Naïve Bayes (MNB).

2.1. Support Vector Machines

Support Vector Machines (SVMs) were developed by Vapnik (2000) based on the structural risk minimization principle from statistical learning theory. Statistical learning theory, the backbone of SVMs, provides a new framework for modeling learning algorithms, merges the fields of ML and statistics, and inspires algorithms that overcome many theoretical and computational difficulties. In recent years, SVMs have found a wide range of real-world applications, including face detection from images (Osuna et al. 1997), object recognition (Blanz et al. 1996), speaker identification (Schmidt and Gish 1996; Moreno and Ho 2003), biomedical data classification (Shoker et al. 2005), and text categorization. The many applications of SVMs for text categorization generated considerable research interest in our study.

Joachims (2001) explains how SVMs can achieve good classification performance despite the high-dimensional feature spaces in text classification. The complexity of text-classification tasks are analyzed and sufficient conditions for good generalization performance are identified. The article also provides a formal basis for developing new algorithms that are most appropriate in specific scenarios. The disadvantage of SVMs is that the classification result is purely dichotomous, and no probability of class membership is given (Masood and Al-Jumaily 2013). Another disadvantage of SVMs is the “black box” nature of these classifiers, the decisions made by the model are not easily explainable. The model produced does not naturally provide any useful intuitive reasons about why a particular point is classified in one class rather than another. This leads us to explore models that will be easily understood by our customers. We consider Logistic Regression (LR), where the most important features of the model gives us insights into its inner workings and provides direction for improving performance.

2.2. Logistic Regression

Regression modeling is one of several statistical techniques that enable an analyst to predict a response based upon a set of inputs. Linear regression models are commonly used when the range of the response is continuous, and can theoretically take any value. LR, invented in the 19th century for the description of the growth of population and the course of chemical reactions, predicts the probability of an occurrence of an event by fitting data to a logistic curve (Zhang et al. 2011). As the output is restricted to the interval (0, 1), the assumption of an infinite range fails. The logistic function used in this prediction method is useful in that it can take any value from negative infinity to positive infinity as input.

2.3. Naïve Bayes

The Naïve Bayes classifier is a classifier based on Bayes Theorem with the naive assumption that features are independent of each other. The Bernoulli Naïve Bayes model uses a set of binary occurrence features. When classifying a text document for example, the Bernoulli Naïve Bayes model is convenient because we could represent the presence or the absence of the given word in the text with a binary feature. On the other hand, this model does not take into account how often the word occurs in the text. The Multinomial Naïve Bayes model (MNB) uses a set of count-based features, each of which does account for how many times a particular feature, such as a word is observed in a document. MNB and SVM are popular choices (Ikonomakis et al. 2005; Joachims 1998; Sebastiani 2002). Both can efficiently deal with high dimensionality and data sparsity.

2.4. Comparison between Logistic Regression, Naïve Bayes, and Support Vector Machines

Both Naïve Bayes (Eyheramendy et al. 2003) and Logistic Regression (Zhang et al. 2003) are examples of probabilistic algorithms. Here the dependent variable is a category (Cosmetics or Personal Hygiene). We have a set of text as predictors or features, which come from our UPC product descriptions. This is called training data in ML terminology. MNB takes advantage of probability theory and Bayes' Theorem to predict the NAPCS class. The algorithm is probabilistic, meaning we calculate the probability of each class for a given text, and then output the most likely class. These probabilities are determined by using Bayes' Theorem, which describes the probability of a feature, based on prior knowledge of conditions that might be related to that feature.

The parameters of an LR model can be estimated by the probabilistic framework called maximum likelihood estimation. Under this framework, a probability distribution for the response variable (NAPCS code) must be assumed, and then a likelihood function defined that calculates the probability of observing the outcome given the input data and the model. This likelihood function can then be optimized to find the set of parameters that gives the greatest possible probability to the training data.

Support Vector Machines, however, are non-probabilistic classifiers. It has the same goal as MNB and LR. Given training data, find the best SVM model, and use the model to classify new UPC descriptions. The difference is that the optimization problem is finding the hyperplane that best separates UPC text labeled "Cosmetics" from those labeled

“Personal Hygiene”. The ML models take the UPC description “*Diamond of California Shelled Pistachios*” and predict the associated NAPCS code for “*snack foods*” as pictured in [Figure 1](#).

3. Relevant Literature

Word (or n-gram) frequencies are typical units of analysis when working with text collections. The general term n-gram means ‘sequence of length n’. A three-word sequence is called a trigram, a sequence of two words is called a bigram, and a single word is called a unigram. It may come as a surprise that reducing a book to a list of word frequencies retains useful information, but this has been demonstrated in natural language processing (NLP) research ([Pagliardini et al. 2018](#)). Treating texts as a list of word frequencies (a vector) also makes available a vast range of mathematical tools developed for studying and manipulating vectors.

Text feature extraction is the process of transforming what is essentially a list of words into a feature set that is usable by a classifier. In Bag-of-Words feature selection, the document is treated as an unordered list of words. Under this approach, words are ranked solely by their frequencies. In this case, the set of feature vectors can be considered as a matrix where each row is one instance and each column represents a word found in any of the documents. Thus, each cell (i,j) represents the number of times a word appears in the text of the document. It can be noted that this model builds a $n \times m$ matrix where, for our work, n is the number of UPC text descriptions and m is the number of words without repetition that appear in the n descriptions.

In our analysis, we were able to extract features by using an n -gram model to transform the data into feature vectors for use in our models. We gathered word frequencies (or term frequencies) associated with texts into a document-term matrix using the CountVectorizer class from the scikit-learn python package.

The document-term matrix is usually very high dimensional and sparse. It can create issues for ML algorithms during the learning phase. Therefore, it is recommended to

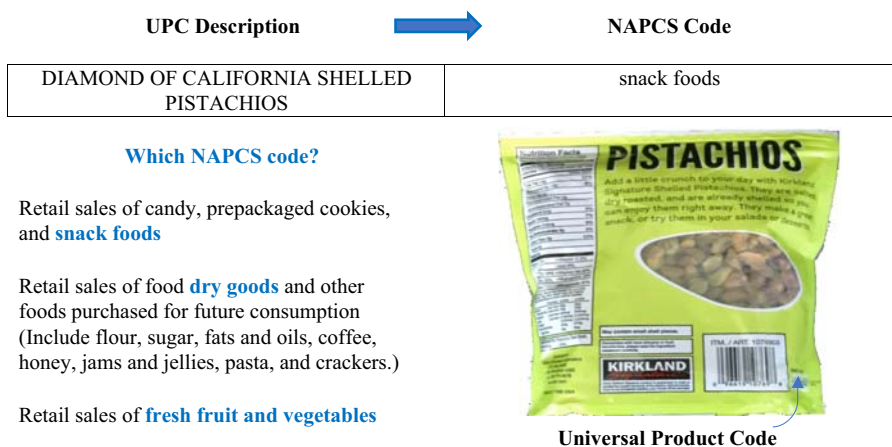


Fig. 1. Example of UPC description Categorization. Given a UPC text description of Pistachios, we seek to predict the associated snack foods NAPCS category. A business might sell pistachios and have trouble deciding what the correct class is: snack foods, dry goods, or fresh fruits and vegetables.

reduce the dimensionality of the data set by either feature selection or dimensionality reduction methods (Wang and Manning 2012). The former selects important features from the original feature set, whereas the latter learns new features from the original set in some other dimension. We will apply Chi-Square (χ^2) and Mutual information (MI) as feature selection methods, and Latent Semantic Analysis (LSA) as a dimensionality reduction technique (Roberts 1996).

3.1. Chi-Square Dictionary

The χ^2 test is used in statistics to test the independence of two events. More precisely in feature selection, it is used to test whether the occurrence of a specific term and the occurrence of a specific class are independent. Given a document D we estimate the quantity $\chi^2(D, t, c)$ for each term, and rank them by their score:

$$\chi^2(D, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \quad (1)$$

where e_t takes the value 1 if the document contains term t and 0 otherwise, and e_c takes the value 1 if the document is in class c and 0 otherwise. The χ^2 statistic is a measure of how much expected counts E and observed counts N deviate from each other. A high value of χ^2 indicates that the hypothesis of independence, which implies that the expected and observed counts are similar, is incorrect. Features with very small probabilities deviate significantly from the independence assumption and are therefore considered important. In this context, χ^2 helps identify the most relevant features in the data, where ranking these features may lead to improved classification performance (Bahassine et al. 2018).

3.2. Mutual Information Dictionary

In information theory, MI measures how much information a word contains about the class. We might not want to use all the features, but just reliable discriminators (Kozareva 2015). We formally define the MI of a word w and a class c as $I(w, c)$ where I is given by:

$$\sum_{e_w \in \{0,1\}} \sum_{e_c \in \{0,1\}} p(e_w, e_c) \log \frac{p(e_w, e_c)}{p(e_w)p(e_c)} \quad (2)$$

Essentially the MI is a way of capturing the degree of dependence between two variables. MI compares the probability of observing e_w and e_c together with the expected joint distribution if e_w and e_c were independent. MI measures the divergence of the actual joint distribution from the expected distribution under the independence assumption. The larger the divergence is, the higher the MI would be. For each feature we compute the MI, and we repeat this analysis with varying feature sizes.

3.3. Latent Semantic Analysis Dictionary

Latent Semantic Analysis (LSA) is a technique for extracting and inferring relations of expected contextual usage of words in documents. LSA takes documents that are semantically similar (talk about the same topics) but are not similar in the vector space, and re-represents

them in a reduced vector space in which they have higher similarity. LSA applies Singular Value Decomposition (SVD) to the term-document matrix (TDM). It factors the TDM into three matrices, to form a smaller, low-rank approximation to the original matrix (Bast and Majumdar 2005). The rank- k LSA model of a TDM, $A \in \mathbb{R}^{m \times n}$, is its ran- k SVD,

$$A_k = U_k \Sigma_k V_k^T \tag{3}$$

where $U_k \in \mathbb{R}^{m \times k}$, $\Sigma_k \in \mathbb{R}^{k \times k}$, $V_k \in \mathbb{R}^{n \times k}$, contain the k leading left singular vectors, singular values, and right singular vectors, respectively. In practice, the reduction in size is usually substantial; from a TDM with tens of thousands of documents and terms, to a low-rank approximation with only a few hundred basis vectors for each document (Roberts 1996).

4. Data

To evaluate our approach to test which ML classifier would perform best in product classification, we analyzed 14,000 UPC product descriptions for 44 NAPCS categories, annotated by U.S. Census analysts. The data was provided by a business in the Health and Personal Care Stores sub-sector (class code 446). These data have disproportionate class labels in the response variable. Most of the data is distributed between the cosmetics and personal hygiene NAPCS codes. After removing categories with under 15 products we are left with 16 NAPCS categories. The distribution of our UPC description data is pictured in Figure 2.

Text-based data is inherently unstructured and must be converted to a structured format for predictive modeling or other type of analysis. This was done by applying text processing techniques. Punctuations and numbers were removed from UPC text. Next, all of the letters were converted to lowercase. Another common preprocessing step is the

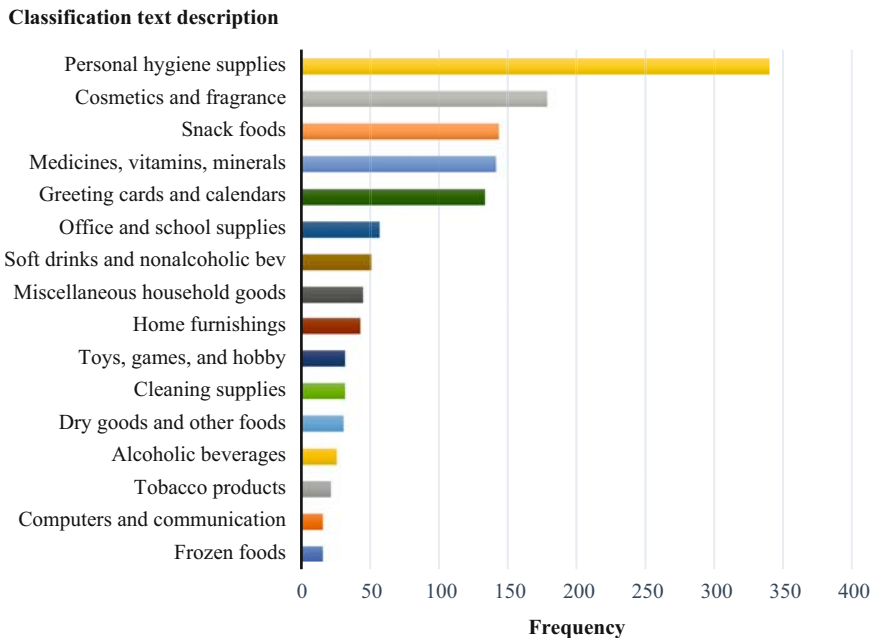


Fig. 2. Product Distribution. This is the distribution of the test data set that was used to measure how well the model generalizes to unseen data.

removal of white space. It is typically the result of all the left over spaces or tabs that were not removed along with the words that were deleted. All white space was removed.

A further preprocessing technique is the removal of stop words. They are words which are filtered out before or after processing of natural language data (text). Any group of words can be chosen as the stop words for a given purpose. Stop words are words that are so common in a language that their information value is almost zero, that is, they do not carry significant information. We would not want these words taking up space in our database, or taking up valuable processing time. Some examples are “a”, “about”, “be”, “do”. Therefore, it is recommended to remove them before further analysis.

5. Implementing Machine Learning Algorithms

The most widely used library for implementing ML algorithms in Python is scikit-learn. This library is a Python module integrating a wide range of state-of-the-art ML. This package focuses on bringing ML to non-specialists using a general-purpose high-level language. Emphasis is put on ease of use, performance, documentation, and API consistency.

A well-fitted model should not just provide good prediction accuracy on the data it was fitted to, it should also generalize to data not yet seen. We can estimate this generalization accuracy with a technique called cross-validation. The simplest form of cross-validation is as follows: the data are separated into a training set and a test set. The algorithm is fit on the training set and the accuracy (e.g., the percent is correctly classified) is evaluated on the test set, giving an estimate of how the fit generalizes.

All classifiers have various parameters which can be tuned to obtain optimal performance. Tuning is performed for varying values of the tuning parameters, searching for those that give the best generalization accuracy (Guenther and Schonlau 2016). This can be done by choosing a small number of possible values to test for each parameter, and trying all possibilities on the grid of their combinations. This is known as a grid search. In the context of ML, hyperparameters are parameters whose values are set prior to the commencement of the learning process. In scikit-learn, hyperparameter tuning can be conveniently done with the GridSearchCV estimator. It takes as input an estimator (such as accuracy) and a set of candidate hyperparameters. Cross-validation scores are then computed for all hyperparameter combinations, in order to find the best one. In this research, we tune the LR, MNB, and SVMs with GridSearchCV.

For LR, we use the `sklearn.linear_model.LogisticRegression` package in the scikit-learn library.

The parameters are as follows:

- **penalty:** It specifies the norm used in penalization. It can be ‘l1’, or ‘l2’. The default value is ‘l2’.
- **C:** It is the inverse of the regularization strength. Smaller values specify stronger regularization.

We first observe that setting the parameter C is crucial as performance drops for inappropriate values of C. The LR regularization parameter was set in the range of ($C = 10^{-4}, 10^{-3}, \dots, 10^5, 10^6$). A large C can lead to an overfit model, while a small C can lead to an underfit model. We used GridSearchCV with 10-fold cross-validation to tune C in this hyperparameter space.

The package used for SVM classification in the scikit-learn library is `svm.SVC`.

The parameters are as follows:

- `C`: It is the regularization parameter, C , of the error term.
- `kernel`: It specifies the kernel type to be used in the algorithm. It can be ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’, or callable. The default value is ‘rbf’.
- `degree`: It is the degree of the polynomial kernel function (‘poly’) and is ignored by all other kernels. The default value is 3.
- `gamma`: It is the kernel coefficient for ‘rbf’, ‘poly’, and ‘sigmoid’. If *gamma* is ‘auto’, then $\frac{1}{n}$ features will be used instead.

Training SVMs with a linear kernel is faster than with any other kernel. When you train an SVM with a linear kernel, you only need to optimize the C regularization parameter. When training with other kernels, you also need to optimize the *gamma* parameter, which means that performing a grid search will usually take more time. Therefore, linear kernels are indeed very well suited for text-categorization. It should be kept in mind, however, that this is not the only solution. In some cases, using another kernel might be better. The recommended approach for text classification is to try a linear kernel first, because of its advantages. An SVM with a linear kernel is similar to logistic regression. Therefore, in practice, the benefit of SVMs typically comes from using non-linear kernels to model non-linear decision boundaries. In this study, in an effort to get the best possible classification performance, it was of interest to try the other kernels to see if accuracy was improved.

We did a set of experiments with different kernel functions such as the linear, RBF, polynomial, and sigmoid in order to see the quality of generalization for each kernel function. Using sklearn’s SVM implementation `svm.SVC`, we apply a grid-search to find the best pair (C , *gamma*) for each kernel function using 10-fold cross-validation. In order to increase efficiency, we try exponentially growing sequences of (C , *gamma*) to identify good parameters ($C = 2^{-5}, 2^{-3}, \dots, 2^{15}$; $\textit{gamma} = 2^{-15}, 2^{-12}, \dots, 2^{12}$). After the optimal (C , *gamma*) is found, the training data is trained using the SVMs with different kernels and the best parameters to generate the final models. After testing our SVM algorithm with various kernel transformations, we identified the linear kernel as the most efficient kernel that resulted in the highest classification results.

For MNB, we use the `sklearn.naive_bayes.MultinomialNB` package in this scikit-learn library.

The parameters are as follows:

- `alpha`: It specifies the smoothing parameter. The default value is ‘0’.

The alpha parameter controls the level of smoothing applied in the training set. This can be useful when items in the test set would have zero probability based on the training set. There is no good rule of thumb for setting this parameter, so the experiment included several values within a parameter grid search.

6. Experiments and Results Data

Feature selection is simply selecting and excluding given features without changing them. Dimensionality reduction transforms features into a lower dimension. [Table 1](#) shows a

Table 1. Summary of feature selection and classification methods.

Feature reduction methods	Abbreviation	Classifiers	Abbreviation
Chi-square feature selection	χ^2	Support vector machine	SVM
	χ^2	Logistic regression	LR
	χ^2	Multinomial naïve bayes	MNB
Mutual information feature selection	MI	Support vector machine	SVM
	MI	Logistic regression	LR
	MI	Multinomial naïve bayes	MNB
Latent semantic analysis dimensionality reduction	LSA	Support vector machine	SVM
	LSA	Logistic regression	LR

summary of the feature selection and dimensionality reduction techniques that are considered for each classifier.

6.1. Results

We define classification accuracy as the percentage of UPC codes for which the classification agreed with the known categories. A UPC code description, whose fitted state differs from the ground truth label, is defined as an error. Classification accuracy, precision, recall, and F-scores were used as performance metrics. Precision captures the fraction of UPC codes classified as cosmetics that are truly cosmetics. Recall captures the fraction of codes that are truly cosmetic, that are found by the model. The F-score is the weighted harmonic mean of precision and recall.

We also calculated Cohen's Kappa coefficient to measure agreement, beyond chance, between the fitted results and the ground truth data (Grandini et al. 2020). Kappa values range between -1 (all text incorrectly classified) and 1 (all text correctly classified). A Kappa value equal to zero indicates a performance no better than random. This was especially important since our data set has severe class imbalance, a classifier could obtain high accuracy by always guessing the most frequent class. Table 2 shows the accuracy rates obtained on various feature sets, across all considered classification and feature selection models. For the LSA dimensionally reduced data, it no longer makes sense to use MNB, since the features are no longer valued in positive integers. However, we can still use SVM and LR for classification (Tong and Koller 2002).

Performance at feature set size 1,100 showed that all models received a boost of 5 percentage points or more using the LSA model, with SVM and LR tied for best. The frequency based Unigram model, Chi-Square, and MI model achieved very good results with a feature set of 1,100, but LSA demonstrated much better results than any other model.

LR, SVM, and MNB displayed a large increase in accuracy at the feature set size of 4,100. With the exception of the bigram model, and the combined unigram and bigram model, all our methods surpassed 90% accuracy with the feature set of 4,100. The optimal feature set size, however, seems to be at 9,100, the entire vocabulary, where we continue to see growth in accuracy without any overfitting. SVM was the overall best performer as the feature set size increased, marginally beating LR and MNB by at most two percentage points. Table 3 summarizes performance metrics for the frequency based unigram model. Precision, Recall, and the F-scores are also all above 90%. All of the Kappa values indicate

Table 2. Accuracy Comparison. We can observe that generally, SVM and LR perform Better than MNB.

Classifier	100 words	1,100 words	4,100 words	7,100 words	9,100 words
Chi-square feature selection with unigrams					
SVM	.55	.83	.94	.96	.96
LR	.55	.83	.93	.95	.95
MNB	.49	.79	.92	.94	.94
Mutual information feature selection with unigrams					
SVM	.64	.87	.95	.95	.96
LR	.64	.87	.94	.94	.95
MNB	.57	.83	.93	.94	.94
LSA dimensionality reduction with unigrams					
SVM	.76	.92	.95	.96	.96
LR	.77	.92	.95	.95	.95
Unigrams					
SVM	.62	.87	.95	.95	.96
LR	.62	.86	.94	.95	.95
MNB	.60	.83	.93	.93	.93
Bigrams					
SVM	.47	.62	.71	.75	.76
LR	.47	.62	.70	.75	.76
MNB	.47	.62	.71	.75	.77
Unigrams+bigrams					
SVM	.61	.84	.93	.95	.95
LR	.61	.85	.93	.94	.95
MNB	.60	.83	.92	.93	.93

Table 3. Classification results using 9,100 features. SVM slightly outperforms the other classifiers.

Classifier	Precision	Recall	F-score	Kappa
Unigrams				
SVM	.96	.96	.96	.94
LR	.95	.95	.95	.94
MNB	.93	.93	.93	.92

that the model results were not due to chance. While dimension reduction wasn't useful for this data set, this serves as an important benchmark for future data collection.

6.2. Error Analysis

Examples closer to the decision boundary are frequently misclassified, that is they are more difficult to identify. Figure 3 shows some misclassified examples. The UPC Description *No7 Lip Balm Coral Blossom* was predicted to belong to the personal hygiene NAPCS code. The lip balm pictured in Figure 3 closely resembles the packaging of lipstick. It may be useful in the future to consider the brand of a company like *No7* (a cosmetics brand), to more accurately predict NAPCS.

We also noticed errors due to annotator disagreement, subject experts determined *Colorsilk H/C Light Blonde* was mislabeled as Cosmetics. Respondent and annotator inconsistencies will be eliminated by creating an official gold standard.

To validate our model further, we looked at which features it is using to make decisions. Figure 4 shows the most important abbreviations in the UPC text data for the cosmetics NAPCS. Cover Girl (cg), Maybelline (mayb), Revlon (rev), are major cosmetics brands. This demonstrates our model is making understandable decisions.

UPC Description	Actual	Predicted
NEUT MEN SKN CLRNG ACNE WSH 5.1OZ	5001425000	5001450000
CALDESENE PROTECTING POWDER 5OZ	5001425000	5001450000
COLORSILK BTTRCRM NAT BRWN 60/51N	5001425000	5001450000
NO7 LIP BALM CORAL BLOSSOM .09OZ	5001450000	5001425000
DR FRED SUMMIT SKN/WHITNR 202 2OZ	5001450000	5001425000
COLORSILK H/C LIGHT BLONDE #81	5001450000	5001425000




Fig. 3. Misclassified Examples. Most of the model’s mistakes were between the cosmetics NAPCS code (5001450000) and the personal hygiene code (5001425000). Personal hygiene was predicted as cosmetics nine times. Cosmetics was predicted as personal hygiene five times.

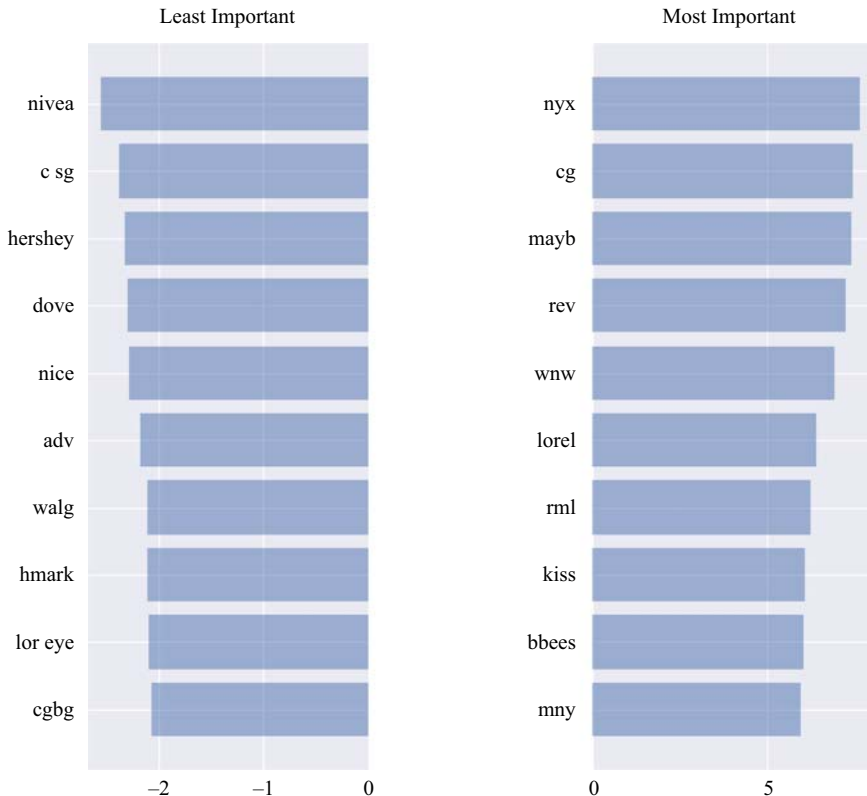


Fig. 4. Importance Scores for Cosmetics NAPCS. On the left we have the ten least important features, the ten most important are shown on the right. The most important feature was nyx, the least important was nivea.

7. Conclusions

The economic census – based on a representative sample of approximately four million businesses from about 400 industries – is vital to understanding how the nonagricultural sectors of the US economy are performing. Statistics from the economic census are used by policymakers and trade and business associations, as well as individual business owners. In 2017 the economic census was updated to collect data on more than 8,000 products through the NAPCS classification system. Due to the nature of this survey change, which included detailed product data collection, response rates were lower than expected.

The challenge of collecting the same product data in efficient, innovative, and less costly ways motivated this study. Respondents have found it burdensome to report data from a long, prespecified list of potential products. To this end, we turn to modern methods and techniques for data collection and classification. We consider integrating alternate data sources for survey collection processes, and an automated survey questionnaire system.

We examine three commonly used ML models for text classification; LR, MNB, and SVM in order to predict NAPCS codes. We found the best performing model is SVM. All of the experiments were carried out on a 2.7GHz Intel 4-Core i7-6820 CPU with 16 GB of RAM, using scikit-learn software and Microsoft Windows 10 operating system. Very high processing speeds were achieved in the learning phase. The processing time was under 180 ms on a 2.7 GHz PC. The best performing SVM model achieves a good accuracy above 95%, but there is room for improvement.

The year 2018 was an inflection point for truly groundbreaking new developments for deep learning approaches to NLP. Traditionally, NLP models were trained after random initialization of the model parameters (also called weights). Transfer learning is a concept in deep learning where you take knowledge gained from one problem and apply it to a similar problem. A technique where instead of training a model from scratch, we use models pre-trained on a large data set and then fine-tune them for specific NLP tasks (Malte and Ratadiya 2019).

The Google AI paper BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding collected honors from the ML community. Google researchers (Devlin et al. 2019) present a deep bidirectional Transformer model that redefines the state-of-the-art for eleven NLP tasks, surpassing human performance in the challenging area of question answering. Language models learn the sequential nature of language, the nuanced flow of words and sentences, and the context in which words appear (Malte and Ratadiya 2019). These models have a much more nuanced understanding of language than prior approaches.

BERT (Devlin et al. 2019), ULMFiT (Howard and Ruder 2018), and GPT (Radford et al. 2018) have shown excellent results on multiple NLP tasks like sentiment analysis, sentence similarity, and document multi-label classification. Relying on deep neural networks, various research groups trained these new models with tremendous amounts of data and compute power.

The hope is that transfer learning for NLP will significantly reduce the effort to create systems dealing with text classification. Unfortunately, in practice, it is not so trivial. BERT is a huge model with more than 100 million parameters. We not only need a GPU to

fine tune it, but also at inference time, a CPU is not enough. Using large models in a production environment requiring large amounts of compute and memory, is currently impractical for our National Statistics Office (Zafirir et al. 2019). In future work, we hope to explore these models.

LR remains the clear choice when the primary goal of model development is to look for possible causal relationships between independent and dependent variables, and a modeler wishes to easily understand the effect of predictor variables on the outcome given that the model equation is also provided. This will be useful to both analysts reviewing our predictions, and also future performance assessments.

We believe we can achieve better consistency, and response by automating classification for US businesses. We think this is a more precise strategy as indicated by our high accuracy rates. This research yields promising results and potentially reduces administration cost in survey processing. We have demonstrated a more efficient methodology for classifying NAPCS that can also be considered by other North American countries, and other official statistics agencies. This work also lays the foundation for not collecting this information only every five years, but with the timeliness that intelligent predictions provide.

8. Future Work

Our next steps include obtaining UPC data from other businesses such as a clothing and electronics store. We plan to test the models on additional data sets. If these methods continue to perform well, we will consider system-to-system agreements with several companies. Our vision is to have more companies opting into the system-to-system arrangement over time.

While there are thousands of papers published each year on how to design and train models, there is surprisingly less research on how to manage and deploy such models in production. The result of model development is typically trained models that can be used to render predictions as part of some application or service (Lee et al. 2018). The final phase of rendering predictions is often referred to as prediction serving, model scoring, or inference. Prediction serving requires integrating ML software with other systems including user-facing application code, live databases, and high-volume data streams. Due to domain-specific requirements, prediction serving has been widely studied in areas such as ad-targeting and content-recommendation (Agarwal et al. 2015). These targeted approaches do not address the full set of system challenges necessary to developing high value ML applications. We require specialized solutions engineered for our unique customer needs.

Our future work includes providing practical lessons for developing ML applications, whether you are developing your own prediction serving system or using off-the-shelf software. We also seek to incorporate the Census general-purpose low-latency prediction serving system for automatic product classification. Census currently uses an application programming interface (API), in conjunction with a web-based survey instrument, to apply ML with Logistic Regression to survey responses in real-time. The current system in production has achieved low latency, high throughput and graceful performance degradation under heavy load. Looking toward the 2022 Economic Census, we want to reduce respondent burden while providing the data of greatest need to stakeholders.

9. References

- Agarwal, D., B. Long, J. Traupman, D. Xin, and L. Zhang. 2015. "LASER: A Scalable Response Prediction Platform for Online Advertising." In Proceedings of the 7th ACM International Conference on Web Search and Data Mining, February 2014, New York, NY, USA: 173–182. DOI: <https://doi.org/10.1145/2556195.2556252>.
- Bahassine, S., A. Madani, M. Al-Sarem, and M. Kissi. 2018. "Feature selection using an improved Chi-square for Arabic text classification." *Journal of King Saud University-Computer and Information Sciences*: 1319–1578. DOI: <https://doi.org/10.1016/j.jksuci.2018.05.010>.
- Bast, H., and D. Majumdar. 2005. "Why spectral retrieval works." In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 2005, Salvador, Brazil: 11–18. DOI: <http://doi.acm.org.ezproxy.lib.unimelb.edu.au/10.1145/1076034.1076040>.
- Blanz, V., B. Scholokopf, H. Bulthoff, C. Burges, V.N. Vapnik, and T. Vetter. 1996. "Comparison of view-based object recognition algorithms using realistic 3D models." In Proceedings of International Conference on Artificial Neural Networks-ICNN'96, July 1996, Berlin, Germany: 251–256. DOI: https://doi.org/10.1007/3-540-61510-5_45.
- Chen, J., and D. Warren. 2013. "Cost-sensitive learning for large-scale hierarchical classification of commercial products," In Proceedings of 22nd Conference on Information and Knowledge Management (CIKM2013) 1351–1360. DOI: <https://doi.org/10.1145/2505515.2505582>.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "Bert: Pretraining of deep bidirectional transformers for language understanding." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, June 2019, Minneapolis, USA: 4171–4186. DOI: <https://www.aclweb.org/anthology/N19-1423/>.
- Eyheramendy, S., D. Lewis, and D. Madigan. 2003. "On the Naïve Bayes model for text categorization." In Proceedings of the 9th Workshop on Artificial Intelligence, January 2003, Key West, USA: 93–100. Available at: <http://proceedings.mlr.press/r4/eyheramendy03a/eyheramendy03a.pdf>.
- Grandini M., E. Bagli, and G. Visani. 2020. "Metrics for Multi-Class Classification: an Overview." Available at: arXiv preprint arXiv:2008.05756.
- Guenther, N., and M. Schonlau. 2016. "Support vector machines." *Stata Journal* 16: 917–937. DOI: <https://doi.org/10.1177/1536867X1601600407>.
- Howard, J., and S. Ruder. 2018. "Universal Language Model Fine-tuning for Text Classification." In Proceedings of ACL.
- Ikonomakis, M., S. Kotsiantis, and V. Tampakas. 2005. "Text classification using machine learning techniques." *WSEAS Transactions on Computers* 8(4): 966–974. DOI: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.9153&rep=rep1&type=pdf>.
- Joachims, T. 1998. "Text categorization with Support Vector Machines: Learning with many relevant features." In: *Machine Learning: Lecture Notes in Computer Science*, vol 1398. Springer: Berlin and Heidelberg.
- Joachims, T. 2001. "A statistical Learning model of text classification for Support Vector Machines." In Proceedings of the 24th Annual International ACM SIGIR Conference

- on Research and Development in Information Retrieval: SIGIR 2001, September 9–13, 2001, New Orleans, LA, USA: 128–136.
- Kozareva, Z. 2015. “Everyone likes shopping! Multi-class product categorization for e-commerce.” In *Proceedings of Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*: 1329–1333. DOI: <https://doi.org/10.3115/v1/N15-1147>.
- Lee, Y., A. Scholari, B. Chun, D. Santambrogio, M. Weimer, and M. Interlandi. 2018. “PRETZEL: Opening the black box of Machine Learning Prediction Serving.” *13th USENIX Symposium on Operating Systems Design and Implementation*. October 2018, Berkeley, USA: 611–626. DOI: <https://dl.acm.org/doi/10.5555/3291168.3291213>.
- Malte, A., and P. Ratadiya. 2019. “Evolution of transfer learning in natural language processing.” Available at: arXiv preprint arXiv:1910.07370 [cs.CL] Accessed 20 August 2019.
- Masood, A., and A. Al-Jumaily. 2013. “Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms.” *International Journal of Biomedical Imaging*: 1–22. DOI: <https://www.hindawi.com/journals/ijbi/2013/323268/>.
- Moreno, P., and P. Ho. 2003. “A New SVM Approach to Speaker Identification and Verification Using Probabilistic Distance Kernels.” In *Eurospeech*: 2965–2968. DOI: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.2378&rep=rep1&type=pdf>.
- Osuna, E., R. Freund, and F. Girosi. 1997. “Support Vector Machines: Training and Applications”, A.I. Memo No. 1602, Artificial Intelligence Laboratory, MIT. DOI: <https://ieeexplore.ieee.org/document/622408>.
- Pagliardini, M., P. Gupta, and M. Jaggi. 2018. “Unsupervised learning of sentence embeddings using compositional n-gram features.” *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, June 2018, New Orleans, USA* : 528–540. DOI: <https://www.aclweb.org/anthology/N18-1049>.
- Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever. 2018. “Improving language understanding by generative pre-training.” In *The Handbook of Contemporary Semantic Theory*. editor: S. Lappin. Blackwell, Cambridge MA & Oxford.
- Roberts, C. 1996. “Anaphora in Intensional Contexts.” In *The Handbook of Contemporary Semantic Theory*, edited by S. Lappin. Cambridge MA & Oxford: Blackwell.
- Schmidt, M., and H. Gish. 1996. “Speaker identification via support vector classifiers.” In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, May 1996, Washington, USA: 105–108. DOI: <https://ieeexplore.ieee.org/document/540301>.
- Sebastiani, F. 2002. “Machine Learning in Automated Text Categorization.” *ACM Computing Surveys* 34: 1–47. DOI: <https://doi.org/10.1145/505282.505283>.
- Shoker, L., S. Sanei, and J. Chambers. 2005. “Artifact removal from electroencephalograms using a hybrid BSS-SVM algorithm.” In *IEEE Signal Processing Letters*: 721–724. DOI: <https://doi.org/10.1109/LSP.2005.855539>.
- Thompson, K., and Y. Ellis. 2015. Exploratory Data Analysis of Economic Census Products: Methods and Results. In *JSM Proceedings, Survey Research Methods*

- Section, American Statistical Association, Seattle, WA, August 7–13. Alexandria, USA: 828–842. DOI: <http://www.asasrms.org/Proceedings/y2015/files/233942.pdf>.
- Tong, S., and D. Koller. 2002. “Support vector machine active learning with applications to text classification.” In *The Journal of Machine Learning Research (Volume 2)*: 45–46. DOI: <https://doi.org/10.1162/153244302760185243>.
- Vapnik, V.N. 2000. *The Nature of Statistical Learning Theory*, (2nd edition). New York: Springer.
- Wang, S., and C.D. Manning. 2012. “Baselines and bigrams. Simple, good sentiment and topic classification.” In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics: 90–94.
- Zafir, O., G. Boudoukh, P. Izsak, and M. Wasserblat. 2019. “Q8BERT. Quantized 8Bit BERT.” Available at: arXiv:1910.06188 [cs.CL]. Accessed 30 November 2019.
- Zhang, J., R. Jin, Y. Yang, and A.G. Hauptmann. 2003. “Modified Logistic Regression. An Approximation to SVM and Its Applications in Large-Scale Text Categorization,” In *Proceedings of the 20th International Conference on Machine Learning*. August 21–24, Washington D.C., USA: 888–895. Available at: arXiv:1910.06188 [cs.CL] (accessed 30 November 2019).
- Zhang, M., G. Johnson, and J. Wang. 2011. “Predicting Takeover Success Using ML Techniques.” *Journal of Business & Economics Research (JBER)*.

Received May 2019

Revised November 2019

Accepted March 2020

A Product Match Adjusted R Squared Method for Defining Products with Transaction Data

*Antonio G. Chessa*¹

The occurrence of relaunches of consumer goods at the barcode (GTIN) level is a well-known phenomenon in transaction data of consumer purchases. GTINs of disappearing and re-introduced items have to be linked in order to capture possible price changes.

This article presents a method that groups GTINs into strata ('products') by balancing two measures: an explained variance (R squared) measure for the 'homogeneity' of GTINs within products, while the second expresses the degree to which products can be 'matched' over time with respect to a comparison period. The resulting product 'match adjusted R squared' (MARS) combines explained variance in product prices with product match over time, so that different stratification schemes can be ranked according to the combined measure.

MARS has been applied to a broad range of product types. Individual GTINs are suitable as products for food and beverages, but not for product types with higher rates of churn, such as clothing, pharmacy products and electronics. In these cases, products are defined as combinations of characteristics, so that GTINs with the same characteristics are grouped into the same product. Future research focuses on further developments of MARS, such as attribute selection when data sets contain large numbers of variables.

Key words: Consumer price index; product relaunch; stratification.

1. Introduction

The increased availability of electronic transaction data sets for the consumer price index (CPI) offers possibilities to national statistical institutes (NSIs) to enhance the quality of index numbers. More refined methods can be applied that deal with the dynamics of consumption patterns in a more appropriate way than traditional fixed-basket methods. For instance, multilateral methods can be used to specify monthly weights based on actual sales at the most detailed item level and new items can be directly included in index calculations (De Haan and Van der Grient 2011; Ivancic et al. 2011; Krsinich 2014; Chessa 2016; Chessa et al. 2017; ABS 2017; Diewert and Fox 2017; Van Loon and Roels 2018).

Electronic transaction or scanner data sets contain expenditures and quantities sold of items purchased by consumers at physical or online sales points of a retail chain. The sales data are often aggregated by retailers into weekly sales for each individual item, which is uniquely identified by its Global Trade Item Number (GTIN); that is, the barcode.

¹ CPI department, Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, the Netherlands. Email: ag.chessa@cbs.nl.

Acknowledgments: This research was funded by a grant assigned by Eurostat, for which the author would like to express his gratitude. The author wishes to thank the associate editor and two referees for their comments and useful suggestions, and also colleagues of various NSIs for discussions on this topic.

Transaction data sets also contain characteristics, such as brand and package volume, of the items sold. While traditional price collection methods typically record prices of several tens of items in shops, electronic transaction data sets may contain several tens of thousands of items for a single retail chain.

GTINs represent the most detailed item level in electronic transaction data sets. Each item has a unique barcode. This essentially means that NSIs are given a set of tightly defined items. The ratio of monthly expenditure and quantity sold yields a transaction price, which can be followed for each item from month to month. However, items may be removed from the market and reintroduced with modified packaging, for instance, in order to fit a retailer's new product line. Quality characteristics of such 'relaunch' items may remain the same, but the barcodes may change after reintroduction, and also the prices compared with those under the previous GTINs. The barcodes of the old and new, reintroduced items have to be linked in order to capture price changes under such relaunches.

Typical market segments that feature relaunches are pharmacy items, clothing and electronics. Rates of item churn may reach high levels, since new item assortments are frequently introduced to replace their predecessors. GTINs of relaunch items have to be linked, which means that broader item definitions are needed in such situations. A generic concept is therefore introduced in this article to denote groups of linked items, which is called *product*. However, a product may also be equivalent to a single item, which is a logical choice in situations where relaunches do not occur.

Old and new GTINs can be linked manually for small samples of items. However, this becomes infeasible when NSIs aim at processing all GTINs each month, or at least those GTINs that account for a high percentage of total expenditure. To date, a method for linking GTINs of relaunch items or, in more general terms, for defining products, that is both broadly applicable and efficient does not appear to exist. Recent studies from different NSIs have shown a need for such a method (Bilius et al. 2018; Hov and Johannesen 2018; Keating and Murtagh 2018). Finding a generic and efficient method is the objective of this article.

Section 2 shows several examples of product types with different dynamics of GTINs that enter and leave an assortment. This section gives an initial, rough impression of the possible impact of different choices with regard to product definition on a price index. The central element of this article, the method MARS for defining products, is described in Section 3. MARS (product Match Adjusted R Squared) has been applied to different types of products, with different rates of churn: food items, clothing, pharmacy products and electronics. Some results are shown in Section 4.

Section 5 treats specific methodological and practical topics concerning product definition. Transaction data sets usually contain a small number of product attributes (often up to four or five in our experience). Are the available attributes sufficient for defining products? An important question from a practical perspective is how MARS could be applied in a production environment. Conclusions are presented in Section 6.

2. Assortment Dynamics

As was mentioned in the previous section, certain types of products are affected more by item relaunches than other product types. This section gives several examples with

different rates of churn. Combining GTINs based on common characteristics is one possible way of linking GTINs of relaunch items. This section also gives an initial impression of the impact of linking versus not linking on a price index, which serves to highlight the importance of the problem of product definition.

The focus in this article is primarily on transaction data. Four product types from data sets of four different Dutch retail chains are considered: milk, cheese and eggs of a supermarket chain, infant garments of a department store chain, hair care products of a pharmacy chain and televisions of an electronics retailer. About four years of data are used for the first two product types and three years of data for the other two product types.

The dynamics of products leaving and entering an assortment over time can be measured in different ways. Chessa et al. (2017) quantify the percentages of existing, leaving and entering products in each month with respect to the preceding month for different types of products. A similar measure is used in this section, which is modified on two points:

- (1) The comparison or base month is fixed, and is taken to be the first month of a 13-month time window (December of the previous year), and
- (2) The share of ‘existing products’ in the total number of products sold in a month is taken as a measure of assortment dynamics. Existing products are products that are sold both in the base month and in the current month.

These two choices can be translated into the following formal notation. Quantities of an item i sold in month t are denoted by $q_{i,t}$ and G_t is the set of items sold in month t . The comparison or base month is denoted as month 0. Let $G_{0,t}$ be the set of items/GTINs that are sold both in the base month and in (current) month t . The measure of dynamics proposed in this article does not merely count numbers of products, but quantifies the numbers sold. This choice expresses the extent of churn more appropriately. For instance, a high number of new products with low sales is not necessarily problematic, in the sense that linking old and new GTINs hardly affects a price index in such situations because of their low expenditure shares.

The proposed measure of dynamics is defined as follows at GTIN level:

$$\frac{\sum_{i \in G_{0,t}} q_{i,t}}{\sum_{i \in G_t} q_{i,t}}. \quad (1)$$

The numerator is equal to the number of items sold in month t that were also sold in the base month, and the denominator is equal to the total number of items sold in month t . It is easy to see that this measure is equal to 1 when there are no new items in month t , while it decreases when the sales quantities of new items increase. High values of the ratio therefore mean that the existing items prevail in the sales; in other words, the items sold in month t match well with the items sold in the base month. For this reason, Expression (1) will be referred to as the ‘degree of product match’ in month t .

Other choices could be made for different aspects, such as a different base period and to include disappearing products as well. Admittedly, a better choice for the base period would be to take the whole previous year instead of a single month. Products may leave temporarily. A longer period would therefore be recommendable for seasonal products. But for non-seasonal items we do not expect significant differences, as was also noted in Chessa (2018, 23–25).

In the above definition of product match, adding the quantities sold for disappearing items that were still sold in the base month to the denominator of Expression (1) would not influence the results when comparing different stratification schemes, since the denominator would be the same in every scheme. In other definitions of product match, such as the version based on numbers of products in [Chessa \(2018\)](#), disappearing items had negligible effect on the results.

Examples of product match are shown in [Figure 1](#) for the four types of products mentioned above. The graphs clearly illustrate how much product match can vary across different types of products. Rates of churn are relatively low for milk, cheese and eggs. Most items that were sold in the base month are still sold at the end of a year, as the existing items dominate the sales. The shares of existing items in the sales quantities for hair care drop to about 70% at the end of each year, so that new items account for about 30%. In this case, we are less confident of choosing GTINs as products. Relaunches are known to occur in this market segment ([Chessa 2013](#)).

The other two product types, televisions and infant garments, show extremely low product match values at the end of each year. Item turnover reaches very high levels in the course of a year. Almost entirely new product lines are introduced each year, which practically replace the previous ones. Infant garments are influenced by fashion trends, which may offer an explanation for the high turnover rates and the rapidly decreasing product match.

Traditional bilateral matched model approaches are hard to use at GTIN level under such circumstances, because of the poor continuity of GTINs over time. This is also the case for more sophisticated methods like multilateral methods, since these methods are not able to identify price changes either when relaunches occur. A separate method is developed for handling this problem, which is the purpose of MARS.

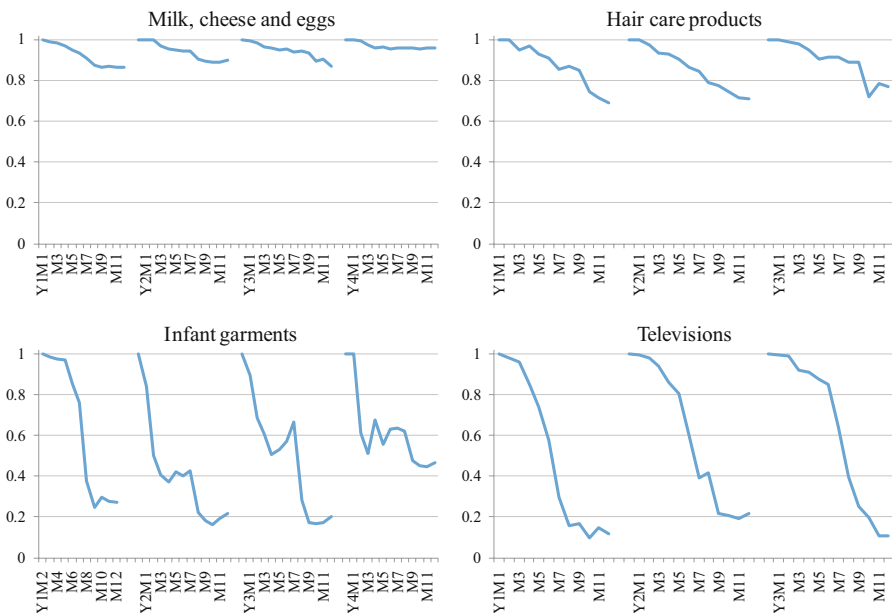


Fig. 1. Degree of product match for four product categories.

Figure 2 shows price indices when each GTIN is taken as a separate product. These indices are calculated with the ‘QU method’, a term that is used to denote a family of ‘quality adjusted’ or ‘generalised’ unit value methods (Von Auer 2014). The multilateral Geary-Khamis method is an example of the class of QU methods; it is used for processing transaction data in the Dutch CPI since January 2016 and is also the index method used in this article (Chessa 2016). The terms QU and Geary-Khamis are used interchangeably in this article for the same index method.

Making no distinction between GTINs represents the other extreme of the stratification spectrum. In that case, all GTINs would be considered of the same quality. Expenditures and sales quantities are summed over all GTINs within a product category. The ratio yields a weighted average price, known as ‘unit value’ (ILO et al. 2004). The unit value indices are shown as well in Figure 2.

The graphs show considerable differences between the two indices, especially for televisions. Existing models usually decrease in price after being introduced. New models are often more expensive than their predecessors. The index at GTIN level does not consider any of these higher prices as price changes with respect to older models. This explains why this index decreases. Higher prices of new models are seen as price increases from a unit value perspective, which explains the behaviour of the unit value index. New products may have higher prices because of relaunches, but also because they differ in terms of quality. Shifts in buying behaviour towards more expensive, higher quality products are also considered as price increases by the unit value. Similar explanations for the differences can be given for the other three product groups.

Although the indices represent two extreme cases of product stratification, the differences nevertheless make clear that product definition may have a substantial impact

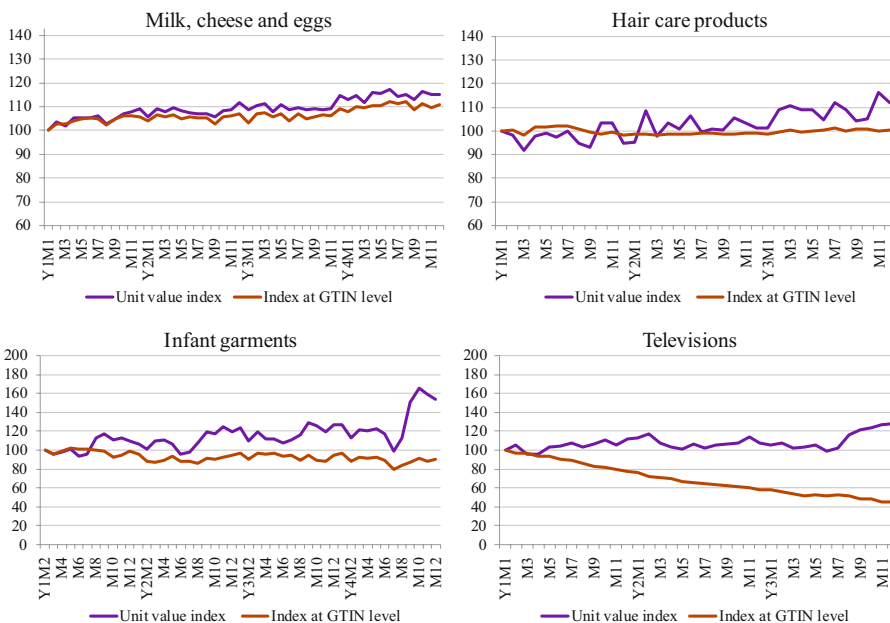


Fig. 2. Price indices at GTIN level and unit value indices for the four product categories (1st month = 100).

on a price index. We are therefore dealing with an influential choice aspect, which requires a balanced and efficient method in order to make careful choices.

3. Product Definition With MARS

3.1. Preliminary Remarks and Terminology

We start this section by introducing some terminology. The term *item* was already introduced at the beginning of this article, which is used here interchangeably with GTIN. Different items, that is, with different barcodes but not necessarily in terms of quality, may have to be linked when relaunches occur. The more generic term *product* is introduced to denote a set of one or more GTINs, which share certain quality characteristics. A product can therefore also be viewed as a combination of characteristics. The latter term is used here as a specific ‘value’ of the more generic term *attribute* or *variable*. For example, ‘screen size’ is an attribute of televisions, and ‘42 inch’ is a specific characteristic.

Items are subdivided into products, and the set of all products forms a *partition* of GTINs. The term *stratification* is used in this article as well for partition, where ‘partition’ is the formal mathematical term. In a partition, each item is assigned to exactly one product, such that products are pairwise disjoint (different products do not have items in common). GTINs may also be chosen as different products, as was illustrated in the previous section, so the set of GTINs is one of the possible partitions. In the examples with unit values there is only one product, which contains all GTINs of a product category.

GTINs should provide a suitable level of stratification for product categories with high degrees of product match. GTINs also represent the most detailed level of product homogeneity in transaction (scanner) data sets. GTINs can therefore be considered as a serious stratification candidate for milk, cheese and eggs. We will return to this in Section 4, when the results of the method MARS are presented. How to select a suitable level of stratification for product categories with low degrees of product match at GTIN level is less obvious. Broader defined products will increase product match, but homogeneity may be adversely affected.

There are different ways of partitioning GTINs into products. Available attributes of GTINs can be used for this purpose. Different selections and combinations of attributes give rise to different partitions. Each product in a partition contains GTINs with the same characteristics. For example, the attributes brand, screen size and screen type yield one partition of the set of televisions. The characteristics ‘Samsung’, ‘between 51 and 59 inch’ and ‘Ultra HD’ define a specific product of this partition. The GTINs within each product are then considered to be of the same or comparable quality. An attribute may be selected or not. This means that a set of GTINs can be partitioned into 2^n ways for n attributes, to which the partition with GTINs as distinct products can be added (thus yielding $2^n + 1$ partitions).

3.2. Formalisation of MARS

From the introduction to this section it may be clear that numerous ways of partitioning a set of GTINs exist, each of which may have a different impact on product match and homogeneity. The aim is to find a method that balances these two properties in an optimal

way in some sense. Measures of product match and homogeneity will be set up in order to operationalise this idea. The two measures are eventually combined, which allows evaluation and ranking of GTIN partitions.

First, some notation is introduced in addition to the notation already used in Section 2. We denote a partition by K and use k to indicate an element of a partition, which we have called a ‘product’. In theory, different partitions K_t can be defined each month t , for example by changing the set of attributes. However, such dynamic cases are highly complex and are probably not yet well-understood in terms of price index calculation. This study therefore deals with situations where products, once defined, are kept fixed for some period. A time window of 13 months would be in line with CPI convention, as product definitions are typically reviewed and possibly revised at the end of each year. We can therefore drop the time dimension from the notation for partitions.

We denote the degree of product match for a partition K in month t with respect to the base month by μ_t^K and the degree of product homogeneity by R_t^K . It is useful to consider desirable properties for these two measures. Attribute selection and product definition are in fact new subjects in the processing of large electronic data sets for index calculation, so that an initial attempt is made below at defining properties.

- **Property 1.** For two partitions K and K' such that K' is a refinement of K , so every element of K' is a subset of an element of K , the degree of product match of the refinement K' cannot be larger than the product match of K , that is: $\mu_t^{K'} \leq \mu_t^K$.
- **Property 2.** For two partitions K and K' such that K' is a refinement of K , the refinement K' is at least as homogeneous as K . In formal terms, we have $R_t^{K'} \geq R_t^K$ for all t .

It is reasonable to expect that broader defined products will increase product match, or at least stay the same, while the opposite is the case for homogeneity. This is in fact what the two properties say. Measures that satisfy both properties are defined below.

3.2.1. Product Match

Expression (1) applies to GTINs as separate products, so a generalisation is needed. We introduce $K_{0,t} \subseteq K$ for the set of products that are sold both in base month 0 and a second month t , with $t \geq 0$. In practical applications, month 0 will usually be December of the previous year and t a month in a 13-month window that runs until December of the present year. Let q_t^k denote the number of items sold for product k in month t ; in formal terms: $q_t^k = \sum_{i \in G_t} q_{i,t} 1_k(i)$, where the indicator function $1_k(i)$ takes the value 1 if item i belongs to product k and is equal to 0 otherwise. The degree of product match of partition K in month t is defined as follows:

$$\mu_t^K = \frac{\sum_{k \in K_{0,t}} q_t^k}{\sum_{i \in G_t} q_{i,t}}. \tag{2}$$

It is easily verified that this measure satisfies Property 1. Note that $0 \leq \mu_t^K \leq 1$ for all K .

3.2.2. Product Homogeneity

A homogeneous product is a product that consists of items of the same quality. Finding a measure for the homogeneity of a set of different products boils down to finding a method that expresses their quality differences. Hedonics is an approach that comes to mind when reflecting on this complex problem. Although this class of methods has been broadly studied, it is certainly not without limitations (Chessa et al. 2017).

Index methods usually express differences among products in terms of prices. This article also takes item prices to set up a measure of product homogeneity. Alternative choices have not yet been studied; some suggestions for further research are mentioned in Subsection 5.3. We introduce the following notation for prices. Let the price of item i in month t be denoted by $p_{i,t}$, and let \bar{p}_t^k denote the unit value for product k in month t , that is:

$$\bar{p}_t^k = \frac{\sum_{i \in G_t} p_{i,t} q_{i,t} \mathbf{1}_k(i)}{\sum_{i \in G_t} q_{i,t} \mathbf{1}_k(i)}, \quad (3)$$

where the denominator is equivalent to q_t^k . The unit value over all items in month t is denoted by \bar{p}_t :

$$\bar{p}_t = \frac{\sum_{i \in G_t} p_{i,t} q_{i,t}}{\sum_{i \in G_t} q_{i,t}}. \quad (4)$$

MARS uses the proportion of explained variance in product prices, relative to the total variance in item prices, as a measure of product homogeneity. The contribution of each product or item is weighted by the quantities sold. This yields the following weighted R squared measure:

$$R_t^K = \frac{\sum_{k \in K} q_t^k (\bar{p}_t^k - \bar{p}_t)^2}{\sum_{i \in G_t} q_{i,t} (p_{i,t} - \bar{p}_t)^2}. \quad (5)$$

More precisely, this is in fact a measure of heterogeneity between products. The complementary measure uses the price variance of GTINs within products. We want this measure to be as low as possible and, consequently, the explained variance as high as possible. Higher values of R_t^K thus denote better homogeneity.

Note that $R_t^K = 0$ when all items are combined into one product and $R_t^K = 1$ when each item is a separate product. Expression (5) satisfies Property 2. Together with the previously mentioned properties, this implies that $0 \leq R_t^K \leq 1$ for all K .

An alternative homogeneity measure could be defined by using coefficients of variation of the products in a partition. These statistics are commonly used in price statistics, not only in the CPI for data analyses, but also in PPPs. However, it can be shown that coefficients of variation do not satisfy Property 2 in general.

3.2.3. MARS: Combining Product Match and Homogeneity

The method MARS aims at evaluating and ranking item partitions. To this end, the measures of homogeneity and product match will be combined. We thus obtain an R squared measure that is adjusted for the degree of product match, which explains the acronym MARS. Some guidance on suitable functions for the combined measure could be provided by considering properties of rank orderings of partitions.

Transaction data sets, but also other data sources like web scraped data or traditionally collected data, are usually incomplete. For example, the available product variables are a subset of attributes that characterise items, and data sets are usually delivered by retailers in some aggregate form (e.g. aggregated over outlets). The ranking of partitions should not be affected by the degree of detail and aggregation of the data. These considerations motivate the following property.

- **Property 3.** For any two sets of partitions \mathcal{K}' and \mathcal{K} of the same set of items, with $\mathcal{K}' \subset \mathcal{K}$, the ordering of partitions on \mathcal{K}' should be preserved on larger sets \mathcal{K} that include \mathcal{K}' .

Property 3 has direct implications for the form of the combined measure of homogeneity and product match. For instance, an arithmetic mean of Expressions (2) and (5) does not necessarily satisfy Property 3. A multiplicative form does satisfy this property, which is the choice made for the method MARS. In relation to this it is useful to note that Expressions (2) and (5) represent measurements on different ratio scales. Item partitions can be compared and ordered in a meaningful way by taking the product of the homogeneity and product match measures. In this article, R squared and degree of product match are thus combined as follows, for every month t :

$$M_t^K = R_t^K \mu_t^K. \tag{6}$$

A multiplicative function also has the characteristic that partitions with either low values for R squared or degree of product match will be ranked as unattractive stratification options.

It will be clear that $0 \leq M_t^K \leq 1$, since Expressions (2) and (5) also have this property. Expression (6) allows us to evaluate and rank item partitions, such that the partition with the highest value of M_t^K is preferred. MARS yields values in every month t , so, in theory, the ranking of partitions may differ from month to month. The values of MARS in different months have to be combined in some way in order to produce one ranking. Different methods can be envisaged, which will be described and compared in Section 4.

One approach to overcome this is to use a price index to deflate prices and then combine the deflated prices of each product over all months. However, we prefer to stick to the approach proposed in this article, since the addition of a price index would make the method more complex, computationally more intensive and also dependent on index method. The method presented in this article can be combined with any index method, which is a major advantage since different index methods are normally used in the CPI for different forms of price collection.

A separate remark is made for the partition where all items are combined into one product (unit value case). The multiplicative form of measure (6) implies that $M_t^K = 0$ for

all t , which means that the single product partition will always be rejected. If this is found to be a limitation, then a simple remedy could be to increase both the numerator and the denominator of Expression (5) by some constant, say 1. This yields a monotonic transformation of R-squared, which therefore still satisfies Property 2, takes values in (0, 1] and preserves the value 1 for the GTIN level. Also Property 3 is still satisfied. However, in practice we do not expect that we have to use modified measures.

3.2.4. Example

We illustrate MARS with an example with a small number of items and only one month of data. Consider three GTINs, say A, B and C. The prices and quantities of the GTINs in some month are given in Table 1. The ‘status’ of each GTIN is also given: GTIN A was sold in the base month and is still sold but is about to leave at a reduced price, GTIN B is new and GTIN C is a regularly sold item. GTIN B could be seen as a relaunch of GTIN A.

We could also specify attributes in order to construct partitions. But, for simplicity, attributes and characteristics are excluded from this example since the number of GTINs is very small. Three GTINs can be partitioned in five ways:

1. A-B-C: each GTIN is a different product,
2. AB-C: GTINs A and B are combined into one product (AB), while C is a separate product,
3. A-BC: B and C are combined into one product, while A is a separate product,
4. AC-B: A and C are combined into one product, while B is a separate product, and
5. ABC: the three GTINs are combined into one product.

The results are shown in Table 2. A clear preference emerges for partition AB-C, in which exiting GTIN A is linked to new GTIN B, while ‘persisting’ GTIN C is treated as a separate product. This partition maximises the degree of product match, while it makes a minor concession in terms of homogeneity. We can thus say that MARS has picked up the relaunch. Product match is also maximised by linking new GTIN B to C, but this partition (A-BC) greatly affects homogeneity.

Table 1. Prices, quantities and status of the three GTINs.

GTIN	Price	Quantity	Status
A	2.00	1	Exiting
B	4.00	20	New
C	2.00	40	Sold in both months, not exiting

Table 2. Results of MARS for the five partitions.

Partition	R squared	Product match	Combined
A-B-C	1	0.672	0.672
AB-C	0.929	1	0.929
A-BC	0.008	1	0.008
AC-B	1	0.672	0.672
ABC	0	1	0

4. Results for Different Product Types

This section presents the results of MARS for the data of the four product categories that were introduced in Section 2. The method is applied to each of the three or four years of data. The product attributes that are available in the four transaction data sets are used to set up partitions, which are evaluated and ranked with MARS, including the partition in which each GTIN is a separate product.

In this article, products are defined either by combinations of product characteristics or as GTINs. A mixture of these two choices, by using existing GTINs as products when these are not about to disappear and attributes for linking disappearing and new GTINs, is also interesting. This option results in tighter products. As the corresponding price indices hardly differ from those that will follow in this section, this hybrid option is omitted. Details can be found in Chessa (2019, 14–16).

The available product attributes are shown in Table 3. Several remarks can be made about the attributes. GTIN classifiers denote GTIN groups defined by retailers. This information is also contained in the data, which is used by NSIs to facilitate the mapping from GTIN to COICOP in the CPI. The most detailed classifiers could also be used as additional attributes, which is done in this study. For example, the broadest of the two classifiers of milk, cheese and eggs contains seven classes. The second classifier is a further refinement. For example, ‘dairy beverages’ is subdivided into milk, buttermilk and several other groups. The classifiers for hair care distinguish between conditioners and shampoo, and the most detailed classifier mainly by hair type (e.g. normal, dry).

Apart from package volume, the other attributes are categorical variables. This is also the case for screen size (televisions), which is expressed as a range (e.g. from 28 to 32 inch). Screen type in fact means display technology (e.g. OLED, Ultra HD). Because of the level of detail used by the retailer to specify colours for clothing, we decided to compress colours into three classes (white, black and coloured). The other attributes were used as specified by the retailers. The example with colour shows that different choices can also be made in terms of how the specified characteristics (i.e. the different colours) are used. Using them as specified or further compressing the range of colours increases the number of partitions. This illustrates again the complexity of the problem of product definition from a combinatorial perspective.

Partitions have been set up by using the product attributes in Table 3. These partitions, together with the partition in which every GTIN represents a separate product, are evaluated and ranked with MARS. The method MARS yields a score in each month for every partition. The monthly scores are combined into a single score, which is eventually used to rank the partitions.

Table 3. Product attributes in the four transaction data sets.

Product category	Variables/attributes
Milk, cheese, eggs	<i>Brand, package volume, 2 GTIN classifiers</i>
Infant garments	<i>Type of garment, volume (#items), fabric, sleeve length, colour, fit, size</i>
Hair care	<i>Brand, package volume, 2 GTIN classifiers</i>
Televisions	<i>Brand, screen size, screen type</i>

In this section, the MARS scores of the last three months are taken to compute an average score for each partition. The idea behind this choice is that the effects of churn become more apparent towards the end of a year, since the products that are sold in the base month will dominate sales in the first months of a year. The results are shown in [Table 4](#). An obvious question is whether the rankings of the partitions will change when the three-month period is extended. We will come back to this later in this section.

These results invite us to make a number of remarks:

- The partition ‘GTINs as products’ is only chosen for milk, cheese and eggs, and emerges as the best partition in each year,
- The results tell us that products for clothing, hair care and televisions should be defined by sets of attributes. The three product categories are characterised by moderately to rapidly decreasing degrees of product match at GTIN level (high rates of churn), and
- For milk, cheese and eggs, hair care and televisions, the selection of attributes is quite stable over the years. The results for infant garments show more variability.

These findings probably summarise what results could be expected beforehand in a broad sense. GTINs are an appropriate choice for milk, cheese and eggs, and also for other COICOP 01 items (not shown here) but not for the other product categories because of the higher rates of churn.

GTINs are not selected as products for hair care, infant garments and televisions. The degrees of product match at GTIN level are very low for infant garments and televisions in the second half of each year ([Figure 1](#)). New GTINs may thus have to be linked to GTINs that leave the stores. The results show that in most cases a small set of attributes is selected and used for linking. The number of attributes selected in the third and fourth year for infant garments is clearly larger than in the first two years. Apparently, the degree of product match increases for tighter defined products in the third and fourth year.

The variability in the sets of selected attributes raises an important question on how MARS could be used in CPI production. In practice, decisions about product definition are made for the next year. The item partitions in this section are obtained by using the data of the corresponding year. It is therefore of tremendous practical interest to know the extent to which the price indices presented in this section change when the best partition of the current year is used to compute an index with the data of the next year. We will return to this question in Subsection 5.2.

[Figure 3](#) shows the monthly MARS scores for various GTIN partitions. Several remarks can be made based on these graphs. First, the GTIN scheme has the highest MARS scores

Table 4. Partitions selected by MARS for the four product categories in each year.

Product category	Year 1	Year 2	Year 3	Year 4
Milk, cheese, eggs	GTINs	GTINs	GTINs	GTINs
Infant garments	Type, volume, sleeve length	Type, volume, colour, fit	All attributes	All attributes, except colour
Hair care	All attributes	Brand, volume	Brand, volume	
Televisions	Screen size, screen type	Screen size, screen type, brand	Screen size, screen type, brand	

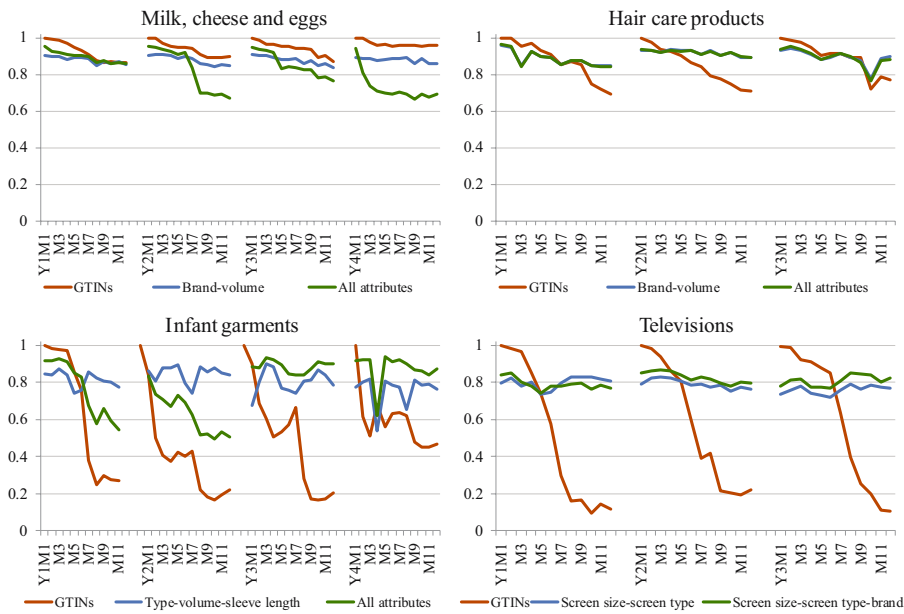


Fig. 3. MARS scores for different partitions of the four product categories.

in the first months of each year, which confirms the idea expressed above to shift the focus towards the second half of a year in terms of evaluating and ranking partitions. As was noted previously, this is most apparent for hair care, infant garments and televisions because of the high churn rates.

Second, the MARS scores in the second half of each year lead to the same ordering of the partitions in almost all months in the graphs of Figure 3. This means that extending the three-month period up to six months in order to calculate an average MARS score for every partition would hardly change the results for the data in this article. The same conclusion was obtained in a previous study (Chessa 2018).

Third, the partition that results from selecting all attributes for milk, cheese and eggs shows a particular behaviour in the second and fourth year. It reaches much lower MARS scores than the other two partitions in the second half of these two years. Product match turns out to be much lower than at GTIN level. This should not happen, since the GTIN level is a refinement of any partition based on attributes. This means that Property 1 in Subsection 3.2 is violated. The rapid decreases in product match for the ‘all attributes’ partition are caused by changes in the names of some GTIN classifiers. Products with changed names are considered as new products, which in reality are existing products with new classifier descriptions.

This example shows the practical usefulness of Property 1 and that MARS can also be used as a data monitoring tool. More generally, this also shows that stratifying items by attributes requires high quality metadata, which have to be checked each month prior to compiling the CPI. For example, retailers may decide to abbreviate names of GTIN classifiers (e.g. “bt. milk” instead of “buttermilk”). Such changes in names have to be identified in a timely manner also in order to assign GTINs to their correct COICOP.

The price indices that correspond with the highest ranked partitions in each year (Table 4) are compared with the index for GTINs as products and the unit value index. The first two indices are calculated with the previously mentioned QU method (end of Section 2). This is done on 13-month time windows. Longer-term index series are obtained by linking the 13-month indices to the indices in the base month (December of the previous year).

The three indices are shown in Figure 4. The index for milk, cheese and eggs at GTIN level is the same as the index for the best partition. The indices for the best partitions for the other product categories show large differences with the indices at GTIN level and the unit value index. The partitions are based on attributes, which enable picking up price differences between new and exiting GTINs with the same characteristics. This results in higher indices compared with the indices at GTIN level for hair care, infant garments and televisions.

The indices for the best partitions for infant garments and televisions also differ substantially from the unit value indices. The differences between these two indices for hair care are smaller. Although products are defined as combinations of characteristics, this does not mean that we should expect the corresponding indices to behave like a unit value index. Products may be tightly defined in such partitions, as is indicated by the figures in Table 5. In relation to this, note that MARS allows new and disappearing products to occur also at broader product levels than GTIN.

In some cases, the average number of GTINs per product is quite large. This may be caused by the relatively small number of attributes in the transaction data sets. On the other hand, it cannot be excluded that broader groups are sufficiently homogeneous. The

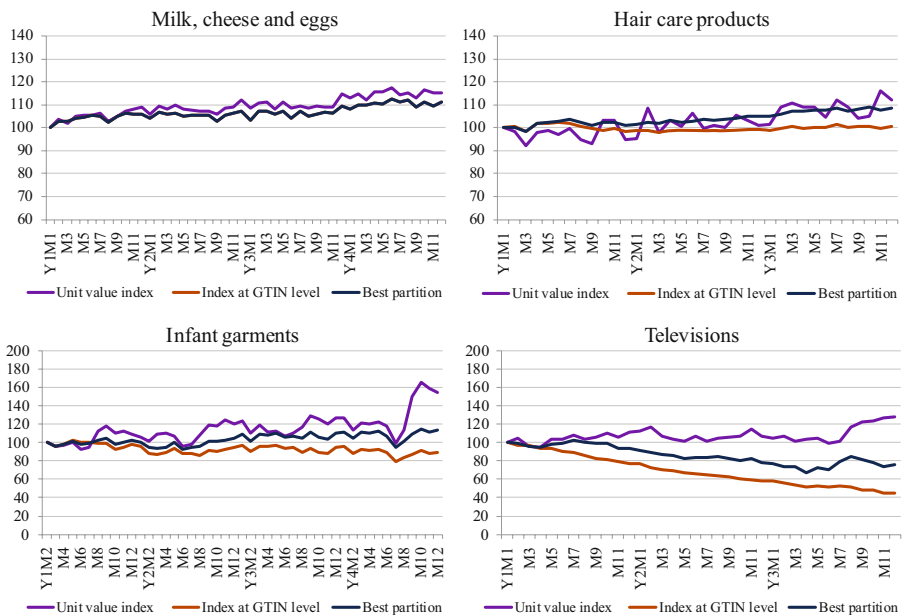


Fig. 4. Price indices for the yearly best partitions, compared with the indices at GTIN level and the unit value indices.

Table 5. Average number of GTINs per product.

Product category	Year 1	Year 2	Year 3	Year 4
Milk, cheese, eggs	1	1	1	1
Infant garments	20.3	24.6	4.2	5.8
Hair care	3.0	5.4	6.0	
Televisions	21.7	6.6	7.1	

data for infant garments contain seven attributes. The partition with four attributes reaches high MARS scores and clearly dominates the partition based on all available attributes in the first two years (Figure 3). Intuitively, this could be interpreted as a sign that the clothing data contain enough attributes. This topic will be discussed further in Subsection 5.1.

Apart from the question of whether the available product attributes are sufficient, it is also interesting to ask whether the price indices for the best partitions are stable. Sensitivity analyses were carried out by varying the weights for the two measures of MARS. Homogeneity and product match are in fact assigned the same weight in Expression (6). Tighter products are favoured by increasing the weight of homogeneity.

An increase of this weight from 0.5 to 0.75, that is, making homogeneity three times more important than product match, hardly changes the results. Brand would be selected for televisions also in the first year and all attributes would be selected for hair care in the second year. No changes were found for the other two product categories. The price indices are not shown since these hardly changed. A previous study on consumer electronics showed stable results for a similar sensitivity analysis (Chessa 2018).

5. Topics of Further Research

5.1. Information on Product Attributes

This section discusses methodological questions related to the availability of product attributes in transaction data sets. Roughly, two situations may occur: data sets containing a limited number of attributes (the situation in this article) and data sets that contain a large number of attributes. Although transaction data sets typically contain up to four or five attributes, there are NSIs, like Statistics Netherlands, that recently received data sets with tens of product attributes. The request and acquisition of more detailed data is driven by a growing experience with the use of transaction data in the CPI and retailers' growing familiarity with data acquisition and usage by NSIs. Maintaining relationships between NSIs and retailers is an important part of this process.

The two situations concerning data availability lead to the following questions:

- (1) In cases with small numbers of product attributes, is the number of attributes sufficient for defining products? Are there ways to establish whether more attributes are needed?
- (2) If the number of attributes is large, say in the order of tens of attributes, the number of stratification schemes quickly becomes too large to enumerate and evaluate. The question is how large solution spaces could be handled efficiently.

5.1.1. Small Numbers of Attributes

First, note that the question about sufficiency of product attributes becomes more relevant when the rate of churn increases (i.e. when product match decreases). GTINs are suited as products for stable assortments, in which case the question is not relevant. The COICOP milk, cheese and eggs is an example of this (Figure 3). MARS results in a selection of attributes for the other three product categories. Not all attributes are needed in most years, which could be interpreted as an indication that the available attributes are sufficient and also that products can be defined with a limited number of attributes. However, Chessa (2019, 17) showed that adding attributes from web scraped data for televisions leads to a different set of attributes and a significantly different price index.

This example shows that the question on data sufficiency is a difficult one. Suggestions to deal with this problem are given below:

- Adding attributes from a second source, such as web scraping. Building and maintaining web scrapers and analysing data requires time. If the main purpose is to use product characteristics, then it should be sufficient to scrape such information once a month. If information about a lot of attributes is available on websites, then the question below has to be addressed, on how variables can be selected from a large set of product attributes,
- Alternative approaches could be considered if web scraping is not feasible in the short term. A practical approach could be to ask CPI analysts to inspect whether products with the highest expenditure shares should be stratified further. More refined stratifications could then be evaluated with MARS,
- The second suggestion may be time-consuming. Another approach could be to generate product refinements in an automatic way by introducing a dummy attribute with a number of ‘values’ (dummy characteristics). The idea then is to draw a value for every GTIN at random, which leads to a further product refinement. This process can be repeated a number of times and the resulting stratification schemes can be evaluated with MARS. Improved MARS scores indicate a need for collecting information about additional attributes. A brief illustration of this idea can be found in Chessa (2018, 32–33).

5.1.2. Large Numbers of Attributes

NSIs have various possibilities to extend sets of product attributes in transaction data. This can be done by web scraping, but also by requesting additional information from retailers. Data sets containing tens of attributes are already being received by NSIs, which raises the question as to how such numbers can be handled for product definition. Stratification schemes based on, say up to about 10 attributes, can all be evaluated separately, but this quickly becomes inefficient for larger numbers of attributes.

The problem addressed in this article is a combinatorial optimisation problem, in which Expression (6) is the objective function to be maximised over a set of item partitions. A number of optimisation methods are worth investigating:

- *The best-first search* method selects the variable that gives the largest improvement in the objective function at each iteration step. The procedure is repeated with the

- remaining variables until the objective function does not improve anymore (Russell and Norvig 2003). This method does not guarantee a global optimum, but is attractive for its simplicity and speed, as n variables require at most $n(n + 1)/2$ evaluations,
- *Branch and bound* searches for an optimal stratification scheme by setting up a rooted tree, where a branch specifies whether a variable is either selected or not. Each node, or subtree, represents a subset of stratification schemes, with the full set contained in the root. Upper bounds of the objective function to be maximised are calculated at each node, which are compared with the best solution found so far. If an upper bound is smaller than the current value of the objective function then the corresponding subtree can be excluded from further exploration. For details, see Land and Doig (1960) and Little et al. (1963), and
 - Another method that may be worth considering is *simulated annealing* (Kirkpatrick et al. 1983; Granville et al. 1994). This is an iterative method that decides whether to stay in the current ‘state’ (i.e. stratification scheme) or move to a new state with a certain probability. The transition probabilities are specified such that a move to a scheme with a lower MARS score than the current one has positive probability, in order to avoid being trapped in local optima. The probability of accepting a scheme with a lower MARS score is gradually decreased as the number of iteration steps increases.

To get a first impression of the performance of best-first search (BFS), the method was applied to the four data sets used in this article, although cases with more attributes are obviously more interesting. The method trivially gives the same results for milk, cheese and eggs, as the GTIN level gives the highest MARS scores. However, BFS gives the same results also for the other three data sets in each year. The data on infant garments contain seven attributes. The selection of attributes changes from year to year (Table 4), but BFS yields the same results.

BFS can be applied in a simple and efficient way, and the excellent results obtained for the data sets in this article make it a very appealing and promising method. As it may get stuck in a local maximum, it is worth combining BFS with the other two methods mentioned above. BFS could be applied at each node of the tree in a branch and bound method. BFS can also be used to generate an initial state in a simulated annealing algorithm and could also be used in subsequent stages. These combined strategies are worth investigating in future research.

5.2. Practical Issues

An important question is how MARS could be applied in a balanced way in production. Different factors should be considered, such as:

- The ‘retail dimension’. Should MARS be applied to each retail chain separately or is it possible to combine the data of different retail chains for the same product category?,
- The ‘product dimension’. Also in this case the question is which level of aggregation would be feasible under resources and time constraints, and
- The ‘time dimension’. This refers to the frequency of maintenance of product definitions during a year.

Applying MARS to each retail chain for the same aggregate is probably very time consuming from a production perspective. People working in a production environment have a limited amount of time that they can spend on analysing the results of product stratification. To give an idea, there are more than 50 COICOPs at the most detailed publication level (five-digit COICOP) for COICOP 01 (food and non-alcoholic beverages). The number of applications of MARS would thus become quite large only for supermarkets. Applying MARS to the pooled data of all retail chains for the same product category seems a better idea. Retail chains should be distinguished when applying MARS, which can be done by treating retail chain as a distinct attribute.

A similar problem arises in the product dimension. Applying MARS below COICOP-5 level would result in a large number of applications. The proposal of combining data of retail chains and applying MARS at COICOP-5 level therefore seems a balanced proposal.

Another important question is how often product definitions should be checked and possibly revised during a year. Will once a year, at the end of a year, be sufficient? Or should more checks be carried out? Once a year is most compatible with current CPI routine and also saves time. A higher frequency has the advantage of timely signalling sudden changes in product assortments, like changes in product churn.

The results in Section 4 showed that product definitions may change over time. Those for infant garments even change each year. Product definitions were established by making use of the data of the same year. In practice, decisions have to be taken for the next year. An interesting question therefore is to what extent the indices will change if these are based on the product definitions established in the preceding year. Figure 5 shows that the resulting price indices are accurate. The results also suggest that it may be sufficient to check product definitions at the end of a year. Some monitoring, say after half a year, is nevertheless useful.

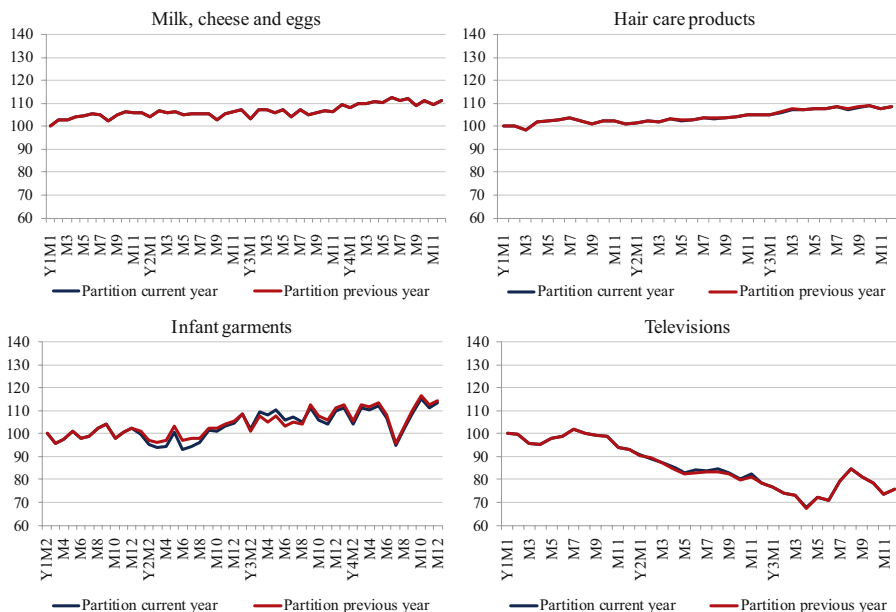


Fig. 5. Price indices when using the product definitions of the preceding year.

5.3. Possible Alternatives for MARS

Taken from a more general perspective, the idea behind MARS is to find similarity measures for items within strata (homogeneity) and over time (product match). A number of specific choices lead to the method MARS:

- MARS does not make direct use of attribute values in the homogeneity measure, but rather uses a mapping of the vector of characteristics of each item onto a one-dimensional variable, which is the item's price in this study,
- The (squared) Euclidean distance of an item to the average price of its corresponding stratum is taken here as an item's similarity measure, which are combined for different items into an overall homogeneity measure, and
- The measure of product match makes use of the numbers of items sold.

The reasons for choosing item prices and numbers of items sold were given in Section 2 and Subsection 3.2. However, these choices may also have shortcomings:

- Prices and numbers sold may vary substantially from month to month, for example during sales periods. Such variations may affect the ranking of item partitions, and
- The choice of item prices has the advantage that distance measures between items can be easily defined and calculated, but on the other hand results in indirect comparisons of the characteristics of different items.

The results in Section 4 showed that the ranking of partitions remains stable within a year for each product type, based on the MARS scores in the last six months of each year. The variability in the rankings increases over subsequent years. The set of selected attributes may change from year to year, but this could be a sign of new developments in a market segment (electronics is a typical example). Such developments should be detected and accounted for in both product definition and index calculation.

Nevertheless, it is worth studying alternatives for the choices made in MARS. Inspiration may be gained from the vast literature on cluster analysis and similarity measurement. The following suggestions may serve as a starting point:

- One direction could be to study more robust choices in the current method. For example, the number of items sold could be replaced by the number of items within products. This choice may be less sensitive to relaunches, but has the advantage of being more robust in sales periods. An additional advantage is that the number of items within item strata can also be applied to other data sources, like web scraped data,
- Alternative distance measures could be considered, such as the maximum norm instead of the Euclidean norm, and
- Alternatives to item prices in the homogeneity measure could be considered, by directly comparing the vectors of characteristics of different items. This increases the complexity of the problem, since product attributes may reflect different types of measurement (e.g. numerical, categorical or binary). Transaction data contain different types of attributes, which is also the case for the data used in this study (Table 3, Section 4). The literature on similarity measurement contains studies that focus on a specific type of variable, like numerical variables (Maftciu-Scail 2013) and

categorical variables (Boriah et al. 2008), and also studies that combine similarity measures for different types of variables (Hoffmann et al. 2015).

6. Final Remarks

Price collection is traditionally carried out by following prices of representative items that satisfy product definitions. This approach is feasible for relatively small product samples. So, historically, the problem of product definition is not new. But it needs a new, more efficient treatment when NSIs consider switching to big electronic data sets and aim at processing a significant part of such data sets or even to their full extent.

The method MARS facilitates such a transition, as this method can be used to automate the problem of product definition to a high degree. The need for a higher degree of automation and data processing was not the only motivation behind the development of MARS. The method also contributes to formalising the notion of “product”, which has been an unfilled gap in price statistics so far.

MARS yields tightly defined products for stable assortments, while products are defined as broader strata, based on common characteristics, as churn rates increase. The latter result was also found in a previous study on consumer electronics (Chessa 2018). MARS is able to identify relaunches, so that price changes associated with this phenomenon will be picked up. The results in this article and in Chessa (2018) also show that the price indices for the stratification scheme with the highest MARS scores are quite stable. The sensitivity analyses that have been carried out hardly affect the results of MARS and the corresponding indices.

MARS has been extensively applied to transaction data. Web scraping is a rapidly growing field, which motivates applying MARS also to web scraped data. This requires finding proxies for quantities sold (e.g. number of scraped product prices per month, see Chessa and Griffioen 2019). The suggestions mentioned in Subsection 5.1 for dealing with small and large numbers of product attributes and the suggestions in Subsection 5.3 for alternative versions of MARS are also worth investigating in future research.

7. References

- ABS. 2017. “Making Greater Use of Transactions Data to Compile the Consumer Price Index.” 15th Meeting of the Ottawa Group on Price Indices, 10–12 May 2017, Eltville am Rhein, Germany. Available at: <https://www.bundesbank.de/en/homepage/making-greater-use-of-transactions-data-to-compile-the-consumer-price-index-australia-635722> (accessed November 2019).
- Bilius, Å., O. Ståhl, and C. Tongur. 2018. “Coverage Bias and the Effect of Re-launches in Scanner Data: A Coffee Index.” Meeting of the Group of Experts on Consumer Price Indices, 7–9 May 2018, Geneva, Switzerland. Available at: https://www.unece.org/finance/admin/DAM/stats/documents/ece/ces/ge.22/2018/Sweden_poster_2_ppt.pdf (accessed November 2019).
- Boriah, S., V. Chandola, and V. Kumar. 2008. “Similarity Measures for Categorical Data: A Comparative Evaluation.” In Proceedings of the 2008 SIAM International Conference on Data Mining, 24–26 April 2008, Atlanta, Georgia, United States: 243–254. DOI: <https://doi.org/10.1137/1.9781611972788.22>.

- Chessa, A.G. 2013. "Comparing Scanner Data and Survey Data for Measuring Price Change of Drugstore Articles." Workshop on Scanner Data for HICP, 26–27 September 2013, Lisbon, Portugal. Available at: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_sem_lista&tipo=r&detalhe=165101941 (accessed November 2019).
- Chessa, A.G. 2016. "A New Methodology for Processing Scanner Data in the Dutch CPI." *Eurostat Review on National Accounts and Macroeconomic Indicators* 2016(1): 49–69. Available at: https://ec.europa.eu/eurostat/cros/content/new-methodology-processing-scanner-data-dutch-cpi-antonio-g-chessa_en (accessed November 2019).
- Chessa, A.G. 2018. *Product Definition and Index Calculation with MARS-QU: Applications to Consumer Electronics*. The Hague: Statistics Netherlands. Available at: <https://circabc.europa.eu/sd/a/16b279bd-04d1-44bd-8972-a537f09f7c59/Report%20Grant%202017-18%20Objective%201C.PDF.pdf> (accessed May 2021).
- Chessa, A.G. 2019. "MARS: A Method for Defining Products and Linking Barcodes of Item Relaunches." 16th Meeting of the Ottawa Group on Price Indices, 8–10 May 2019, Rio de Janeiro, Brazil. Available at: https://eventos.fgv.br/sites/eventos.fgv.br/files/arquivos/u161/product_definition_with_mars_chessa_og19.pdf (accessed November 2019).
- Chessa, A.G., and R. Griffioen. 2019. "Comparing Scanner Data and Web Scraped Data for Consumer Price Indices." *Economie et Statistique/Economics and Statistics* 509: 49–68. DOI: <https://doi.org/10.24187/ecostat.2019.509.1984>.
- Chessa, A.G., J. Verburg, and L. Willenborg. 2017. "A Comparison of Price Index Methods for Scanner Data." 15th Meeting of the Ottawa Group on Price Indices, 10–12 May 2017, Eltville am Rhein, Germany. Available at: [http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/\\$FILE/A%20comparison%20of%20price%20index%20methods%20for%20scanner%20data%20-Antonio%20Chessa,%20Johan%20Verburg,%20Leon%20Willenborg%20-Paper.pdf](http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/A%20comparison%20of%20price%20index%20methods%20for%20scanner%20data%20-Antonio%20Chessa,%20Johan%20Verburg,%20Leon%20Willenborg%20-Paper.pdf) (accessed November 2019).
- De Haan, J., and H.A. van der Grient. 2011. "Eliminating Chain Drift in Price Indices Based on Scanner Data." *Journal of Econometrics* 161: 36–46. DOI: <https://doi.org/10.1016/j.jeconom.2010.09.004>.
- Diewert, W.E., and K.J. Fox. 2017. *Substitution Bias in Multilateral Methods for CPI Construction Using Scanner Data*. Vancouver: The University of British Columbia. Discussion paper 17-02. Available at: [https://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/\\$FILE/Substitution%20bias%20in%20multilateral%20methods%20for%20CPI%20construction%20using%20scanner%20data%20-Erwin%20Diewert,%20Kevin%20Fox%20-Paper.pdf](https://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/Substitution%20bias%20in%20multilateral%20methods%20for%20CPI%20construction%20using%20scanner%20data%20-Erwin%20Diewert,%20Kevin%20Fox%20-Paper.pdf) (accessed May 2021).
- Granville, V., M. Krivanek, and J.-P. Rasson. 1994. "Simulated Annealing: A Proof of Convergence." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16: 652–656. DOI: <https://doi.org/10.1109/34.295910>.
- Hoffmann, U., A. da Silva, and M. Carvalho. 2015. "Finding Similar Products in E-commerce Sites Based on Attributes." In Proceedings of the 9th Alberto Mendelzon International Workshop on Foundations of Data Management, 6–8 May 2015, Lima, Peru. Available at: <http://ceur-ws.org/Vol-1378/> (accessed November 2019).

- Hov, K., and R. Johannessen. 2018. "Using Scanner Data for Sports Equipment." Meeting of the Group of Experts on Consumer Price Indices, 7–9 May 2018, Geneva, Switzerland. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Norway_-_session_1.pdf (accessed November 2019).
- ILO, IMF, OECD, UNECE, Eurostat, and The World Bank. 2004. *Consumer Price Index Manual: Theory and Practice*. Geneva: ILO Publications. DOI: <https://doi.org/10.5089/9787509510148.069>.
- Ivancic, L., W.E. Diewert, and K.J. Fox. 2011. "Scanner Data, Time Aggregation and the Construction of Price Indexes." *Journal of Econometrics* 161: 24–35. DOI: <https://doi.org/10.1016/j.jeconom.2010.09.003>.
- Keating, J., and M. Murtagh. 2018. "Quality Adjustment in the Irish CPI." Meeting of the Group of Experts on Consumer Price Indices, 7–9 May 2018, Geneva, Switzerland. Available at: <https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Ireland.pdf> (accessed November 2019).
- Kirkpatrick, S., C.D. Gelatt Jr, and M.P. Vecchi. 1983. "Optimization by Simulated Annealing." *Science* 220 (4598): 671–680. DOI: <https://doi.org/10.1126/science.220.4598.671>.
- Krsinich, F. 2014. "The FEWS Index: Fixed Effects with a Window Splice – Non-Revisable Quality-Adjusted Price Indices with No Characteristic Information." Meeting of the Group of Experts on Consumer Price Indices, 26–28 May 2014, Geneva, Switzerland. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/New_Zealand_-_FEWS.pdf (accessed November 2019).
- Land, A.H., and A.G. Doig. 1960. "An Automatic Method of Solving Discrete Programming Problems." *Econometrica* 28: 497–520. DOI: <https://doi.org/10.2307/1910129>.
- Little, J.D.C., K.G. Murty, D.W. Sweeney, and C. Karel. 1963. "An Algorithm for the Traveling Salesman Problem." *Operations Research* 11: 972–989. DOI: <https://doi.org/10.2307/1910129>.
- Maftiu-Scai, L.O. 2013. "A New Dissimilarity Measure between Feature-Vectors." *International Journal of Computer Applications* 64: 39-44. DOI: <https://doi.org/10.5120/10730-5734>.
- Russell, S.J., and P. Norvig. 2003. *Artificial Intelligence: A Modern Approach* (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall. DOI: <https://doi.org/10.1017/s0269888900007724>.
- Van Loon, K., and D. Roels. 2018. "Integrating Big Data in the Belgian CPI." Meeting of the Group of Experts on Consumer Price Indices, 7–9 May 2018, Geneva, Switzerland. Available at: <https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Belgium.pdf> (accessed November 2019).
- Von Auer, L. 2014. "The Generalized Unit Value Index Family." *Review of Income and Wealth* 60: 843–861. DOI: <https://doi.org/10.1111/roiw.12042>.

Received June 2019

Revised December 2019

Accepted April 2020

Variance Estimation after Mass Imputation Based on Combined Administrative and Survey Data

Sander Scholtus¹ and Jacco Daalman¹

This article discusses methods for evaluating the variance of estimated frequency tables based on mass imputation. We consider a general set-up in which data may be available from both administrative sources and a sample survey. Mass imputation involves predicting the missing values of a target variable for the entire population. The motivating application for this article is the Dutch virtual population census, for which it has been proposed to use mass imputation to estimate tables involving educational attainment. We present a new analytical design-based variance estimator for a frequency table based on mass imputation. We also discuss a more general bootstrap method that can be used to estimate this variance. Both approaches are compared in a simulation study on artificial data and in an application to real data of the Dutch census of 2011.

Key words: Data integration; finite-population bootstrap; categorical data; census tables.

1. Introduction

Administrative data are being used ever more frequently in the production of official statistics (Bethlehem 2008). In many cases, available registers cannot meet all demands made by users of official statistics and therefore have to be supplemented by other data sources, most notably sample surveys. To ensure that the resulting statistics are of sufficient quality, it is necessary to evaluate their accuracy – in particular, their variances. In general, estimating the variance of an estimator based on combined administrative and survey data is not a trivial task.

As a running example, we consider variance estimation for frequency tables in the Dutch decennial virtual population census. Since 1981, the Dutch census tables have been estimated by re-using data from existing sources rather than collecting data with a dedicated questionnaire. Nowadays, most variables needed for the census are available from administrative sources with (near-)complete population coverage. An exception occurs for *educational attainment*, which is observed partly in education registers and partly in the Labour Force Survey (LFS). For about seven million Dutch persons (of a total of 17 million), educational attainment is not observed.

¹ Statistics Netherlands, Department of Process Development and Methodology, P.O. Box 24500, 2490 HA The Hague, the Netherlands. Emails: sshs@cbs.nl and jdas@cbs.nl

Acknowledgments: The views expressed in this article are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. We would like to thank our colleagues Arnout van Delden, Jeroen Pannekoek, Eric Schulte Nordholt, and Ton de Waal for helpful discussions about this topic and for commenting on early versions of this article. We would also like to thank the Associate Editor and three anonymous referees for their review comments that have greatly improved this article.

In the Dutch censuses of 2001 and 2011, the “repeated weighting method” was used to handle missing data during estimation of census tables (Schulte Nordholt et al. 2004, 2014). It is known that repeated weighting has practical limitations for the estimation of large numbers of high-dimensional frequency tables, and these limitations were indeed encountered during the 2011 census (Daalmans 2018). As an alternative approach, it has been proposed for the next census to use mass imputation to predict all missing values of educational attainment (De Waal et al. 2018). This naturally leads to the question how to determine the precision of estimated entries in frequency tables based on mass-imputed data.

In principle, one may try to use analytical approximations to estimate the variance of statistical tables after mass imputation and, by extension, other estimators based on combined administrative and survey data. For instance, Knottnerus and van Duin (2006) developed an analytical variance approximation for estimated census tables based on repeated weighting. For the particular case of estimated frequency tables, we will present a new formula for the design-based variance of mass-imputed estimates in Section 2. This formula is of interest for a class of problems in which a categorical variable is mass-imputed, using one or more auxiliary variables. It is assumed here that a certain model is applied that predicts, for each record, the probabilities for each of the categories of the target variable. For instance, this can be a logistic regression model. Imputations for missing values are then drawn based on these predicted probabilities.

In general, a drawback of analytical variance estimation is that a separate expression has to be derived for each estimator and each data configuration. In fact, for some estimators based on combined data, deriving an adequate variance expression may prove to be impossible. Therefore, a resampling method such as the bootstrap may be a more attractive option. In Section 3, a generic bootstrap method will be described for this purpose. Unlike the analytical approach of Section 2, this bootstrap approach does not depend on the specific context of mass imputation of a categorical variable and could therefore easily be applied to many other estimators based on combined administrative and survey data.

Both approaches have been tested on a small artificial population in a simulation study (Section 5). In Section 6, we will describe an application of both approaches to real data of the Dutch virtual census 2011. Both applications in Sections 5 and 6 make use of a specific imputation model that has been proposed for the next Dutch census, the “continuation-ratio model”; this model will be outlined in Section 4. In particular, we will present a new analytical expression for the large-sample variances of predicted probabilities based on this model. Some conclusions follow in Section 7.

2. Analytical Variance Estimation for Mass-Imputed Tables

We will now consider the estimation of frequency tables based on mass imputation in more detail. Let $\theta_{hc} = \sum_{i \in U} h_i y_{ci}$ denote the true count in a particular cell of a table involving our target variable. Here, U is the target population and y_c is an indicator variable such that $y_{ci} = 1$ if person i belongs to category c of the target variable and $y_{ci} = 0$ otherwise ($c = 1, \dots, C$, where C denotes the total number of categories). Furthermore, h is a similar indicator variable for the cross-classification of all other variables in the table, that is, $h_i = 1$ if person i contributes to this cell according to these other variables and $h_i = 0$ otherwise.

As is common in official statistics, we will consider the target population and the values of variables for units in this population as fixed. All variables other than the target variable represented by y_1, \dots, y_C are supposed to be completely observed for all units in the population. The target variable is partially observed. In general, it could be observed for some units in an administrative source and for other units in a sample survey. We suppose that missing values on the target variable are imputed throughout the population (mass imputation; see, e.g., De Waal et al. 2011). After mass imputation, θ_{hc} is estimated as follows:

$$\hat{\theta}_{hc} = \theta_{hc1} + \hat{\theta}_{hc2} = \sum_{i \in U_1} h_i y_{ci} + \left(\sum_{i \in S_2} h_i y_{ci} + \sum_{i \in U_2 \setminus S_2} h_i \tilde{y}_{ci} \right). \quad (1)$$

Here, U_1 consists of all persons in U for which the target variable is observed in a register. From the remaining subpopulation, $U_2 = U \setminus U_1$, a probability sample S_2 is available with observed values of the target variable. Finally, for all $i \in U_2 \setminus S_2$, the indicator y_{ci} is unknown and replaced by an imputation \tilde{y}_{ci} in Equation (1).

We will consider the register part of the population, U_1 , to be fixed. In general, the sample S_2 may be a subsample of a sample S drawn from U . Note that the size of the register overlap $S_1 = S \cap U_1$ is then random in general.

In the application to be discussed in Section 6, Equation (1) is used to estimate a table in the Dutch virtual census, where y_1, \dots, y_C represent levels of educational attainment. In this case, the imputations are based on a model that is estimated only on data from S_2 , as the register data are known to be selective. The results in this section can be applied to any frequency table involving a categorical variable that is mass-imputed. This includes the special case where only sample survey data are available ($U_1 = \emptyset$).

In general, we suppose that the missing values of y_{ci} are imputed by drawing – independently for each person $i \in U_2 \setminus S_2$ – a vector $(\tilde{y}_{1i}, \dots, \tilde{y}_{Ci})$ from a multinomial distribution with predicted probabilities $(\hat{p}_{1i}, \dots, \hat{p}_{Ci})$, so that for each i exactly one of the values \tilde{y}_{ci} is equal to 1 and the other values are equal to 0. The predicted probabilities \hat{p}_{ci} are obtained from an imputation model estimated on the observed distribution of y_c in the sample S_2 . An example of an imputation model will be discussed in Section 4.

The uncertainty in $\hat{\theta}_{hc} - \theta_{hc}$ comes from three different sources: the random selection of persons in $U_2 \setminus S_2$ that have to be imputed, the uncertainty in the predicted probabilities \hat{p}_{ci} for these persons, and finally the stochastic drawing of the imputations \tilde{y}_{ci} based on these predicted probabilities. All of these uncertainties are directly or indirectly related to the selection of the probability sample S_2 and therefore depend on the design of this sample.

In principle, it is possible to evaluate the variance of $\hat{\theta}_{hc} - \theta_{hc}$ analytically. Here, we will make the simplifying assumption that the predicted probabilities \hat{p}_{ci} in S_2 satisfy the following condition, for all $c = 1, \dots, C$:

$$\sum_{i \in S_2} h_i (\hat{p}_{ci} - y_{ci}) = 0. \quad (2)$$

Under this assumption, the following variance formula is derived in Appendix (Subsection 8.1):

$$\text{var} \{ \hat{\theta}_{hc} - \theta_{hc} \} = E \left\{ \sum_{i \in U_2 \setminus S_2} h_i \hat{p}_{ci} \{ 1 - \hat{p}_{ci} \} \right\} + \sum_{i \in U_2} \sum_{j \in U_2} h_i h_j \text{cov} \{ \hat{p}_{ci}, \hat{p}_{cj} \}. \quad (3)$$

For many commonly used models, Equation (2) may be reasonable, at least as an approximation. For instance, it is satisfied exactly when imputation is based on a logistic regression model, fitted by maximum likelihood to a simple random sample S_2 , which includes the cross-classified variable h as a predictor (see, e.g., [Agresti 2013](#), 192–193). However, there are also many situations where Equation (2) will not hold exactly; for instance, if S_2 in the previous example has a complex survey design, the sampling weights are used when estimating the logistic regression model and these weights are not contained in the space spanned by the predictor variables. In Appendix (Subsection 8.1), it is argued that Equation (3) will be a reasonable approximation to the true variance even when Equation (2) does not hold, provided that the sampling fraction of S_2 from U_2 is sufficiently small (see the Appendix for details). In the next section, a bootstrap method will be described that can be used instead of Equation (3) if the use of Equation (2) is problematic.

To complete the specification of $\text{var}(\hat{\theta}_{hc} - \theta_{hc})$, an expression is needed for $\text{cov}(\hat{p}_{ci}, \hat{p}_{cj})$ in Equation (3). The precise form of this covariance depends on the imputation model and the sample design. For the proposed imputation model for educational attainment in the Dutch census, an approximate expression for $\text{cov}(\hat{p}_{ci}, \hat{p}_{cj})$ will be derived in Section 4 and Appendix (Subsection 8.2); there it is also discussed under which conditions Equation (2) holds (approximately) for this model. Evaluating the resulting variance formula may be computationally challenging in practice.

Based on the observed data, the variance in Equation (3) could be estimated as follows:

$$\widehat{\text{var}}(\hat{\theta}_{hc} - \theta_{hc}) = \sum_{i \in U_2 \setminus S_2} h_i \hat{p}_{ci} (1 - \hat{p}_{ci}) + \sum_{i \in U_2} \sum_{j \in U_2} h_i h_j \widehat{\text{cov}}(\hat{p}_{ci}, \hat{p}_{cj}). \quad (4)$$

Here, $\widehat{\text{cov}}(\hat{p}_{ci}, \hat{p}_{cj})$ denotes an estimator of $\text{cov}(\hat{p}_{ci}, \hat{p}_{cj})$. Again, the precise form of this term depends on the imputation model and the sample design. Note that the first term in Equation (4) follows from the fact that any random variable is an unbiased estimator for its own expectation.

Note that in [Scholtus \(2018\)](#) an alternative variance estimator was derived, without Equation (2) but with the additional technical assumption that \hat{p}_{ci} is stochastically independent of the sample inclusion indicator of S_2 for all units. Based on the simulation study presented there and in Section 5 of this article, it appears that this assumption can be failed in practice and as a result this alternative formula may underestimate the true variance. Note that this independence assumption is not easy to check based on the observed data, unlike Equation (2).

As far as we are aware, the above design-based variance of a mass-imputed estimator for categorical data has not been mentioned before in the literature. [Valliant et al. \(2000\)](#) discussed estimators of this type from a model-based perspective, but focused mainly on numerical target variables. [Kim et al. \(2020\)](#) consider variance estimation for a more general imputation model, for a situation that they call ‘mass imputation’ but which does not involve imputation of missing data throughout the target population.

3. A Bootstrap Method

The classical bootstrap (Efron 1979; see, e.g., Efron and Tibshirani 1993, for an introduction) uses resampling with replacement from an original sample to approximate the sampling distribution of a target estimator. This method cannot be used directly here, as it does not account for finite-population sampling. In fact, mass imputation is meaningful only in the context of a finite population.

Different extensions of the bootstrap to finite-population sampling have been developed; see Mashreghi et al. (2016) for a recent overview. For estimators that involve weighting or imputation, a particularly useful extension is based on generating pseudo-populations. This methodology was developed by Gross (1980), Booth et al. (1994), Canty and Davison (1999), and Chauvet (2007). At Statistics Netherlands, Kuijvenhoven and Scholtus (2011) applied this type of bootstrap method to combined register and LFS data on educational attainment, for various estimators based on weighting. Here, we will describe a slight extension of their method that can accommodate more general estimators.

As a generalisation of Equation (1), suppose that the estimator of interest is $\hat{\theta} = t(S, U_1)$, for some functional $t(\cdot)$. The underlying finite-population parameter is θ . Let π_i denote the inclusion probabilities of sample S , and write the design weights $w_i = 1/\pi_i$ as $w_i = \lfloor w_i \rfloor + \varphi_i$, with $\lfloor w_i \rfloor \in \mathbb{N}$ and $\varphi_i \in [0, 1)$. Here, $\lfloor z \rfloor$ denotes the integer part of $z \in \mathbb{R}$, that is, the largest integer that is smaller than or equal to z . The bootstrap algorithm consists of the following steps:

- For each $a = 1, \dots, A$ do the following:
 1. Create a pseudo-population \hat{U}_a^* by taking ω_i copies of each unit $i \in S$, where the random inflation weight ω_i is chosen to be $\omega_i = \lfloor w_i \rfloor$ with probability $1 - \varphi_i$ and $\omega_i = \lfloor w_i \rfloor + 1$ with probability φ_i . (Note that Kuijvenhoven and Scholtus (2011) proposed to obtain these random inflation weights by applying Fellegi's method for controlled random rounding. This has the nice property that $|\hat{U}_a^*| = |U|$ holds with certainty, rather than just in expectation.)
 2. For each $b = 1, \dots, B$ do the following:
 - a. Draw a sample S_{ab}^* from \hat{U}_a^* according to the same design that was used to draw S from U . For $k \in \hat{U}_a^*$ the inclusion probability is chosen to be $\pi_k^* \propto \pi_i$, with i the unit in the original sample S of which unit k is a copy. Here, the proportionality constant is chosen so that $\sum_{k \in \hat{U}_a^*} \pi_k^* = |S|$ holds.
 - b. Analogously to the original estimation procedure yielding $\hat{\theta} = t(S, U_1)$, construct the bootstrap replicate $\hat{\theta}_{ab}^* = t(S_{ab}^*, U_1)$.
 3. Compute the variance estimate for $\hat{\theta} - \theta$ based on pseudo-population \hat{U}_a^* as $v_a(\hat{\theta} - \theta) = (B - 1)^{-1} \sum_{b=1}^B \left(\hat{\theta}_{ab}^* - \bar{\hat{\theta}}_a^* \right)^2$, with $\bar{\hat{\theta}}_a^* = B^{-1} \sum_{b=1}^B \hat{\theta}_{ab}^*$.
- Compute the final variance estimate for $\hat{\theta} - \theta$ by averaging over the pseudo-populations: $\widehat{\text{var}}_{boot}(\hat{\theta} - \theta) = A^{-1} \sum_{b=1}^A v_a(\hat{\theta} - \theta)$.

The outer for loop of this algorithm is intended to reduce the noise due to the random assignment of integer-valued inflation weights to units with non-integer sampling weights in Step 1. Previous results in Chauvet (2007) and Kuijvenhoven and Scholtus (2011) suggest that this additional for loop may have little added value in practice (i.e., choosing $A = 1$ leads to variance estimates of a similar accuracy as choosing $A > 1$). It can

certainly be avoided in the special case that all w_i are integer-valued. For variance estimation, $B = 200$ replicates are often considered sufficient in the bootstrap literature (Efron and Tibshirani 1993 sec. 6.4). If the sample S is based on a multi-stage design, an extended version of the above pseudo-population approach may be used to account for clustering (Chauvet 2007; Mashreghi et al. 2016).

The contents of Step 2b depend on the original estimation procedure. For $\hat{\theta}_{hc}$ in Equation (1), based on mass imputation, in this step we basically re-estimate the imputation model and use this to impute the missing values in the pseudo-population. In general, the bootstrap sample S_{ab}^* may contain copies of units from $S \setminus S_2$, that is, units that overlap with the register part of the population. In analogy with the original imputation procedure, only the subset of units in S_{ab}^* that originate from S_2 , say S_{2ab}^* , is used to re-estimate the imputation model. Similarly, only the missing values for the subset of units in the pseudo-population \hat{U}_a^* that originate from S_2 , say \hat{U}_{2a}^* , are subject to imputation. The missing values in \hat{U}_{2a}^* occur for those units that are not contained in the bootstrap sample S_{2ab}^* , that is, $\hat{U}_{2a}^* \setminus S_{2ab}^*$. The register part of the pseudo-population is not imputed in the bootstrap procedure, as the contribution of the register part to $\hat{\theta}_{hc}$ is considered fixed. In summary, Step 2b consists here of the following steps:

- Define \hat{U}_{2a}^* as the subpopulation of \hat{U}_a^* consisting of copies of units from S_2 , and define $S_{2ab}^* = S_{ab}^* \cap \hat{U}_{2a}^*$.
- Use S_{2ab}^* to re-estimate the imputation model for y_1, \dots, y_C .
- Impute the missing values of y_1, \dots, y_C in $\hat{U}_{2a}^* \setminus S_{2ab}^*$ using the re-estimated model.
- Compute the replicate $\hat{\theta}_{hc,ab}^* = \sum_{k \in U_1} h_k y_{ck} + \left(\sum_{k \in S_{2ab}^*} h_k y_{ck} + \sum_{k \in \hat{U}_{2a}^* \setminus S_{2ab}^*} h_k \tilde{y}_{ck} \right)$ analogously to Equation (1).

Note that only the contribution of the non-register part to $\hat{\theta}_{hc,ab}^*$ varies between replicates.

The bootstrap method is straightforward to implement and can in fact re-use most of the code that was created to compute the original estimates. It is a computationally intensive method. A potentially useful aspect is that the time-consuming parts of the above bootstrap algorithm have to be performed only once. For instance, with mass imputation, the mass-imputed pseudo-populations could be stored and used to compute a variance estimate for any estimator $\hat{\theta}_{hc}$ by generating the replicates $\hat{\theta}_{hc,ab}^*$ ‘on the fly’. Since each pseudo-population consists of copies of units in the original sample S , all relevant information can be stored in a matrix of $|S|ABC$ integers; see Scholtus (2018) for details.

4. The Continuation-Ratio Model for Imputation

In Sections 5 and 6 we will present applications of the variance estimation methods of Sections 2 and 3 to simulated and real data on educational attainment. In these applications, use is made of the imputation approach proposed for educational attainment in the next Dutch virtual census. In the present section we will briefly discuss the underlying imputation model.

The imputation approach is based on logistic regression. Since educational attainment has $C > 2$ categories, the binomial logistic regression model cannot be applied directly. To account for the fact that educational attainment is an ordinal variable, De Waal et al.

(2018) proposed to use an extension of logistic regression known as the continuation-ratio model.

The continuation-ratio logistic regression model (Agresti 2013 sec. 8.3.6) consists of $C - 1$ ordinary binomial logistic regression models. Each of these binomial models refers to the conditional probability q_{ci} that person i does not attain a higher level than a particular level c , given that this person at least reached level c ($c = 1, \dots, C - 1$):

$$q_{1i} = P(y_{1i} = 1 | \mathbf{x} = \mathbf{x}_i).$$

$$q_{ci} = P(y_{ci} = 1 | y_{1i} = \dots = y_{(c-1)i} = 0, \mathbf{x} = \mathbf{x}_i), \quad (c = 2, \dots, C - 1).$$

Here, \mathbf{x}_i denotes a vector of auxiliary variables used in the model. Note that each conditional probability q_{ci} refers to a binary choice ($y_{ci} = 1$ or $y_{ci} = 0$). The continuation-ratio logistic regression model thus consists of a sequence of models of the form:

$$\log \left(\frac{q_{ci}}{1 - q_{ci}} \right) = \boldsymbol{\beta}_c^T \mathbf{x}_i, \quad (c = 1, \dots, C - 1). \quad (5)$$

Agresti (2013) noted that maximum likelihood estimates of all parameters in the continuation-ratio model can be obtained by estimating the $C - 1$ binomial logistic regression models in Equation (5) separately, each of them being estimated on the subset of the sample that satisfies the relevant condition of the form $y_{1i} = \dots = y_{(c-1)i} = 0$. From the estimated model parameters, the conditional probability that a person with characteristics \mathbf{x}_i has education level c may then be predicted by

$$\hat{q}_{ci} = \frac{\exp(\hat{\boldsymbol{\beta}}_c^T \mathbf{x}_i)}{1 + \exp(\hat{\boldsymbol{\beta}}_c^T \mathbf{x}_i)}, \quad (c = 1, \dots, C - 1). \quad (6)$$

Subsequently, predictions for the marginal probabilities $p_{ci} = P(y_{ci} = 1 | \mathbf{x} = \mathbf{x}_i)$ as used in Section 2 can be derived by the following recursive relation:

$$\begin{aligned} \hat{p}_{1i} &= \hat{q}_{1i}, \\ \hat{p}_{ci} &= \hat{q}_{ci} \left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki} \right), \quad (c = 2, \dots, C - 1), \\ \hat{p}_{Ci} &= 1 - \sum_{c=1}^{C-1} \hat{p}_{ci}. \end{aligned} \quad (7)$$

To account for finite-population sampling – possibly with a complex survey design – pseudo maximum likelihood estimation can be used (Chambers and Skinner 2003; see also Appendix, Subsection 8.2). Under this approach, a large-sample approximation to $\text{cov}(\hat{p}_{ci}, \hat{p}_{cj})$ for the predicted probabilities from the continuation-ratio model is derived in Appendix (Subsubsection 8.2.2). The computation of these approximate covariances involves a recursive algorithm over $c = 1, \dots, C$. Moreover, to evaluate Equations (3) or (4) this algorithm would need to be run for each pair ($i \in U_2, j \in U_2$) or at least each pair with $h_i = h_j = 1$, which can be computationally challenging for populations of realistic

size. In practice, some efficiency may be gained by noting that $\text{cov}(\hat{p}_{ci}, \hat{p}_{cj}) = \text{cov}(\hat{p}_{ck}, \hat{p}_{cl})$ whenever $\mathbf{x}_i = \mathbf{x}_k$ and $\mathbf{x}_j = \mathbf{x}_l$. (Scholtus 2018).

In general, Equation (2) need not hold exactly for the continuation-ratio model with $C > 2$. Sufficient conditions under which Equation (2) holds exactly are provided in Appendix (Subsubsection 8.2.4). There is also argued that, for large samples, Equation (2) should hold approximately in practice provided that the imputation model contains h , the variables that define the sampling design of S_2 , and the interaction of h with these variables.

5. Simulation Study

An extensive test of the bootstrap method from Section 3 for various estimators based on weighting with combined register and survey data was conducted in Kuijvenhoven and Scholtus (2011). This bootstrap method has been in use at Statistics Netherlands since 2010 to estimate variances for publications on educational attainment based on weighting. In this section, we describe the results of a simulation study for an estimator based on mass imputation. We also compare the bootstrap variance estimates to the analytical estimates from Equation (4).

All computations were done in the R environment for statistical computing. The *survey* package (Lumley 2018) was used for pseudo maximum likelihood estimation. A fast implementation of the analytical variance estimator was created using the *data.table* package (Dowle et al. 2019).

As a basis for this study, we used the data of the synthetic Samplonia population (see, e.g., Bethlehem 2009). A target population of size $N = 5 \times 745 = 3725$ was created by concatenating five copies of all persons aged over 14 in Samplonia. In this simulation, there were no register data, so $U_1 = \emptyset$; and $U = U_2$. The sample $S = S_2$ was drawn according to a simple random sampling design without replacement, with sample size $n = N/5 = 745$.

Mass imputation of educational attainment for persons in $U_2 \setminus S_2$ was based on a simplified version of the imputation approach proposed for the Dutch census, outlined in Section 4. In this simulation study, *educational attainment* was classified into $C = 3$ categories, labelled as ‘low’, ‘medium’, and ‘high’. The continuation-ratio model was applied with auxiliary information of the form *gender* \times (*age* + *income*). Here, *gender* consisted of two classes, *age* consisted of three levels, and *income* was used as a continuous variable. This model was based on the available variables in the Samplonia data set; it should be noted that the imputation model proposed for the Dutch census uses different auxiliary variables (see Section 6).

The target frequency table in this study consisted of a cross-classification of *age* and *educational attainment* (both with three levels). Table 1 below shows the true population counts (left panel) and approximate true standard deviations of the mass-imputed Equation (1) for these counts (right panel). The latter were obtained by drawing 20,000 samples from the population and for each of them estimating the model, applying mass imputation and tabulating the target estimates.

Next, we simulated 100 samples from the population and estimated the variances by two approaches:

- using the analytical variance estimator (4), with $\widehat{\text{cov}}(\hat{p}_{ci}, \hat{p}_{cj})$ based on the large-sample approximation in Appendix (Subsection 8.2);

Table 1. True counts and simulated true standard deviations for an artificial population.

Age (years)	True counts			True standard deviations		
	Educational attainment			Educational attainment		
	Low	Medium	High	Low	Medium	High
Young (15–35)	330	795	400	34.5	42.2	36.8
Middle (36–55)	115	560	480	22.3	36.8	36.1
Old (56+)	120	525	400	22.8	35.6	34.5

- using the bootstrap algorithm of Section 3, with $A = 1$ (as no rounding was necessary here) and $B = 200$.

Table 2 shows the mean estimated standard deviations and (in brackets) their standard deviation across 100 simulations. The table shows that both approaches produced estimated standard deviations that were close to their true values on average. From the values in brackets it can be seen that the estimated standard deviations from the analytical approach were more precise than those of the bootstrap approach (i.e., there was less variation between simulated samples).

We found that Equation (2) was approximately satisfied in these samples; this is as expected, since the imputation model included the variable *age* as a predictor and all observations had the same design weight. Finally, a note on computation times: across 100 simulated samples the average computation time was about 2.4 minutes per sample for the bootstrap method and just under five seconds per sample for the analytical method.

6. Application to Real Data

6.1. Setup

In this application we estimate variances for one table for the Dutch Population and Housing Census 2011. The table under consideration contains the Dutch population by *Geographic area* (12 categories), *Sex* (two categories), *Age* (eight categories) and *Educational attainment* (six categories), which makes up $12 \times 2 \times 8 \times 6 = 1,152$ cells in

Table 2. Mean and standard deviation of estimated standard deviations for estimated counts after mass imputation, based on 100 simulations.

Age (years)	Estimated analytical st. dev.			Estimated bootstrap st. dev.		
	Educational attainment			Educational attainment		
	Low	Medium	High	Low	Medium	High
Young (15–35)	34.1 (1.3)	41.8 (0.7)	36.6 (1.2)	34.1 (2.2)	41.9 (2.3)	36.4 (2.0)
Middle (36–55)	22.5 (1.7)	36.8 (0.8)	36.1 (1.0)	22.7 (2.4)	36.6 (2.0)	36.0 (2.1)
Old (56+)	22.8 (1.8)	35.4 (0.8)	34.3 (1.0)	22.5 (1.9)	35.2 (2.1)	34.5 (2.2)

total. *Geographic area*, *Sex* and *Age* are available from central population registers that fully cover the target population of the census. *Educational attainment* is available for a subset of the population.

In this evaluation study *Educational attainment* is estimated from the Educational Attainment File (EAF), with reference day January 1, 2011. The EAF is a database that includes data from multiple registers and LFS data from multiple years. Data from the EAF can be matched to the population registers at the micro level. The EAF is planned to be used for the 2021 census.

The EAF contains a ‘register part’ and a ‘non-register part’, which include 9,363,909 and 7,291,890 persons, respectively. These two parts refer to the people for which register information on educational attainment is and is not available (U_1 and U_2 in the notation of Section 2). The focus of this simulation study is entirely on the non-register part. For 340,472 out of 7,291,890 persons, educational levels are available from an LFS. The missing observations are imputed at the micro level, using the continuation-ratio model that was introduced in Section 4. The imputed data are used to estimate counts and variances for the aforementioned table: *Geographic area* \times *Sex* \times *Age* \times *Educational attainment*.

It should be noted that for this study we had only limited information about the origin of the sample data S_2 . First, we did not have any information about the larger sample S from which S_2 was obtained by removing the overlap with U_1 . Second, the data in S_2 are an integrated sample of several LFS rounds and we only had the final recalibrated sampling weights but no information about the underlying sampling design and the way different years were combined. For the purpose of this study, we approximated the design of S_2 by that of a simple random sample without replacement from U_2 . This should give reasonable results for comparing the analytical and bootstrap methods, as the same approximation was used for both approaches. For a future application to the real Dutch census, a better approximation will be made which accounts for the complex survey design of the LFS.

The EAF-based data set was enriched with information from other data sources that are included in the system of Social Statistical Data sets (SSD). The variable *Income* (six categories: five quantiles and unknown/not available) was used as a stratification variable in the imputation model; that is to say, a separate continuation-ratio model was estimated for each income class. *Income* has been chosen because it has a relatively strong association with *Educational attainment* (Daalmans 2017). The variables *Age*, *Geographic area* and *Sex* that are contained in the target table were also used as an auxiliary variable for most of the cases. However, for the lowest class of education and three strata of income only *Sex* was taken as auxiliary variable. This choice will be explained further on.

As before, the bootstrap method in Section 3 was implemented with $B = 200$ and $A = 1$ and all computations were done using R and the *survey* and *data.table* packages.

6.2. Results

The computation time was much longer for the bootstrap method than for the analytical method: about 21.5 hours versus 48 minutes and 26 seconds. Note that the bootstrap computations could easily be parallelised across multiple processors to save time. However, we did not do this here.

For the analytical Equation (4), for each cell of the target table a double sum needs to be computed of all people that have the same combination of values for the auxiliary variables in the imputation model, that is, *Age*, *Sex* and *Geographic area* and stratification variable *Income*. The maximum number of items over which this sum extends was 2,937 and the average number was 296. The largest double sum contained approximately nine million elements (2937^2), which does not seem too problematic from a computational point of view.

A first question is whether 200 bootstrap iterations are enough. To answer that question, we consider the ratio of the estimated standard deviations after 200 and 190 iterations. Table 3 below shows percentiles of these ratios for the 1,152 cells of the target table

For all but a few cells the difference is smaller than $\pm 3\%$. For more than half of the cells the difference is even below $\pm 1\%$. These results also show that there is no structural upward or downward trend of the standard deviations between the 190th and 200th iteration.

We now give an impression about the results of the bootstrap procedure. The coefficient of variation (CV) has been computed for each cell, that is, the ratio of the estimated standard deviation to the mean, that is, the average cell count over the 200 bootstrap samples. As we consider the non-register part of the EAF only, the mean is derived from the non-register part of the EAF.

In Figure 1, the CV is plotted against the sample size of the cell. As expected, cells with the least number of sample survey observations have the highest CVs.

We now turn to the main question of this simulation exercise: the difference in results for the standard deviations between the bootstrap method and the analytical approximation. First, we plot the CVs of both approaches against each other in Figure 2, subdivided for each (imputed) educational category.

The graph shows similarities between the CVs of both methods, especially for the categories 3 and 4 that are the most frequently occurring. To assess this further, percentiles of the ratios of the CVs are given in Table 4. These ratios compare the CVs of the analytical method with those of the bootstrap method and have been derived from all cells, that is, all educational levels.

The difference between the analytical and the bootstrap method is reasonably small. For instance, the median value is 0.97, the 90th percentile is 1.79 and the 10th percentile is 0.59. This gives us some empirical evidence that the analytical procedure gives close approximations to the true variance.

In Figure 3, density plots are shown of the ratio of the CVs of the analytical and bootstrap method. The solid black line represents the density of all cells in the table. The other lines represent the density for subsets of 20% of cells in the table, where the cells are ordered by sample size from smallest (quintile 1) to largest (quintile 5). As expected, the largest deviations occur for cells with relatively few observations. Interestingly, it appears that the analytical method tends to underestimate the CV slightly compared to the bootstrap method for all quintiles except the first one, where it has a tendency to overestimate the CV.

Table 3. Percentile of the ratios of standard deviations after 200 and 190 iterations.

Percentile	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99
SD-ratio	0.981	0.985	0.987	0.992	0.998	1.007	1.014	1.020	1.035

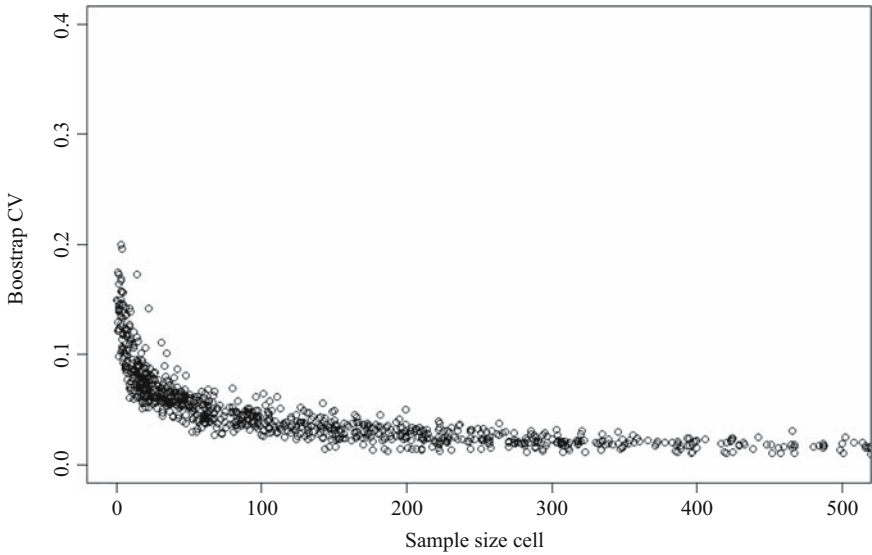


Fig. 1. Coefficient of variation for the bootstrap method.

A possible explanation for this underestimation in cells where the sample size is sufficiently large is that the condition in Equation (2) does not hold in this application, as the imputation model contains the main effects of *Sex*, *Geographic area* and *Age* but not their interactions. Thus, variance Equation (4) neglects a source of uncertainty; see also the discussion at the end of Appendix (Subsection 8.1). To illustrate this further, we also

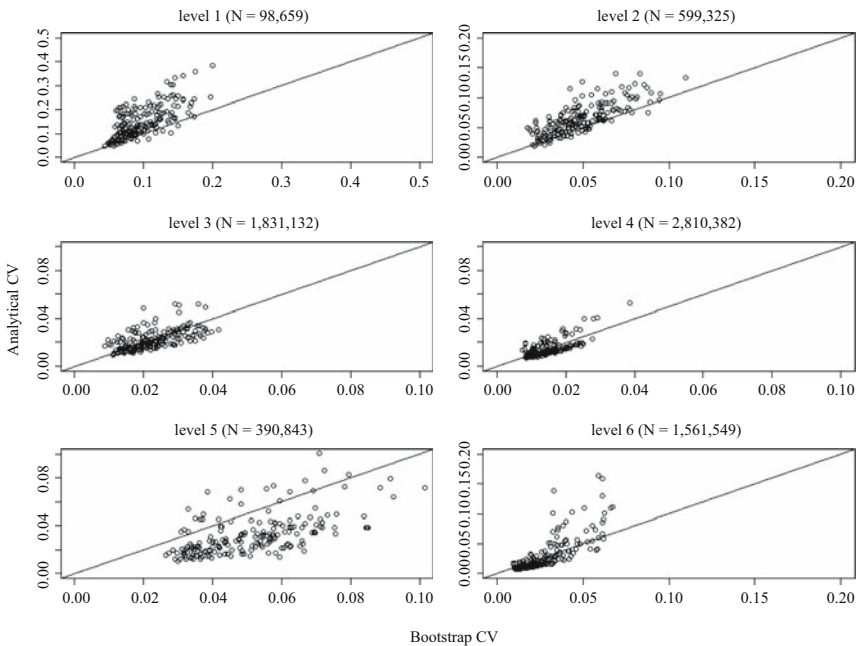


Fig. 2. Coefficient of variation; analytical method against bootstrap.

Table 4. Percentiles of the ratios of CVs based on the analytical and bootstrap method.

Percentiles	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99
CV-ratios	0.356	0.494	0.589	0.748	0.967	1.360	1.785	2.086	2.760

computed the analytical CVs and bootstrap CVs for the table *Geographic area* \times *Educational attainment*, that is, one of the marginal tables of the original table. For this lower-dimensional table, Equation (2) is expected to hold approximately in this application. In Figure 4, the ratio of the CVs of the analytical and bootstrap method is plotted against the sample size per cell for the 72 cells in this table. It is seen that here the two approaches are in closer agreement and that the ratio of the CVs tends to one for cells with large sample sizes.

Finally, it should be noted that for some strata the number of observations was too small for a reliable estimation of all regression coefficients. Initially, *Sex*, *Geographic area* and *Age* were included in the model for each stratum. For some strata, this led to extremely high analytically derived standard deviations, when compared to the bootstrap. The problem was especially apparent for the highest income classes. In these strata, low educational attainment barely occurs. As a consequence, the regression coefficients for estimating the probability of low educational levels could not be reliably estimated. Due to near-multicollinearity of the auxiliary variables, extremely high standard deviations were returned for some coefficients by Equation (9) in Appendix (Subsubsection 8.2.1). The impact of this diminished after reducing the regression model, that is, after using only *Sex* as an auxiliary variable in certain strata. This shows that the analytical approach is sensitive to model selection. It also appears that the analytical variance estimates are more sensitive to model over-parametrisation than the mass-imputed estimates themselves.

Model selection is an important step in an application to the census anyhow, due to the availability of a large number of auxiliary variables. All variables that appear in the target

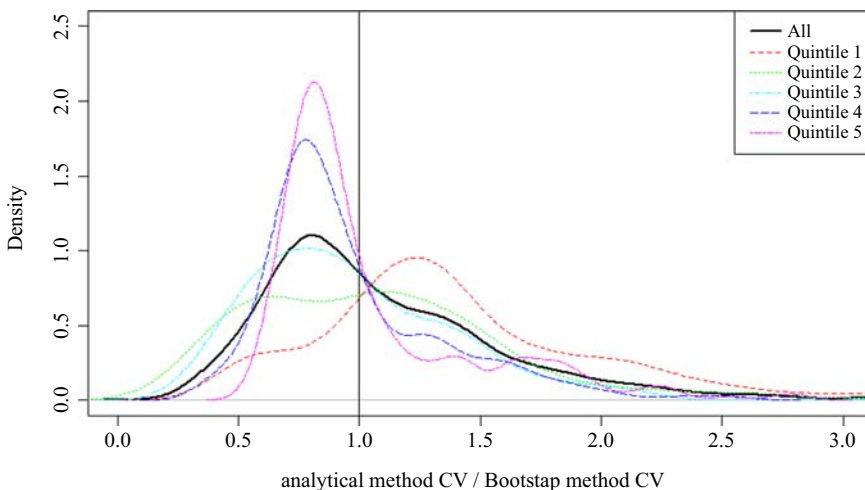


Fig. 3. Ratio of coefficient of variation according to the analytical and bootstrap methods; density of all cells and density by quintiles based on sample size per cell.

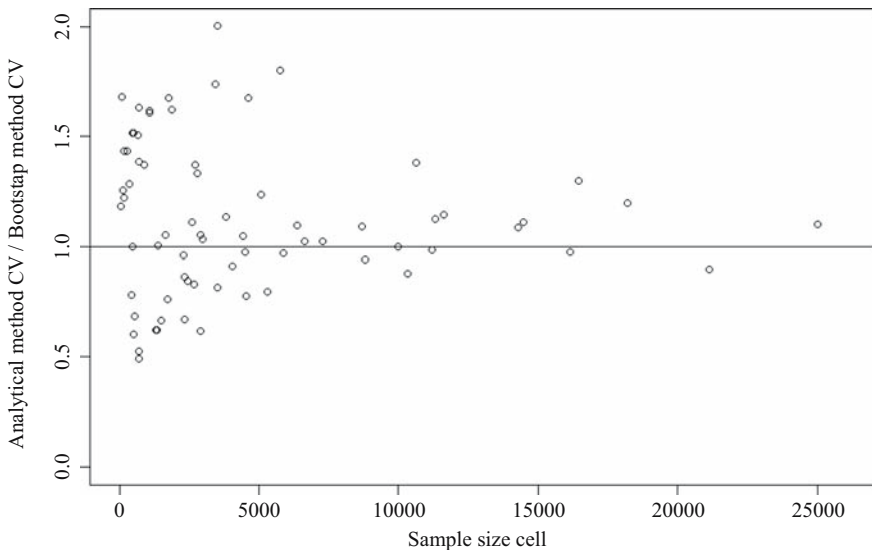


Fig. 4. Ratio of coefficient of variation according to the analytical and bootstrap methods (marginal table: geographic area \times educational attainment).

tables to be produced can be used. These are: *Age*, *Country of citizenship*, *Place of usual residence*, *Household status*, *Industry / sector of economic activity*, *Size of locality*, *Place of birth*, *Place of usual residence one year prior to the census*, *Sex*, *Status in employment*, and *Year of arrival in the country*. Some of these have a large number of categories, which often include ‘rare’ cases. An example from the Dutch census is the category “Oceanic” for *Country of citizenship*. On the one hand, it is attractive to include as many auxiliary variables in the imputation model as available, because a relation between a target variable (to be imputed) and some other variable can only be properly measured if that other variable is taken as an auxiliary variable in the imputation model. Also, a model with few auxiliary variables is more likely to be mis-specified, leading to bias in the mass-imputed estimator. On the other hand, as we have seen before, including many auxiliary variables in the imputation model might lead to a large variance, because the number of observations can be insufficient for reliable estimation of the regression coefficients. Hence, a trade-off must be made between variance and (potential) bias and it can be inevitable to aggregate some categories of auxiliary variables in an imputation model. Several automatic procedures can be implemented for this purpose. In the analytical approach, one might manually ‘optimise’ the aggregation level of the auxiliary variables for the sample survey at hand. In the bootstrap approach, one must take into account that different resamples are obtained at each iteration. Therefore, the bootstrap methods require some automation in the selection of the imputation model.

7. Discussion and Conclusion

This article has developed a new design-based variance formula for estimated frequency tables after mass imputation of one of the variables. A case is considered in which one target variable is observed for a sample of the population and where all missing values for the target population need to be imputed. Furthermore, it is postulated that the imputations

are based on a model that predicts, for each record, the probabilities for the different categories of the target variable.

The above problem is relevant for the Dutch census. One of the census variables, educational attainment, cannot be observed for the entire target population. All non-observed values can be imputed, for instance by using logistic regression or one of its variants. Several frequency tables need to be produced for the Dutch census that cross-classify educational attainment with other variables that are observed for the whole population. The analytical expressions in this article can be used to estimate the variances of the cells of these tables. Although the case of the Dutch census has been mentioned often in this article, the problem is more generally relevant for other applications in which a categorical variable is imputed for an entire target population.

In Section 2 we presented an analytical variance expression that allows for a broad class of imputation methods. Each imputation method gives rise to a specific completion of the expression. The special case of the continuation-ratio model was considered in Section 4. We have derived a new analytical variance approximation for this specific model.

As an illustration, the analytical approximation was applied to a Dutch census table in Section 6. The results have been compared to outcomes of a bootstrap method, adapted from [Kuijvenhoven and Scholtus \(2011\)](#), which can be applied more generally to estimators based on combined data sources. The results demonstrated that the analytical variance approximation works reasonably well. For the large majority of cases, the estimated standard deviation was between 0.5 and 1.5 times the standard error of the bootstrap method. Moreover, we have demonstrated that both methods can be applied to a data set of a significant size, that is, approximately seven million imputations.

The analytical variance estimator was derived under the simplifying assumption that Equation (2) holds, at least approximately. This condition can be verified in practice: if the left-hand-side of Equation (2) is small in comparison to $\hat{\theta}_{hc2}$, then the analytical variance approximation proposed here should work well. In situations where this condition does not hold, our analytical approach is likely to underestimate the true variance for cells with a sufficiently large sample size (see Appendix, Subsection 8.1). The question of the design-based variance of the mass-imputed Equation (1) is relevant for our statistical office and, we believe, for the wider official statistical community. As far as we are aware, there is no analytical method currently available in the literature to derive this variance in a general setting, which avoids introducing Equation (2) or some other assumption that might be violated in practice. Alternatively, the bootstrap method from Section 3 may be used regardless of whether Equation (2) holds.

In general, the computation time can be a limiting factor for the estimation of variances for large data sets. The analytical approximation requires the computation of a double sum. The number of terms in this sum can become quite large. However, the bootstrap method can be expected to have an even larger computational burden for many applications. The bootstrap method requires work that is equivalent to imputing the entire target population many times. In our case study, approximately seven million records needed to be imputed 200 times. This led to a computation time that was about a factor 30 larger than for the analytical approximation. This illustrates that, although the bootstrap is

more flexible than the analytical approach, analytical approximations may still be useful and even necessary for variance estimation in practice.

Our application also revealed that the analytical approximation of the variances is more sensitive to model over-parametrisation than the mass-imputed estimates themselves. This stresses the importance of using the analytical approximation in combination with a proper model selection method. When this is neglected, there is a risk of a serious over-estimation of variances.

An alternative variance estimation approach that was not considered in this study is multiple imputation (Rubin 1987). Similar to our bootstrap method, this would require imputing the missing values of educational attainment throughout the population several times. In fact, the bootstrap algorithm of Section 3 could be adapted to be used for multiple imputation. A practical advantage of multiple imputation compared to direct bootstrapping would be that it requires fewer replicates. A limitation of multiple imputation, compared to the approaches considered here, is that it cannot be used to estimate the variance of an arbitrary estimator, but only that of the associated multiple imputation estimator. Thus, in this approach the variance estimation method actually guides the choice of the estimator itself, which then has to be based on a multiply-imputed file. For practical reasons there are currently no plans to generate more than one imputation per person in the official microdata of the Dutch virtual census, so multiple imputation is not an option for variance estimation in this application. For this reason, it was not considered here. However, it may be interesting to compare multiple imputation to our analytical and bootstrap method in a more extensive future study.

Future work may focus on extending the variance estimation methods considered here. The analytical approach could be extended to the case of imputing multiple categorical variables, or a combination of categorical and numerical variables. An extension of the analytical approach is also needed to handle imputation methods that are not based on a parametric model, such as hot deck imputation. Finally, it may be interesting to develop variance estimation techniques – either analytical or by means of resampling – that can account for uncertainty in the measurement of register-based variables, including the effects of micro-integration (Bakker 2011) when overlapping data are available from a register and a survey.

8. Appendix

8.1. Derivation of Variance Formula (3)

Conditioning on a realisation of the random sample S_2 and using a standard decomposition formula for conditional variances, we can write the variance of the estimator $\hat{\theta}_{hc}$ in Equation (1) as:

$$\text{var}(\hat{\theta}_{hc} - \theta_{hc}) = E\{\text{var}(\hat{\theta}_{hc} - \theta_{hc}|S_2)\} + \text{var}\{E(\hat{\theta}_{hc} - \theta_{hc}|S_2)\}.$$

Substituting the definitions of θ_{hc} and $\hat{\theta}_{hc}$ from Equation (1) and using the fact that imputations for different persons are independent, we obtain (with $\theta_{hc2} = \sum_{i \in U_2} h_i y_{ci}$):

$$\begin{aligned}
\text{var}(\hat{\theta}_{hc} - \theta_{hc}) &= E\{\text{var}(\hat{\theta}_{hc2} - \theta_{hc2}|S_2)\} + \text{var}\{E(\hat{\theta}_{hc2} - \theta_{hc2}|S_2)\} \\
&= E\left\{\text{var}\left(\sum_{i \in U_2 \setminus S_2} h_i(\tilde{y}_{ci} - y_{ci})|S_2\right)\right\} \\
&\quad + \text{var}\left\{E\left(\sum_{i \in U_2 \setminus S_2} h_i(\tilde{y}_{ci} - y_{ci})|S_2\right)\right\} \\
&= E\left\{\sum_{i \in U_2 \setminus S_2} h_i \text{var}(\tilde{y}_{ci} - y_{ci}|S_2)\right\} + \text{var}\left\{\sum_{i \in U_2 \setminus S_2} h_i E(\tilde{y}_{ci} - y_{ci}|S_2)\right\} \\
&= E\left\{\sum_{i \in U_2 \setminus S_2} h_i \hat{p}_{ci}(1 - \hat{p}_{ci})\right\} + \text{var}\left\{\sum_{i \in U_2 \setminus S_2} h_i(\hat{p}_{ci} - y_{ci})\right\}.
\end{aligned}$$

For the third equality it was used that $h_i^2 = h_i$. For the last equality it was used that $E(\tilde{y}_{ci}|S_2) = \hat{p}_{ci}$ and $\text{var}(\tilde{y}_{ci}|S_2) = \hat{p}_{ci}(1 - \hat{p}_{ci})$ based on a multinomial distribution. The first term in this expression is easily estimated in practice (cf. Equation (4)), so we leave this term as it is and focus on elaborating the second term.

It should be noted that \hat{p}_{ci} , the predicted probability that person i belongs to category c , is obtained from a model that is estimated on the observed data from sample S_2 . Hence, \hat{p}_{ci} is a random variable that depends on S_2 . In general, this complicates the evaluation of the above variance component. However, under the assumption that Equation (2) holds, the second term can be simplified and we find:

$$\begin{aligned}
\text{var}\left\{\sum_{i \in U_2 \setminus S_2} h_i(\hat{p}_{ci} - y_{ci})\right\} &= \text{var}\left\{\sum_{i \in U_2} h_i(\hat{p}_{ci} - y_{ci})\right\} \\
&= \sum_{i \in U_2} \sum_{j \in U_2} h_i h_j \text{cov}(\hat{p}_{ci} - y_{ci}, \hat{p}_{cj} - y_{cj}) \\
&= \sum_{i \in U_2} \sum_{j \in U_2} h_i h_j \text{cov}(\hat{p}_{ci}, \hat{p}_{cj}).
\end{aligned}$$

The last equality follows since we treat the values of y_{ci} in the finite population as fixed here. Hence, Equation (3) is obtained.

More generally, if Equation (2) is not satisfied, it follows from the above derivation that

$$\text{var}(\hat{\theta}_{hc} - \theta_{hc}) = \tilde{V} + V_{S_2} - 2C_{U_2, S_2},$$

where \tilde{V} denotes the variance from Equation (3), $V_{S_2} = \text{var}\{\sum_{i \in S_2} h_i(\hat{p}_{ci} - y_{ci})\}$ and $C_{U_2, S_2} = \text{cov}\{\sum_{i \in U_2} h_i(\hat{p}_{ci} - y_{ci}), \sum_{i \in S_2} h_i(\hat{p}_{ci} - y_{ci})\}$. Since $\sum_{i \in S_2} h_i(\hat{p}_{ci} - y_{ci})$ is an unweighted sample total, it seems intuitively reasonable to expect the contributions of the terms V_{S_2} and C_{U_2, S_2} to $\text{var}(\hat{\theta}_{hc} - \theta_{hc})$ to be negligible if the sampling fraction of S_2 is ‘small enough’, even when Equation (2) does not hold.

To make this argument somewhat more precise, we suppose that S_2 is a simple random sample with $|S_2|/|U_2| \ll 1$. Let N_h denote the number of units in U_2 with $h_i = 1$ and let n_h denote the expected number of units in S_2 with $h_i = 1$. It can be shown that the standard Horvitz-Thompson estimator for θ_{hc} then has a variance of order $O(N_h^2/n_h)$. In practice, the available background variables to impute y_c usually have limited predictive value, so we expect the variance of the mass-imputed estimator to have the same order of magnitude: $\text{var}(\hat{\theta}_{hc} - \theta_{hc}) = O(N_h^2/n_h)$. (We have also confirmed this empirically for the application in Section 6; results not shown here.) Furthermore, if Equation (2) does not hold, V_{S_2} has the same order as

$$\text{var} \left\{ \sum_{i \in S_2} h_i (p_{ci} - y_{ci}) \right\} = n_h^2 \text{var} \left\{ \frac{1}{n_h} \sum_{i \in S_2} h_i (p_{ci} - y_{ci}) \right\} = O(n_h).$$

Here, we used that $\hat{p}_{ci} - p_{ci}$ is of a smaller order than $p_{ci} - y_{ci} = O(1)$ and that $\text{var} \left\{ \frac{1}{n_h} \sum_{i \in S_2} h_i (p_{ci} - y_{ci}) \right\} = O(1/n_h)$; see [Särndal et al. \(1992, 214\)](#) for the latter result. For the covariance term, we find:

$$|C_{U_2, S_2}| \leq \sum_{i \in U_2} h_i \sqrt{\text{var}(\hat{p}_{ci} - y_{ci})} \sqrt{V_{S_2}} = O(N_h \sqrt{n_h}).$$

Thus, if S_2 is a simple random sample and $n_h \ll N_h^{2/3}$, then both V_{S_2} and C_{U_2, S_2} are of negligible order and it follows that $\text{var}(\hat{\theta}_{hc} - \theta_{hc}) \approx \tilde{V}$ even when Equation (2) does not hold exactly.

Furthermore, under the weaker condition $n_h \ll N_h$ the term V_{S_2} is still negligible but C_{U_2, S_2} may not be when Equation (2) does not hold. Note that

$$C_{U_2, S_2} = V_{S_2} + \text{cov} \left\{ \sum_{i \in U_2 \setminus S_2} h_i (\hat{p}_{ci} - y_{ci}), \sum_{i \in S_2} h_i (\hat{p}_{ci} - y_{ci}) \right\}.$$

Provided that the imputation model does not introduce bias, the average prediction error $h_i(\hat{p}_{ci} - y_{ci})$ over U_2 should tend to cancel out for repeated samples. This suggests that we may expect in practice that $\text{cov} \left\{ \sum_{i \in U_2 \setminus S_2} h_i (\hat{p}_{ci} - y_{ci}), \sum_{i \in S_2} h_i (\hat{p}_{ci} - y_{ci}) \right\} < 0$. Therefore, since

$$\text{var}(\hat{\theta}_{hc} - \theta_{hc}) = \tilde{V} - V_{S_2} - 2 \text{cov} \left\{ \sum_{i \in U_2 \setminus S_2} h_i (\hat{p}_{ci} - y_{ci}), \sum_{i \in S_2} h_i (\hat{p}_{ci} - y_{ci}) \right\}$$

for cells where V_{S_2} is negligible but C_{U_2, S_2} is not, we may expect that $\text{var}(\hat{\theta}_{hc} - \theta_{hc}) > \tilde{V}$. That is to say, when Equation (2) is violated and $n_h \ll N_h$ but not $n_h \ll N_h^{2/3}$, then Equation (3) is more likely to underestimate the true variance than to overestimate it.

Although these conclusions have been derived for the case of a simple random sample, we expect that they extend to other sampling designs that are commonly used in practice for social surveys. In practice, the available design variables for social surveys usually are correlated only weakly to the variables of interest, leading to a limited design effect on the order of magnitude of various variance terms.

8.2. Details of the Continuation-Ratio Model

8.2.1. (Pseudo) Maximum Likelihood Theory for the Continuation-Ratio Model

First, suppose that S_2 were a sample of independent, identically distributed observations from the underlying distribution of the continuation-ratio model of Section 4. [Agresti \(2013, sec. 8.3.6\)](#) shows that the log likelihood function may then be written as:

$$\ell = \sum_{c=1}^{C-1} \sum_{i \in S_{2, \geq c}} \{y_{ci}(\boldsymbol{\beta}_c^T \mathbf{x}_i) - \log [1 + \exp(\boldsymbol{\beta}_c^T \mathbf{x}_i)]\}$$

where $S_{2, \geq 1} = S_2$ and $S_{2, \geq c}$ ($c = 2, \dots, C - 1$) denotes the subsample of units in S_2 with $y_{1i} = \dots = y_{(c-1)i} = 0$. As the $C - 1$ terms of ℓ have no parameters in common, it is clear that maximum likelihood estimates can be obtained by solving the following likelihood equations:

$$\begin{aligned} 0 &= \frac{\partial \ell}{\partial \boldsymbol{\beta}_c} = \sum_{i \in S_{2, \geq c}} \left\{ y_{ci} \mathbf{x}_i - \frac{\exp(\hat{\boldsymbol{\beta}}_c^T \mathbf{x}_i)}{1 + \exp(\hat{\boldsymbol{\beta}}_c^T \mathbf{x}_i)} \mathbf{x}_i \right\} \\ &= \sum_{i \in S_{2, \geq c}} (y_{ci} - \hat{q}_{ci}) \mathbf{x}_i \quad (c = 1, \dots, C - 1) \end{aligned}$$

where the last equality follows from Equation (6). Clearly, this is equivalent to fitting each binomial logistic regression model in Equation (5) separately to the associated subsample $S_{2, \geq c}$. Furthermore, it follows from standard inference theory for maximum likelihood that, asymptotically, for any pair of vectors $\hat{\boldsymbol{\beta}}_c$ and $\hat{\boldsymbol{\beta}}_d$ ($c \neq d$),

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}_c \\ \hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \{X^T \Delta_{\geq c} X\}^{-1} & 0 \\ 0 & \{X^T \Delta_{\geq d} X\}^{-1} \end{pmatrix} \right);$$

here, the matrix X contains rows \mathbf{x}_i^T for each unit and $\Delta_{\geq c}$ denotes an associated diagonal matrix with $(\Delta_{\geq 1})_{ii} = q_{1i}(1 - q_{1i})$ and, for $c > 1$,

$$(\Delta_{\geq c})_{ii} = \begin{cases} q_{ci}(1 - q_{ci}) & \text{if } y_{1i} = \dots = y_{(c-1)i} = 0, \\ 0 & \text{otherwise.} \end{cases}$$

For a more detailed derivation, see [Scholtus \(2018\)](#).

In a finite-population setting, instead of the above approach pseudo maximum likelihood estimation can be used ([Chambers and Skinner 2003](#) chap. 2). The original model is then interpreted as a superpopulation model from which the target population has been generated. The parameters of interest (say, $\boldsymbol{\beta}_{U_2, c}$) are now implicitly defined by the likelihood equations that would be obtained if y_{ci} were observed for the entire population

U_2 :

$$\begin{aligned}
 0 &= \sum_{i \in U_{2, \geq c}} \left\{ y_{ci} \mathbf{x}_i - \frac{\exp(\boldsymbol{\beta}_{U_{2,c}}^T \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}_{U_{2,c}}^T \mathbf{x}_i)} \mathbf{x}_i \right\} \\
 &= \sum_{i \in U_{2, \geq c}} (y_{ci} - q_{U_{2,ci}}) \mathbf{x}_i \quad (c = 1, \dots, C - 1),
 \end{aligned}$$

where $U_{2, \geq c}$ is defined analogously to $S_{2, \geq c}$. Note that here we ignore the register part of the population U_1 as there is no need to estimate an imputation model for these units.

To account for the sample design of S_2 , the parameters $\boldsymbol{\beta}_{U_{2,c}}$ are estimated from a weighted version of the likelihood equations:

$$\begin{aligned}
 0 &= \sum_{i \in S_{2, \geq c}} w_{2i} \left\{ y_{ci} \mathbf{x}_i - \frac{\exp(\hat{\boldsymbol{\beta}}_c^T \mathbf{x}_i)}{1 + \exp(\hat{\boldsymbol{\beta}}_c^T \mathbf{x}_i)} \mathbf{x}_i \right\} \tag{8} \\
 &= \sum_{i \in S_{2, \geq c}} w_{2i} (y_{ci} - \hat{q}_{ci}) \mathbf{x}_i \quad (c = 1, \dots, C - 1),
 \end{aligned}$$

with $w_{2i} = 1/\pi_{2i}$ the design weight of unit $i \in S_2$ based on its inclusion probability π_{2i} . (Note: if S_2 is actually a subsample of a larger sample S from U , as discussed in Section 2, then these inclusion probabilities should also reflect the subsampling step.) This provides consistent estimates of $\boldsymbol{\beta}_{U_{2,c}}$ (and $\boldsymbol{\beta}_c$).

It follows from a Taylor series linearisation (Chambers and Skinner 2003, 23) that the large-sample variance of the estimated parameters with respect to the finite population can be estimated by the following 'sandwich' estimator:

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}_{U_{2,c}}) = \left(X^T \hat{\Delta}_{\geq c, w_2} X \right)^{-1} \hat{\Gamma}_c \left(X^T \hat{\Delta}_{\geq c, w_2} X \right)^{-1}. \tag{9}$$

Here, $\hat{\Delta}_{\geq c, w_2}$ is a diagonal matrix with elements $(\hat{\Delta}_{\geq 1, w_2})_{ii} = w_{2i} \hat{q}_{1i} (1 - \hat{q}_{1i})$ and, for $c > 1$,

$$(\hat{\Delta}_{\geq c, w_2})_{ii} = \begin{cases} w_{2i} \hat{q}_{ci} (1 - \hat{q}_{ci}) & \text{if } y_{1i} = \dots = y_{(c-1)i} = 0 \\ 0 & \text{otherwise,} \end{cases}$$

while $\hat{\Gamma}_c$ is a design-based estimator of $\text{var}(\sum_{i \in S_{2, \geq c}} w_{2i} (y_{ci} - q_{U_{2,ci}}) \mathbf{x}_i)$. Since the latter expression is an ordinary Horvitz-Thompson estimator, we can use (Särndal et al. 1992, 48):

$$\hat{\Gamma}_c = \sum_{i \in S_{2, \geq c}} \sum_{j \in S_{2, \geq c}} \frac{\pi_{2ij} - \pi_{2i} \pi_{2j}}{\pi_{2ij}} \mathbf{x}_i (y_{ci} - \hat{q}_{ci}) (y_{cj} - \hat{q}_{cj}) \mathbf{x}_j^T,$$

where π_{2ij} denotes a second-order inclusion probability of S_2 . Finally, by a similar argument it can be shown that $\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}_{U_{2,c}}$ and $\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_{U_{2,d}}$ ($c \neq d$) are still asymptotically independent.

8.2.2. Conditional Probabilities: Large-Sample Covariances

Given the results in Subsubsection 8.2.1, large-sample variances and covariances of the predicted conditional probabilities \hat{q}_{ci} can be derived in a straightforward manner from a first-order Taylor series approximation to Equation (6). For pseudo maximum likelihood estimation it follows that, asymptotically, $\text{cov}(\hat{q}_{ci}, \hat{q}_{dj}) \approx 0$ for $c \neq d$, while $\text{cov}(q_{ci}, q_{cj})$ is estimated consistently by:

$$\widehat{\text{cov}}(\hat{q}_{ci}, \hat{q}_{cj}) = \hat{q}_{ci}(1 - \hat{q}_{ci})\hat{q}_{cj}(1 - \hat{q}_{cj})\mathbf{x}_i^T \left(X^T \hat{\Delta}_{\geq c, w_2} X \right)^{-1} \hat{\Gamma}_c \left(X^T \hat{\Delta}_{\geq c, w_2} X \right)^{-1} \mathbf{x}_j. \quad (10)$$

In particular, it follows that predicted conditional probabilities for different levels of educational attainment are asymptotically independent. We refer to Scholtus (2018) for more details on the derivation.

8.2.3. Marginal Probabilities: Large-Sample Covariances

Before proceeding to the marginal probabilities \hat{p}_{ci} , we introduce the following useful lemmas.

- **Lemma 1.** For any set of four random variables A, B, C and D such that
 - A is independent of B and D , and
 - C is independent of B and D ,

it holds that $\text{cov}(AB, CD) = E(AC)\text{cov}(B, D) + E(B)E(D)\text{cov}(A, C)$.

- **Lemma 2.** For any set of three random variables X, Y and Z such that X is independent of Y and Z , it holds that $\text{cov}(XY, Z) = E(X)\text{cov}(Y, Z)$.

Lemma 1 may be derived by conditioning on A and C . Lemma 2 follows as a corollary of the first lemma by choosing $A = X, B = Y, C = 1$ and $D = Z$.

Denote $\text{cov}(\hat{p}_{ci}, \hat{p}_{dj}) = C_{cdij}$, for $i, j \in U_2$ and $1 \leq c, d \leq C$. To evaluate the second component of variance formula (3), we only need the 'diagonal' terms C_{ccij} . However, to derive expressions for these diagonal terms it will be seen below that we also need to consider the terms C_{cdij} with $c \neq d$.

Since $\hat{p}_{1i} = \hat{q}_{1i}$, $C_{11ij} = \text{cov}(\hat{q}_{1i}, \hat{q}_{1j})$ can be estimated by Equation (10). For $2 \leq c \leq C - 1$, using Equation (7) we can write:

$$C_{ccij} = \text{cov} \left\{ \hat{q}_{ci} \left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki} \right), \hat{q}_{cj} \left(1 - \sum_{l=1}^{c-1} \hat{p}_{lj} \right) \right\}.$$

According to Equation (7), each \hat{p}_{ki} is constructed from just the probabilities $\hat{q}_{1i}, \dots, \hat{q}_{ki}$. Since the conditional probabilities $\hat{q}_{1i}, \dots, \hat{q}_{(c-1)i}$ are known to be asymptotically mutually independent, it follows that \hat{q}_{ci} is asymptotically independent of all $\hat{p}_{1i}, \dots, \hat{p}_{(c-1)i}$ and all $\hat{p}_{1j}, \dots, \hat{p}_{(c-1)j}$ for $j \neq i$. Therefore, we can (asymptotically) apply Lemma 1 to the above expression for C_{ccij} , with $A = \hat{q}_{ci}$,

$B = 1 - \sum_{k=1}^{c-1} \hat{p}_{ki}$, $C = \hat{q}_{cj}$ and $D = 1 - \sum_{l=1}^{c-1} \hat{p}_{lj}$. This yields:

$$C_{ccij} \approx E(\hat{q}_{ci}\hat{q}_{cj}) \operatorname{cov} \left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki}, 1 - \sum_{l=1}^{c-1} \hat{p}_{lj} \right) + E \left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki} \right) E \left(1 - \sum_{l=1}^{c-1} \hat{p}_{lj} \right) \operatorname{cov} (\hat{q}_{ci}, \hat{q}_{cj}),$$

and therefore (for $c = 2, \dots, C - 1$)

$$C_{ccij} \approx \{ \operatorname{cov} (\hat{q}_{ci}, \hat{q}_{cj}) + q_{ci}q_{cj} \} T_{c-1,ij} + \left(1 - \sum_{k=1}^{c-1} p_{ki} \right) \left(1 - \sum_{l=1}^{c-1} p_{lj} \right) \operatorname{cov} (\hat{q}_{ci}, \hat{q}_{cj}), \tag{11}$$

with the short-hand notation

$$T_{c,ij} = \sum_{k=1}^c \sum_{l=1}^c C_{kl ij}, \quad (c = 1, \dots, C - 1)$$

(Note: in a finite-population context, in Equation (11) q_{ci} and p_{ci} should be replaced by $q_{U_2,ci}$ and the associated marginal probabilities $p_{U_2,ci}$. In what follows we will ignore this distinction as it is not essential to the argument). For the remaining case $c = C$, it follows directly from Equation (7) that

$$C_{CCij} = \operatorname{cov} \left(1 - \sum_{k=1}^{C-1} \hat{p}_{ki}, 1 - \sum_{l=1}^{C-1} \hat{p}_{lj} \right) = \sum_{k=1}^{C-1} \sum_{l=1}^{C-1} \operatorname{cov} (\hat{p}_{ki}, \hat{p}_{lj}) = T_{C-1,ij}. \tag{12}$$

It remains to find an expression for $T_{c,ij}$. The following recursion is derived at the end of this subsection: for $c = 1, \dots, C - 1$ it holds asymptotically that

$$T_{c,ij} \approx \sum_{k=1}^c C_{kkij} - \sum_{k=2}^c (q_{ki} + q_{kj}) T_{k-1,ij}, \tag{13}$$

with the convention that the second sum is zero for $c = 1$.

We now have all the required ingredients to estimate all terms $\operatorname{cov} (\hat{p}_{ci}, \hat{p}_{cj}) = C_{ccij}$ that occur in Equation (3), at least for large samples. The following algorithm can be used:

1. Estimate $\operatorname{cov} (\hat{q}_{ci}, \hat{q}_{cj})$ by (10) for all $c = 1, \dots, C - 1$ and define $\hat{C}_{11ij} = \hat{T}_{1,ij} = \widehat{\operatorname{cov}} (\hat{q}_{1i}, \hat{q}_{1j})$.
2. Repeat the following steps for $c = 2, \dots, C - 1$:
 - a. Estimate C_{ccij} in line with (11) by

$$\hat{C}_{ccij} = \{ \widehat{\operatorname{cov}} (\hat{q}_{ci}, \hat{q}_{cj}) + \hat{q}_{ci}\hat{q}_{cj} \} \hat{T}_{c-1,ij} + \left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki} \right) \left(1 - \sum_{l=1}^{c-1} \hat{p}_{lj} \right) \widehat{\operatorname{cov}} (\hat{q}_{ci}, \hat{q}_{cj}).$$

b. Estimate $T_{c,ij}$ in line with Equation (13) by

$$\hat{T}_{c,ij} = \sum_{k=1}^c \hat{C}_{kkij} - \sum_{k=2}^c (\hat{q}_{ki} + \hat{q}_{kj}) \hat{T}_{k-1,ij}.$$

3. Finally, define $\hat{C}_{CCij} = \hat{T}_{C-1,ij}$ in line with Equation (12).

An illustration of the first steps of this algorithm is given in Scholtus (2018). Note that the algorithm manages to avoid a circular argument, because C_{ccij} is estimated in Step 2a using $\hat{T}_{c-1,ij}$ and $T_{c,ij}$ is estimated in Step 2b using only the estimated covariances $\hat{C}_{11ij}, \dots, \hat{C}_{ccij}$.

We conclude this subsection by deriving expression (13). By definition, the expression holds exactly for $c = 1$. Therefore, suppose that $2 \leq c \leq C - 1$. We begin by evaluating the ‘off-diagonal’ terms C_{cdij} with $c \neq d$. First suppose that $d < c$. We can write:

$$C_{cdij} = \text{cov} \left\{ \hat{q}_{ci} \left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki} \right), \hat{p}_{dj} \right\}.$$

Since \hat{q}_{ci} is asymptotically independent of all $\hat{p}_{1i}, \dots, \hat{p}_{(c-1)i}$ and also of \hat{p}_{dj} , Lemma 2 can be applied to this expression, with $X = \hat{q}_{ci}$, $Y = 1 - \sum_{k=1}^{c-1} \hat{p}_{ki}$ and $Z = \hat{p}_{dj}$. This yields:

$$C_{cdij} \approx E(\hat{q}_{ci}) \text{cov} \left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki}, \hat{p}_{dj} \right) = -q_{ci} \sum_{k=1}^{c-1} C_{kdij}, \quad (d < c)$$

Similarly, we obtain for $d < c$ that $C_{dcij} \approx -q_{cj} \sum_{l=1}^{c-1} C_{dlj}$.

These expressions can be substituted in the definition of $T_{c,ij}$, to find:

$$\begin{aligned} T_{c,ij} &= \sum_{k=1}^c \left\{ C_{kkij} + \sum_{k=1}^{l-1} C_{klj} + \sum_{l=k+1}^c C_{klj} \right\} \\ &\approx \sum_{k=1}^c \left\{ C_{kkij} - \sum_{l=1}^{k-1} \left(q_{ki} \sum_{m=1}^{k-1} C_{mlj} \right) - \sum_{l=k+1}^c \left(q_{lj} \sum_{m=1}^{l-1} C_{kmij} \right) \right\} \\ &= \sum_{k=1}^c C_{kkij} - \sum_{k=2}^c \left(q_{ki} \sum_{l=1}^{k-1} \sum_{m=1}^{k-1} C_{mlj} \right) - \sum_{l=2}^c \left(q_{lj} \sum_{k=1}^{l-1} \sum_{m=1}^{l-1} C_{kmij} \right) \\ &= \sum_{k=1}^c C_{kkij} - \sum_{k=2}^c q_{ki} T_{k-1,ij} - \sum_{l=2}^c q_{lj} T_{l-1,ij}, \end{aligned}$$

from which Equation (13) follows. In the third line, we used that the middle term is empty (hence zero) for $k = 1$ and we re-arranged the summation over k and l in the right-most term.

8.2.4. Marginal Probabilities: Equation (2)

Suppose first that the imputation model satisfies the following Assumptions:

1. The model includes the cross-classified variable h as a predictor.
2. The model includes all variables that determine the sampling design of S_2 as predictors.
3. All predictor variables are categorical (coded as dummy variables) and the model contains all higher-order interactions of these predictor variables.

The following lemma establishes in particular that, under these Assumptions, Equation (2) is satisfied exactly by the predicted marginal probabilities \hat{p}_{ci} from the continuation-ratio model.

Lemma 3. *Under Assumptions 2 and 3 it holds that $\sum_{i \in S_2} \mathbf{x}_i \hat{p}_{ci} = \sum_{i \in S_2} \mathbf{x}_i y_{ci}$ for $c = 1, \dots, C$.*

Proof. First, it follows from Equation (8) that in this case, for all $c = 1, \dots, C - 1$,

$$\sum_{i \in S_{2, \geq c}} \mathbf{x}_i \hat{q}_{ci} = \sum_{i \in S_{2, \geq c}} w_{2i}(\mathbf{x}_i \pi_{2i}) \hat{q}_{ci} = \sum_{i \in S_{2, \geq c}} w_{2i}(\mathbf{x}_i \pi_{2i}) y_{ci} = \sum_{i \in S_{2, \geq c}} \mathbf{x}_i y_{ci} = \sum_{i \in S_2} \mathbf{x}_i y_{ci}$$

The second equality follows because, under Assumptions 2 and 3, $\mathbf{x}_i \pi_{2i}$ is contained in the span of \mathbf{x}_i ; the last equality follows because $y_{ci} = 0$ for all $i \in S_2 \setminus S_{2, \geq c}$. In particular, this establishes that $\sum_{i \in S_2} \mathbf{x}_i \hat{p}_{1i} = \sum_{i \in S_2} \mathbf{x}_i y_{1i}$ since $\hat{p}_{1i} = \hat{q}_{1i}$.

Now suppose that $\sum_{i \in S_2} \mathbf{x}_i \hat{p}_{ki} = \sum_{i \in S_2} \mathbf{x}_i y_{ki}$ has been shown to hold for all $k = 1, \dots, c - 1$. For \hat{p}_{ci} (with $c = 2, \dots, C - 1$) we find from Equation (7):

$$\begin{aligned} \sum_{i \in S_2} \mathbf{x}_i \hat{p}_{ci} &= \sum_{i \in S_2} \mathbf{x}_i \hat{q}_{ci} \left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki} \right) = \sum_{i \in S_2} \mathbf{x}_i \hat{q}_{ci} \left(1 - \sum_{k=1}^{c-1} y_{ki} + \sum_{k=1}^{c-1} y_{ki} - \sum_{k=1}^{c-1} \hat{p}_{ki} \right) \\ &= \sum_{i \in S_{2, \geq c}} \mathbf{x}_i \hat{q}_{ci} - \sum_{i \in S_2} \mathbf{x}_i \hat{q}_{ci} \left\{ \sum_{k=1}^{c-1} (\hat{p}_{ki} - y_{ki}) \right\} \\ &= \sum_{i \in S_2} \mathbf{x}_i y_{ci} - \sum_{k=1}^{c-1} \sum_{i \in S_2} \mathbf{x}_i \hat{q}_{ci} (\hat{p}_{ki} - y_{ki}). \end{aligned}$$

Thus, to establish the result for \hat{p}_{ci} , it suffices to show that $\sum_{i \in S_2} \mathbf{x}_i \hat{q}_{ci} (\hat{p}_{ki} - y_{ki}) = 0$ for all $k = 1, \dots, c - 1$. Consider an infinite Taylor series expansion of \hat{q}_{ci} from Equation (6), in terms of $\hat{\beta}_c^T \mathbf{x}_i$ around 0. Each term in this series contains a product of one or more elements of the vector \mathbf{x}_i . By assumption 3 above, all of these products are contained in \mathbf{x}_i itself. One further application of assumption 3 yields that $\mathbf{x}_i \hat{q}_{ci}$ is also contained in the span of \mathbf{x}_i . Hence, it follows by induction that $\sum_{i \in S_2} \mathbf{x}_i \hat{q}_{ci} (\hat{p}_{ki} - y_{ki}) = 0$ for all $k = 1, \dots, c - 1$ and that $\sum_{i \in S_2} \mathbf{x}_i \hat{p}_{ci} = \sum_{i \in S_2} \mathbf{x}_i y_{ci}$.

Finally, to complete the proof of Lemma 3, for $c = C$ we find from Equation (7) that

$$\sum_{i \in S_2} \mathbf{x}_i \hat{p}_{Ci} = \sum_{i \in S_2} \mathbf{x}_i \left(1 - \sum_{k=1}^{C-1} \hat{p}_{ki} \right) = \sum_{i \in S_2} \mathbf{x}_i \left(1 - \sum_{k=1}^{C-1} y_{ki} \right) = \sum_{i \in S_2} \mathbf{x}_i y_{Ci}$$

In practice, Equation (3) will often not be satisfied. Higher-order interaction terms may be excluded from the imputation model if they are not significant, and there may also be

predictor variables on a continuous scale. Consider the following weaker version of Equation (3):

3*) The model includes the interaction of h and the variables of the sampling design.

We argue that, for large samples and provided that Assumptions 2 and 3* are satisfied, Equation (2) should still hold approximately.

Namely, under Assumptions 2 and 3* it follows analogously to Equation (8) that $\sum_{i \in S_2} h_i \hat{p}_{ci}^{mlr} = \sum_{i \in S_2} h_i y_{ci}$ for predicted probabilities \hat{p}_{ci}^{mlr} based on an ordinary multinomial logistic regression model. Furthermore, Agresti (2013, 301) notes that, compared to this multinomial logistic regression model, the main benefits of using a specialised model such as continuation-ratio that reflects the ordinal aspect of the target variable are "improved model parsimony and power" in small samples. If the sample is large enough, the ordinary multinomial logistic regression model should be able to capture the ordinal aspect nearly as well. This suggests that, asymptotically, it should hold that $\sum_{i \in S_2} h_i \hat{p}_{ci} / \sum_{i \in S_2} h_i \hat{p}_{ci}^{mlr} \rightarrow 1$. Hence, when Equation (2) holds exactly in the multinomial logistic regression model, it should also hold approximately in the continuation-ratio model.

9. References

- Agresti, A. 2013. *Categorical Data Analysis* (Third Edition). New York: John Wiley and Sons.
- Bakker, B.F.M. 2011. "Micro-integration: State of the Art." In *ESSnet on Data Integration, Report on WPI: 77–107*. Available at: http://ec.europa.eu/eurostat/cros/content/essnet-di-final-report-wp1_en (accessed October 2020).
- Bethlehem, J. 2008. "Surveys without Questions." In *International Handbook of Survey Methodology*, edited by E.D. de Leeuw, J.J. Hox, and D.A. Dillman: 500–511. New York: Psychology Press.
- Bethlehem, J. 2009. *Applied Survey Methods: A Statistical Perspective*. Hoboken, NJ: John Wiley and Sons.
- Booth, J.G., R.W. Butler, and P. Hall. 1994. "Bootstrap Methods for Finite Populations." *Journal of the American Statistical Association* 89: 1282–1289. DOI: <http://doi.org/10.1080/01621459.1994.10476868>.
- Canty, A.J., and A.C. Davison. 1999. "Resampling-based Variance Estimation for Labour Force Surveys." *The Statistician* 48: 379–391. DOI: <http://doi.org/10.1111/1467-9884.00196>.
- Chambers, R.L., and C.J. Skinner, eds. 2003. *Analysis of Survey Data*. Chichester: John Wiley and Sons.
- Chauvet, G. 2007. *Méthodes de Bootstrap en Population Finie*. PhD Thesis, Rennes: ENSAI. Available at: <http://pastel.archives-ouvertes.fr/tel-00267689/document>. (accessed October 2020).
- Daalmans, J. 2017. *Mass Imputation for Census Estimation*. Discussion Paper, The Hague: Statistics Netherlands. Available at: <http://www.cbs.nl/en-gb/background/2017/11/mass-imputation-for-census-estimation>. (accessed October 2020).

- Daalmans, J. 2018. "Divide-and-Conquer Solutions for Estimating Large Consistent Table Sets." *Statistical Journal of the IAOS* 34: 223–233. DOI: <http://doi.org/10.3233/SJI-170375>.
- De Waal, T., J. Daalmans, and F. Linder. 2018. *Mass Imputation for Census Estimation: Methodology*. Report, The Hague: Statistics Netherlands. Available at: http://ec.europa.eu/eurostat/cros/system/files/admin_wp6_2016_nl.pdf (accessed October 2020).
- De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley and Sons.
- Dowle, M., et al. 2019. *data.table: Extension of data.frame*. R package version 1.12.0. Available at: [http://cran.R-project.org/package = data.table](http://cran.R-project.org/package=data.table). (accessed October 2020).
- Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7: 1–26. DOI: <http://doi.org/10.1214/aos/1176344552>.
- Efron, B., and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap*. London: Chapman & Hall/CRC.
- Gross, S.T. 1980. "Median Estimation in Sample Surveys." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 11–14, 1980: 181–184. Houston, Texas: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/papers/1980_037.pdf (accessed October 2020).
- Kim, J.K., S. Park, Y. Chen, and C. Wu. 2020. "Combining Non-probability and Probability Survey Samples Through Mass Imputation." Unpublished manuscript. Available at: <https://arxiv.org/abs/1812.10694> (accessed October 2020).
- Knottnerus, P., and C. van Duin. 2006. "Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey." *Journal of Official Statistics* 22: 565–584. Available at: <https://www.scb.se/contentassets/ca21efb41fee47d293b-bee5bf7be7fb3/variances-in-repeated-weighting-with-an-application-to-the-dutch-labour-force-survey.pdf> (accessed March 2021).
- Kuijvenhoven, L., and S. Scholtus. 2011. *Bootstrapping Combined Estimators based on Register and Sample Survey Data*. Discussion Paper, The Hague: Statistics Netherlands. Available at: <http://www.cbs.nl/nl-nl/achtergrond/2011/39/bootstrapping-combined-estimator-based-on-register-and-sample-survey-data>. (accessed October 2020).
- Lumley, T. 2018. *survey: Analysis of Complex Survey Samples*. R package version 3.35. Available at: [http://cran.R-project.org/package = survey](http://cran.R-project.org/package=survey). (accessed October 2020).
- Mashreghi, Z., D. Haziza, and C. Léger. 2016. "A Survey of Bootstrap Methods in Finite Population Sampling." *Statistics Surveys* 10: 1–52. DOI: <http://doi.org/10.1214/16-SS113>.
- Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.
- Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Scholtus, S. 2018. *Variances of Census Tables after Mass Imputation of Educational Attainment*. Discussion Paper, The Hague: Statistics Netherlands. Available at: <http://www.cbs.nl/en-gb/background/2018/49/variances-of-census-tables-after-mass-imputation>. (accessed October 2020).
- Schulte Nordholt, E., M. Hartgers, and R. Gircour, eds. 2004. *The Dutch Virtual Census of 2001. Analysis and Methodology*. Voorburg/Heerlen: Statistics Netherlands. Available

- at: <http://www.cbs.nl/en-gb/publication/2005/43/the-dutch-virtual-census-of-2001>. (accessed October 2020).
- Schulte Nordholt, E., J. van Zeijl, and L. Hoeksma (eds.). 2014. *Dutch Census 2011. Analysis and Methodology*. The Hague/Heerlen: Statistics Netherlands. Available at: <http://www.cbs.nl/en-gb/publication/2014/47/dutch-census-2011> (accessed October 2020).
- Valliant, R., A.H. Dorfman, and R.M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley and Sons.

Received May 2019

Revised January 2020

Accepted October 2020

Assessing and Adjusting Bias Due to Mixed-Mode in Aspect of Daily Life Survey

*Claudia De Vitiis*¹, *Alessio Guandalini*¹, *Francesca Inglese*¹, and *Marco D. Terribili*¹

The mixed-mode (MM) designs are adopted by NSIs both to contrast declining response and coverage rates and to reduce the cost of the surveys. However, MM introduces several issues that must be addressed both at the design phase, by defining the best collection instruments to contain the measurement error, and at the estimation phase, by assessing and adjusting the mode effect. In the MM surveys, the mode effect refers to the introduction of bias effects on the estimate of the parameters of interest due to the difference in the selection and measurement errors specific to each mode. The switching of a survey from single to mixed-mode is a delicate operation: the accuracy of the estimates must be ensured in order to preserve their consistency and comparability over time. This work focuses on the methods chosen for the evaluation of the mode effect in the Italian National Institute of Statistics (ISTAT) mixed-mode survey “Aspects of Daily Life – 2017”, in the experimental context for which an independent control single-mode (SM) PAPI sample was planned to assess the introduction of the sequential web/PAPI survey. The presented methods aim to analyze the causes that can determine significant differences in the estimates obtained with the SM and MM surveys.

Key words: Mixed-mode; mode effect; *R*-indicator; instrumental variable approach; propensity score.

1. Introduction

The mixed-mode (MM), that is, the use of different collection techniques in the same survey, is a relatively new approach that ISTAT, as well as other NSIs, is adopting especially for social surveys. Its use is spreading both to contrast declining response rates and to reduce the total cost of the surveys (De Leeuw 2005). The use of different data collection modes, in fact, helps in contacting different types of respondents in the most suitable way for each of them, allowing a gain both in population coverage and response rate. However, it introduces a bias, named *mode effect*, that must be addressed at different levels: in the design phase by defining the best collection instruments to contain the measurement error; in the estimation phase by assessing and treating the bias effects due to the introduction of the mixed-mode data collection, in order to ensure the accuracy of the estimates. The switching of a survey from single to mixed-mode must be planned by aiming to make the MM estimates consistent and comparable with the analogue ones obtained in the previous SM survey editions, to ensure that any changes in the time series are exclusively due to real changes of the observed phenomenon.

¹ Italian National Institute of Statistics (ISTAT), Via Cesare Balbo 16, Rome, 00184, Italy. Emails: devitiis@istat.it, alessio.guandalini@istat.it, fringles@istat.it and terribili@istat.it

Mixed-mode can simultaneously generate nonresponse error (*selection effect*) and measurement error (*measurement effect*). Selection effect occurs when different types of respondents choose different modes to answer. This is not actually a problem, but rather a “desirable” consequence of MM surveys. Measurement effect, on the other hand, refers to the influence of a survey mode on the answers that respondents give, such that one person would give different answers in different modes and is caused by differences in the measurement errors specific to each mode. These errors can be due to differences in interviewer effect and social desirability, primacy and recency effects, recall bias, acquiescence, and so on (De Leeuw 2005; Tourangeau and Yan 2007).

The main problem of MM designs is that the selection and measurement effects are confounded, especially when modes are managed in a sequential way. As a consequence, methods for carrying out the inferential process to disentangle the two effects and obtain unbiased estimates of measurement error are needed. The experimental survey designs, as parallel independent samples or re-interview approaches, can facilitate the estimates of the selection and measurement effects (Buelens et al. 2018).

This article focuses on methods for assessing and adjusting mode effect in a survey setting where an independent single-mode survey is carried out together with a MM survey. The methods chosen are applied to the experimental situation of ISTAT “Multipurpose Survey on Households – Aspects of Daily Life (ADL) – 2017” ISTAT (2019). In the 2017 edition, the mixed-mode approach was used for the first time, as a web mode was added to the traditional PAPI mode in a sequential design. A parallel single-mode (SM) PAPI design was planned to allow for an assessment of mode effect on two independent samples collected with different modes. The goal of the analysis presented is the evaluation of the impact on the final estimates of switching from single to mixed-mode. The field of comparison is a survey that produces several indicators for satisfying both national and European requirements. For this purpose, different components of the total non-sampling error are compared in the two samples, as selection effects due to total nonresponse and measurement errors. The total nonresponse is evaluated for both SM and MM samples through the indicators of the representativeness of response (Schouten et al. 2009; Bethlehem et al. 2011; Schouten et al. 2011) and the response rates, using auxiliary variables.

Selection and measurement effects of some survey variables are estimated on the MM respondents set, considering two different approaches: the instrumental variable approach proposed by Vannieuwenhuyze et al. (2010) in the experimental context where a benchmark single-mode survey is available and the propensity score method, generally used in the observational study (Rosenbaum and Rubin 1983; Vandenplas et al. 2016). Finally, a comparison between estimates obtained using different methods for adjusting selection effect is made on the mixed-mode sample (web-PAPI).

The final aim of this article is to point out methods and analyses useful to ensure the accuracy of a complex inferential process for a MM survey, on the basis of evidences on a real survey.

The article is organized as follows: in Section 2 the survey context is outlined, Section 3 describes the outline of the analyses carried out, Section 4 reports methods and results of the comparison between the single and the MM surveys, Section 5 illustrates methods applied to disentangle selection and measurement effects and results, while Section 6 focuses on the adjustment of mode effects. Lastly, Section 7 outlines some concluding remarks.

2. Survey and Experimental Context

The ADL is carried out each year on a selected sample of about 24,000 households. The sample is spread across nearly 850 Italian municipalities using a two-stage sampling design. The sample of households is selected from the centralized municipal register. In the 2017 edition of the ADL survey, a mixed-mode design was introduced for the first time: a web mode has been added to the traditionally used PAPI mode in a sequential design. In order to analyze the impact of the MM on the estimates of the parameters of interest, a survey design was made in which the sample of each municipality was randomly divided to two sub-samples. In the first sub-sample, of larger size (19,262 households), the mixed web/PAPI mode has been administered sequentially (mixed-mode, MM design), while in the second one (8,014 households), only the PAPI interview has been proposed (single-mode SM, control sample).

The two surveys started simultaneously with an invitation letter sent by ISTAT, about one month before, to invite the households to participate in the survey. In the PAPI single-mode (SM) survey, filling two questionnaires in paper form is required: the basic survey questionnaire administered face-to-face by a municipal interviewer to all individuals in the household and a self-reported questionnaire delivered by the municipal interviewer to each individual of the household. In the sequential MM survey, ISTAT sent an invitation letter to all sample households inviting them to fill the questionnaires in electronic form via the web, using the credentials given by ISTAT in the presentation letter. If the household did not complete the questionnaire on the web, at the end of the period foreseen for online filling (about two months with two reminder letters), a municipal interviewer personally address the same questionnaires to all the individuals.

3. Outline of the Analyses of the Mode Effect

3.1. The Analyses

The analyses carried out on the ADL survey aimed firstly to evaluate the impact of the introduction of MM design on the estimate of the parameters of interest with respect to the single-mode design and, subsequently, to analyze in-depth the reasons that determine significant differences in the estimates obtained with the two designs. For this purpose, the study was developed on several levels of analysis, corresponding to different operational steps. The first level is based on the comparison between the two samples (SM and MM); the second level addresses the evaluation of the components of the mode effect (selection and measurement) of the MM design; and the third level consists of some experiments to adjust for the selection effect in the MM response set.

The analyses were developed thanks to the acquisition of mode insensitive auxiliary variables (socio-demographic) that has been obtained through the linkage of the units of SM and MM samples with administrative data performed through the individual code. The administrative database comes from the Archimede Project, (Integrated archive of economic and demographic microdata, [Garofalo 2014](#); [Ballabio et al. 2018](#)) built for expanding ISTAT information provided by administrative archives to produce longitudinal paths and cross-sectional collections of microdata to be made available to different users. The auxiliary variables acquired on the individuals selected in the samples are: education level, occupation type, tax income, citizenship, geographical area and municipal type.

For some analyses, these structural variables were redefined at the household level, as the household as a whole is involved in the response process and the “choice” of the mode. The household auxiliary variables, defined by aggregating the individual ones, are: higher education level (below/equal/above high school diploma); income class (five quintiles – (1) <EUR 10,508, EUR 10,508–20,281, EUR 20,281–29,778, EUR 29,778–46,079, and EUR >46,079); (2) citizenship (Italian/Foreign household); (3) household type (one-individual under 35, one-individual 35–64, one-individual over 64, two-individuals at least one under 35, two-individuals all over 34, more than two individuals at least one under 25, more than two individuals all of them over 24); (4) municipal type (Metropolitan cities, metropolitan area, other municipalities with <2,000, 2,000–10,000, 10,000–50,000, >50,000 inhabitants); (5) different geographical areas, such as Italian regions (NUTS 2) and macro-areas (NUTS 1: North West, North East, Center, South, Islands).

In the first step of the analysis, tests were performed on the differences in the estimates calculated on SM and MM samples, for a set of relevant survey variables, to highlight the variables suspected of being significantly susceptible to mode effect. Subsequent analyses, based on response rates and some indicators of the representativeness of response, were conducted to study the total nonresponse effect in the two samples, considering a set of the household structural variables. The total nonresponse can introduce a bias effect in the survey estimates and compromise their precision level. While the precision of the estimates is directly related to the response rate, the bias does not only depend on the response rate but also on the difference between respondents and non-respondents. To assess the overall quality of the surveys (especially in terms of the magnitude of the bias), two types of indicators of representative response were used, the global R-indicator and the partial unconditional R-indicator (Schouten et al. 2009; Bethlehem et al. 2011; Schouten et al. 2011). Dissimilarities in both response rates and indicators of representative response could explain, at least partly, differences in the estimates of the survey produced with the SM and MM samples, due to the different composition of SM and MM respondents (selection effects) with respect to a given set of auxiliary variables.

In the second step, the assessment of the mode effect in the MM survey was carried out for some target variables following different methods that make the respondent units to the web and PAPI modes comparable: the propensity score (Rosenbaum and Rubin 1983) is applied using only the MM respondents and a PAPI mode as the reference; the instrumental variable approach proposed by Vannieuwenhuyze (2010) is adopted using the PAPI mode of the SM respondents as a benchmark. The MM estimates can be affected either by the different composition of the web and PAPI respondents or by differences in measurement errors generated from each mode (Hox et al. 2015).

In the third step, some experiments to adjust for MM selection effect were made. In particular, the calibration on fixed proportions of the web and PAPI respondents were applied in order to stabilize the total measurement error over time (Buelens and Van den Brakel 2014). This method was compared with other weighting methods, as propensity score and calibration on demographic known totals (individual structural variables).

In the following scheme, the steps and the methods considered in the study are listed. The structure of “Scheme 1” could be used for (at least) two purposes: (1) to present results from a case study (the ADL-experiment in this case); (2) to describe how different methods can be applied when analyzing a mixed-mode experiment.

Scheme 1. Operational steps of the analysis

	Method	Objective	Assumptions/conditions
First step	1) Tests on the differences in the estimates calculated on the two samples for a set of relevant survey variables	Highlighting the variables for which a suspect of mode effect was significant	Independence between the two samples
	2) Tests on the response rates in the SM and MM sample.	Analysis of the response processes and evaluation of the bias caused by the total non-response	Independence between the two samples; MAR assumption for the response models
	3) Response rates and tests		
	4) Indicators of representativeness		
Second step	5) Instrumental variable approach	Disentangling measurement and selection effects	Representativity assumption
	6) Propensity score		MAR assumption for the response models; Balancing assumption
Third step	7) Weighting methods as propensity score, calibration	To adjust selection effect	Ignorability of selection mechanism; Measurement error negligible
	8) Mode calibration	To stabilize the total measurement error over time	Invariance over time of measurement error

3.2. Analysis Validation: Assessing the Significance of the Estimated Parameters

To properly assess the estimated parameters and effects, their significance must be studied in view of sampling variance. Due to the complexity of the estimated parameters and the complex sampling design adopted in ADL, bootstrapping (Efron 1979; Efron and Tibshirani 1986; Wolter 2007) would seem the most suitable solution. However, despite its widespread use for variance estimation in statistics, there is some controversy when using a method that implies some modelization.

Drawing bootstrap samples entails incorporating additional potential sources of variability when using methods that rely on models estimated on the original sample. In fact, in each bootstrap sample, the model must be estimated, and therefore, variability due to parameter estimation of the model is added. The device suggested by Austin and Small (2014) can be adapted to this context (for instance, when using R -indicators and propensity score). They suggest estimating the model for the whole sample and then selecting bootstrap samples and applying the methods just considering the parameters estimated from the original sample, that is, as the model is given. An alternative method is to use pseudo-replication (Rust and Rao 1996). Each replication is defined by a random weight assigned, according to the sample design, to each unit of the original sample. Also in this case, the model must be estimated just one time and this enables the estimation of the variance without involving additional variance due to the estimation of the model.

The two replication methods provide very similar results. However, for the present purpose the pseudo-replication method has been used and 500 pseudo-replications of the ADL sample, according to the sample design, have been carried out with the aim of properly assessing the estimated parameters and effects.

4. First Step: The Comparison Between the Single and the Mixed-Mode Survey

4.1. Test of Differences Between Estimates

Tests on the differences in the estimates between the SM and MM surveys were performed, through chi-square test, to determine whether there are significant differences between the distributions of the answer with respect to the data collection mode and t-test to determine whether the difference between proportions of individuals for each item is significant with respect to the data collection mode (Martin and Lynn 2011). The hypothesis tests concerned numerous survey indicators. Among these variables, the difference for the distributions of the following variables were found to be significant: life satisfaction; reading books in the last year; use of PC; use of internet; trust in others; frequency of seeing friends. Figure 1 shows the detailed distributions for three variables for which the two independent response sets produced estimates showing the most relevant differences between categories (NR is the nonresponse category).

4.2. Response Rates

The response rates in the SM and MM surveys are calculated on the list of the households sampled, and deleting non-eligible units and “unresolved” units (Eurostat 2014). In the response rate analysis, some household structural variables were used.

To test whether the differences between the response rates in the two independent samples (SM, MM) are statistically significant, a z-test was used, with a confidence level $1-\alpha = 0.99$. In Table 1 the response rates for the two surveys and a Z_c statistic are reported by macro-areas.

The SM and MM surveys achieved different response rates. The response rate was generally slightly higher for the MM survey (74% versus 71% of the SM), while for the web mode the response rate varied greatly between Italian macro-areas (ranging from 36% in the North East to 17% in the South and Islands).

The response rate results were significantly different in the North-West and the North-East. This type of analysis, extended to other structural variables, revealed that the response rates in the two surveys are significantly different when the household presents some characteristics, such as mixed nationality, lower-income class, is composed of just one individual between 35 and 64 years old or of more than two individuals, at least one of them under 26 years old.

Further analyses were conducted to assess whether the response rate distributions are independent of structural variables. The hypothesis of independence between the response and the auxiliary variables was tested through a chi-square test. The results highlighted that these variables influence the response in both surveys, above all the variables that characterize households based on education level, lower income class and mixed-nationality-households, and where the households live, such as northern metropolitan cities.

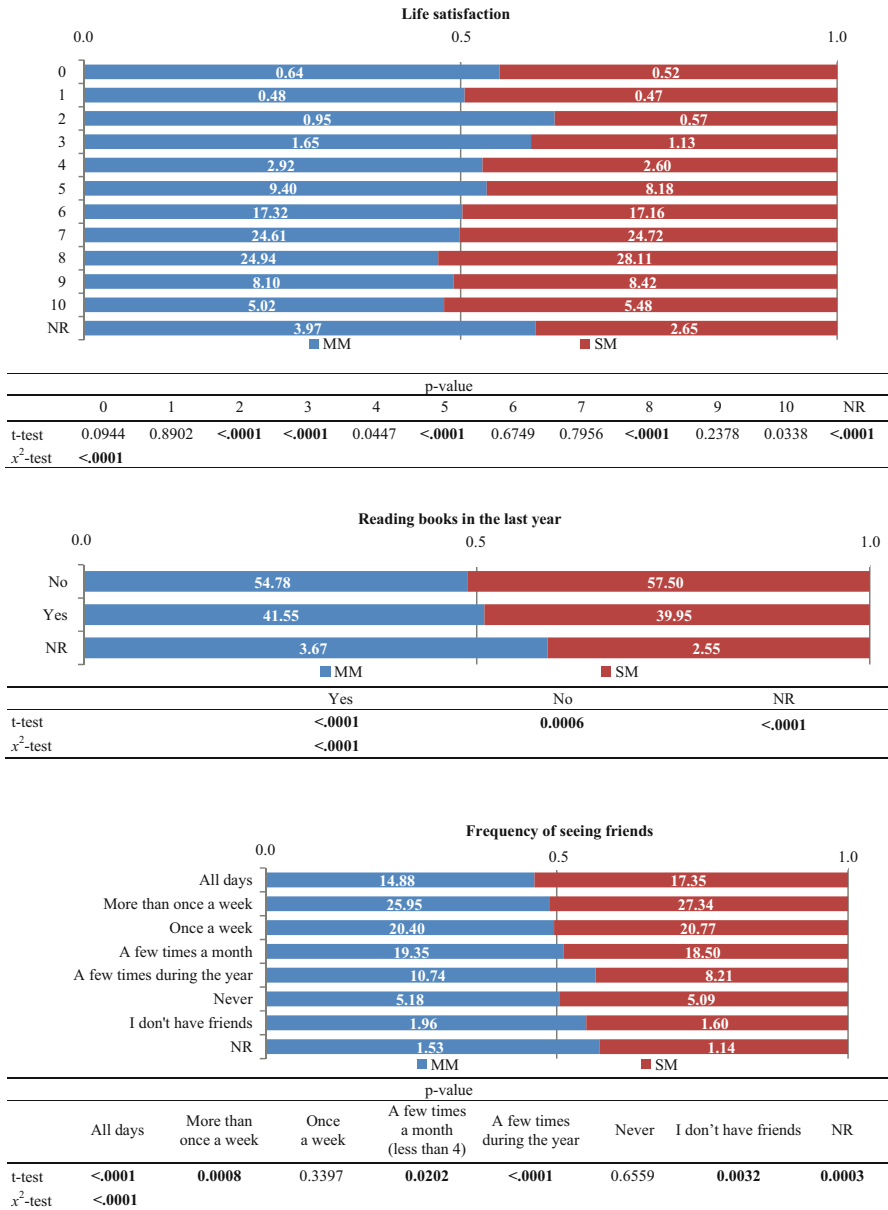


Fig. 1. Distributions of the responses for MM and SM samples for some survey target estimates with p-values of t-test and χ^2 -test on the distributions. NR is the nonresponse category, significant differences at 95% are in bold font.

4.3. Indicators of Representativity

The indicators of representativity, called R-indicators, are based on a measure of the variability of the response propensity, with respect to a given set X of auxiliary variables, known for both respondent and non-respondent units (sample-based auxiliary information). R-indicators may measure deviations from representative response with respect to a vector of available X. The response is representative with respect to some

Table 1. Sample size and response rate in the SM and MM surveys by macro-area.

Macro area	SM		Total MM		Web in MM	
	n	Response rate	n	Response rate	Zc	Response rate
North West	1900	65.9%	4369	71.2%	-4.193*	32.5%
North East	1741	70.2%	4028	73.6%	-2.657*	36.0%
Center	1616	68.6%	3735	70.2%	-1.169	27.8%
South	1781	79.3%	5314	79.4%	-0.090	17.7%
Islands	976	71.3%	1816	74.2%	-1,649	17.3%
ITALY	8014	71.0%	19262	74.0%	-5,091*	26.8%

* The difference between SM and Total MM response rates is statistically significant at a confidence level of 99%.

categorical variable X when the average response probabilities over the classes of X are equal (Schouten et al. 2009).

Schouten et al. (2009) propose two indicators of representativity in the population to compare the quality surveys or the quality of the same survey over time that are based on the standard deviation, $S(\rho_X)$, of the response propensities ρ_X (the response propensity function for the auxiliary variables X): the global R -indicator, $R(\rho_X) = 1 - 2S(\rho_X)$, and the unconditional partial R -indicator, $P_u(Z, \rho_X) = S_B(\rho_X|Z)$. The first indicator describes how the respondent units reflect the target population according to certain characteristics and measure how much the response set in a survey deviates from the representative response. It takes values on the interval $[0, 1]$, with the value 1 indicating the most representative response. The response is not representative if there is much variation in response probabilities, reflected by a large standard deviation.

The unconditional partial R -indicator, instead, can be seen as a measure of the impact of the specified variable on deviations from the representative response (Schouten et al. 2011). This indicator measures the distance to representative response for a single auxiliary variable and is based on the between variance of the response propensities given a stratification with categories of Z : it is non-negative and assumes values less than or equal to 0.5. At the subpopulation (category) level it assumes values between -0.5 and 0.5: a negative value indicates that a subpopulation is under-represented, while a positive value indicates that a subpopulation is over-represented, and the value 0 means that it is represented (Schouten et al. 2011).

The sample-based R -indicators are obtained using a response model (logit or probit regression models) and auxiliary variables known for respondents and nonrespondents are used to estimate the response propensity. Response propensity, ρ_X , is the conditional expectation of R_i (response indicator variable which takes the value 1 if unit i responds and the value 0 otherwise) given the vector of values x_i , of the vector X of auxiliary variables. The estimator of R -indicator, $\hat{R}(\hat{\rho}_X = 1 - 2\hat{S}(\hat{\rho}_X))$, is obtained using the estimator of the variance of the response propensities, $\hat{S}^2(\hat{\rho}_X) = \frac{1}{N-1} \sum d_i (\hat{\rho}_X(x_i) - \hat{\rho}_X)^2$ (d_i is the design weight and $\hat{\rho}_X$ is the weighted sample average of the estimated response propensities). The $\hat{R}(\hat{\rho}_X)$ can be used to derive the standardized absolute maximum bias (Bethlehem et al. 2011), $\hat{B}_m(\hat{\rho}_X) = (1 - \hat{R}(\hat{\rho}_X))/2\hat{\rho}_X$.

The estimator of unconditional partial R -indicator, expressed as $\hat{P}_u(Z, \hat{\rho}_X) = \hat{S}_B(\hat{\rho}_X|Z)$, is obtained using the estimator of the between variance, $\hat{S}_B^2((\hat{\rho}_X|Z)) = \sum_{k=1}^K \frac{N_k}{N} (\hat{\rho}_{X,k} - \hat{\rho}_X)^2$, given a K strata of variable Z that can be included or not in the model

for response. At the level of the subpopulation (strata) $Z = k$, the unconditional partial R -indicator is $\hat{P}_u(Z, k, \hat{\rho}_x) = \hat{S}_B(\hat{\rho}_x|Z = k)$, with the estimator of the variance $\hat{S}_B(\hat{\rho}_x|Z = k) = \sqrt{\frac{\hat{N}_k}{N}(\hat{\rho}_{x,k} - \hat{\rho}_x)}$.

The estimate of response probabilities through the sample introduces additional precision loss in the R -indicator estimates. To evaluate the estimates of the global R -indicator and the unconditional partial R -indicator, the standard error and the confidence intervals must be calculated.

For the considered survey context, the estimates of the indicators of the representative response are obtained starting from different logistic regression models for response, $\log[\rho_x/(1 - \rho_x)] = x'\beta$. The models are defined at the national level (Italy) and for three macro-areas (North, Center, South and Islands), considering as dependent variable the response indicators, $R_{SM,i}$ for the SM sample and $R_{MM,i}$ for the MM sample, and as independent variables the vector of the household structural variables. The covariates of the logit models with a significant regression coefficient are: household type, income class, higher educational level and geographical areas (five macro-areas in the model for the national level and the municipal type in the models for the macro-areas).

Tables 2 to 5 present, for both SM and MM samples, the estimates of $\hat{R}(\hat{\rho}_x)$, the values of the standardized absolute maximum bias proposed by Bethlehem et al. (2011), and the related standard error (SE) with the confidence intervals at 95% (CI 95%). As already discussed in Subsecon 3.2, standard errors are obtained through the pseudo-replication method to take into account the sample design of the survey and correctly estimate the variance of the R -indicators without involving the estimation of the model. Then, confidence intervals are computed using a Gaussian assumption for the sampling variance.

Table 2 shows that, at the national level the MM response set deviates less from the representative response with respect to the SM sample, 0.854 in the first and to 0.812 in the second, confirming a better representativity of the mixed-mode survey. For the North, the values of the R -indicator are similar for the two samples, for the other macro-areas they are very different. The response in these cases is more representative when the MM survey is adopted. Although the web response rates are much lower in the South and Islands, it

Table 2. Global R -indicator estimates for the SM and MM samples in the macro-area and at national level.

Macro-area	SM sample			MM sample		
	Estimate (CI 95%)	SE	Absolute maximum bias	Estimate (CI 95%)	SE	Absolute maximum bias
North	0.847 (0.841, 0.853)	0.00311	0.0516	0.840 (0.832, 0.848)	0.00433	0.0578
Center	0.752 (0.744, 0.760)	0.00428	0.0844	0.842 (0.832, 0.852)	0.00491	0.0572
South and Islands	0.840 (0.833, 0.847)	0.00339	0.0608	0.907 (0.899, 0.915)	0.00429	0.0337
Italy	0.814 (0.810, 0.818)	0.00183	0.0657	0.854 (0.850, 0.858)	0.00214	0.0542

Table 3. Unconditional partial *R*-indicator estimates for the income class and strata.

Variable Strata	SM sample			MM sample		
	Estimate	CI 95%	SE	Estimate	CI 95%	SE
<i>Income class</i>	0.0032	(0.0029, 0.0035)	0.00016	0.0020	(0.0017, 0.0023)	0.00014
< 10508	-0.0449	(-0.0451, -0.0447)	0.00012	-0.0359	(-0.0361, -0.0357)	0.00011
10508-20281	-0.0077	(-0.0077, -0.0077)	0.00001	-0.0016	(-0.0016, -0.0016)	0.00000
20281-29778	0.0142	(0.0142, 0.0142)	0.00002	0.0035	(0.0035, 0.0035)	0.00001
29778-46079	0.0141	(0.0141, 0.0141)	0.00002	0.0123	(0.0123, 0.0123)	0.00002
> 46079	0.0261	(0.0260, 0.0262)	0.00004	0.0238	(0.0237, 0.0239)	0.00003

Table 4. Unconditional partial *R*-indicator estimates for the household typology and strata.

Variable Strata	SM sample			MM sample		
	Estimate	CI 95%	SE	Estimate	CI 95%	SE
<i>Household type</i>	0.0029	(0.0027, 0.0031)	0.00013	0.0019	(0.0017, 0.0021)	0.00008
A	-0.0267	(-0.0268, -0.0266)	0.00007	-0.0276	(-0.0277, -0.0275)	0.00006
B	-0.0328	(-0.0329, -0.0327)	0.00007	-0.0199	(-0.0200, -0.0198)	0.00003
C	-0.0126	(-0.0126, -0.0126)	0.00002	-0.0114	(-0.0114, -0.0114)	0.00001
D	-0.0015	(-0.0015, -0.0015)	0.00001	-0.0080	(-0.0080, -0.0080)	0.00001
E	0.0219	(0.0219, 0.0221)	0.00003	0.0166	(0.0166, 0.0166)	0.00001
F	0.0096	(0.0096, 0.0096)	0.00002	0.0091	(0.0091, 0.0091)	0.00001
G	0.0206	(0.0205, 0.0207)	0.00003	0.0155	(0.0155, 0.0155)	0.00001

seems that in the MM survey the response set better reflects the population of interest with respect to certain characteristics used in the models. At the national level, the upper limit of the bias is higher in the SM survey than in the MM survey. Similar results are shown for the Center and the South and Islands.

The unconditional partial *R*-indicators, $\hat{P}_u(Z, \hat{\rho}_X)$ and $\hat{P}_u(Z, k, \hat{\rho}_X)$, for SM and MM samples are calculated considering the logistic regression model for response defined at the national level (Italy) to evaluate the contribution to the representativeness of the response of the variables “income class”, “household type” and “higher educational level”.

The unconditional partial *R*-indicator for “income class” assumes a higher value in the SM sample rather than in the MM sample: this means that the variable contributes more to the non-representativeness of the response in the SM sample. At subpopulation (strata)

Table 5. Unconditional partial *R*-indicator estimates for the household typology and strata.

Variable Strata	SM sample			MM sample		
	Estimate	CI 95%	SE	Estimate	CI 95%	SE
<i>Higher educational level</i>	0.0002	(−0.0002, 0.0006)	0.00020	0.0003	(−0.0002, 0.0008)	0.00025
A	0.0111	(0.0111, 0.0111)	0.00002	0.0120	(0.0120, 0.0120)	0.00002
B	0.0049	(0.0049, 0.0049)	0.00001	0.0111	(0.0111, 0.0111)	0.00002
C	−0.0034	(−0.0043, −0.0025)	0.00038	−0.0030	(−0.0089, −0.0081)	0.00023

level, it can be noted that for the first and second quintiles of the distribution (lower incomes) there is an under-representation of the two samples, although it is more marked for the respondents of the SM survey. In addition, there is a greater over-representation of households with higher incomes in the SM survey (Table 3).

Table 4 shows the results of the unconditional partial *R*-indicator for “household type” and the following categories: one-individual ≤ 34 (A); one-individual 35–64 (B); one-individual ≥ 65 (C); two-individuals at least one ≤ 35 (D); two-individuals all > 34 (E); more than two individuals at least one ≤ 24 (F); more than two individuals all > 24 (G).

According to the indicator, the variable “household type” contributes more to the non-representativeness of the response in the SM sample (0.00291) than in the MM sample (0.00195). For the subpopulations, it should be noted that for households with one-individual and two individuals aged less than or equal to 35 years (A and D) there is an under-representation of the two response sets, although it is more accentuated for the SM survey except for the last subpopulation. There is still more over-representation for households with two individuals over 34 (E) and households with more than two individuals (F and G) in the SM survey compared to the MM survey (Table 4).

Table 5 refers to the results of the unconditional partial *R*-indicator for the variable “higher educational level” and the categories below high school diploma (A), equal high school diploma (B) and above high school diploma (C). Table 5 shows the contribution to the representative response of the modality “higher educational level”, on the total. It is very low in both response sets. However, it is always higher (in absolute value) in the MM survey than in the SM survey. At the subpopulations level, it should be noted that for households with equal high school diploma (B) there is an over-representation which more accentuated in the MM survey.

5. Second Step: Evaluation of Mode Effect in the MM Sample

5.1. Instrumental Variable Approach

The instrumental variable approach proposed by Vannieuwenhuyze et al. (2010) is useful for the evaluation of both selection and measurement effects in the experimental context with parallel survey designs and a reference mode. The method requires the validity of the comparability and representativity assumptions and the equivalence of the measurement errors between the reference mode and the same mode used in the mixed-mode design.

The instrumental variable approach has been adopted starting from two comparable response sets (SM and MM), in which the SM survey (only PAPI mode) is used as a benchmark. In order to make the SM and MM response sets comparable, that is, to ensure that they are not affected by selection effects due to different response processes, a calibration procedure is adopted separately. That is, the sampling weights are modified to match the distributions of socio-demographic known totals (age class, sex, educational level) at the regional level. Then, on the response set of the SM and MM surveys with adjusted weights, the probability distributions of the survey variable are estimated introducing the instrumental variable I , which identifies the respondents of the SM and the MM surveys (I_{SM} , I_{MM}), and the variable M , that classifies the respondents in the MM survey according to PAPI and web modes (binomial variable that assumes value 1 for PAPI and 0 for web). The instrumental variables I_{SM} and I_{MM} each have a multinomial distribution with p parameter vector: $(\pi_{SM} = \pi_{SM,1}, \dots, \pi_{SM,p})$ where $\pi_{SM,i}$ is the change that $I_{SM} = i$ with $0 \leq \pi_{SM,i} \leq 1$ and $\sum \pi_{SM,i} = 1$ for $(i = 1, \dots, p)$; $(\pi_{MM} = \pi_{MM,1}, \dots, \pi_{MM,p})$ where $\pi_{MM,i}$ is the change that $I_{MM} = i$ with $0 \leq \pi_{MM,i} \leq 1$ and $\sum \pi_{MM,i} = 1$ for $(i = 1, \dots, p)$.

The frequency distributions for I_{SM} and I_{MM} , estimated (calibration weights) from both SM and MM response sets are: the distribution of the PAPI responses from SM, $P(I_{SM})$; the distribution of the PAPI responses from MM, $P(I_{MM}|M = 1)$; the distribution of the web responses from MM, $P(I_{MM}|M = 0)$; the distributions of the PAPI and the web responses from all MM, $P(M = 1)$, $P(M = 0)$.

From these distributions, it is not possible to estimate the selection and measurement effects. In fact, should be necessary to include the not observed distribution $P(I_{SM}|M = 0)$, which represents the counterfactual distribution as if the PAPI respondents in the MM had responded to the web mode (PAPI to web-MM). Starting from the observed distributions, through the total probability formula, $P(I_{SM}) = P(I_{SM}|M = 0) P(M = 0) + P(I_{SM}|M = 1) P(M = 1)$, Vannieuwenhuyze et al. (2010) derive the distribution, $P(I_{SM}|M = 0) = P(I_{SM}) \frac{1}{P(M=0)} - P(I_{SM}|M = 1) \frac{P(M=1)}{P(M=0)}$. From this equation, as $P(I_{SM} = i) = \pi_{SM,i}$, they obtain for each category of I_{SM} , the probabilities $\pi_{SM,i}|M = 0$ for the respondents who chose the web mode.

The estimate of the selection effect is obtained as the difference between the PAPI-MM and the PAPI to web-MM probability distributions, while the estimate of the measurement effect is obtained as the difference between the web-MM and the PAPI to web-MM probability distributions.

The following tables (Table 6 and Table 7) show the results of the application of the instrumental variable approach for some target variables of the ADL survey, such as “reading books in the last 12 months” and “use of PC in the last year”.

Table 6 shows the presence of both selection and measurement effects. The positive selection effect for the No category may be indicative of the fact that the PAPI respondents in the MM sample are more likely to respond No than the web respondents.

Concerning the “Use of PC” (Table 7), for the first category both measurement and selection effects are negative, while with respect to the other categories, they are positive. More individuals respond to other categories when this question is asked in the web mode. The positive selection effect for all categories, except to the first category, indicates that the PAPI respondents in the MM sample are more likely to provide these answers than the web respondents.

Table 6. Probability distributions of reading books in the last 12 months and estimate of the selection and measurement effects.

Category	web-MM	PAPI-MM	PAPI-SM	PAPI to web-MM	Selection effect	Measurement effect
No	0.4086	0.6291	0.5754	0.4813	0.1478	-0.0727
Yes	0.5531	0.3348	0.3989	0.5115	-0.1767	0.0416
Nonresponse	0.0383	0.0361	0.0256	0.0073	0.0288	0.0311

Table 7. Probability distributions of Use of PC and estimate of the selection and measurement effects.

Category	web-MM	PAPI-MM	PAPI-SM	PAPI to web-MM	Selection effect	Measurement effect
Yes, in the last 3 months	0.6430	0.4331	0.5279	0.6945	-0.2614	-0.0515
Yes, from 3 months to 1 year ago	0.0321	0.0265	0.0260	0.0251	0.0014	0.0070
Yes, more than 1 year ago	0.0595	0.0501	0.0484	0.0455	0.0045	0.0140
Never	0.2293	0.4559	0.3734	0.2285	0.2273	0.0008
Nonresponse	0.0360	0.0345	0.0243	0.0063	0.0282	0.0297

5.2. Propensity Score Method

The propensity score (PS) stratification (Rosenbaum and Rubin 1983) can be used to assess the selection and measurement effects. This approach is adopted in observational studies by achieving a balance of covariates between comparison groups. In MM surveys the propensity score can be interpreted as the probability of mode assignment conditional on observed covariates. With adjustments based on PS, the confounding effects of the selection mechanism are mitigated.

The application of this method implied: an estimation of the propensity score model parameters; the definition of subclassification (strata) of respondents based on propensity score; the validation of the balancing assumption, through a chi-square test of the independence between the mode choice and each covariate; for each balanced group, the calculation of weights that equate the weighted proportion of web respondents with the proportion of PAPI respondents in the same stratum.

A logit regression model was used in which the binary response variable is the mode choice web/PAPI is the dependent variable and the structural household variables are the covariates of the model. The covariates of the logit model with a significant regression coefficient are: Italian region, type of municipality, household typology, income class and higher educational level. For eight out of ten of the deciles of the distribution of the predicted probabilities, the independence hypothesis was accepted for all variables.

For each balanced group k , a correction factor, or weight, of the selection effect has been calculated as $w_k = \frac{n_{k,PAPI}/n_{PAPI}}{n_{k,web}/n_{web}}$ (Vandenplas et al. 2016), being $n_{k,T}$ the number of respondents to the mode T (web or PAPI) in the group k . This corrector allows an overall evaluation of the mode effect in the balanced classes: the selection effect is obtained for each target variable, following Vandenplas et al. (2016), as the difference between the web responses adjusted with w_k and the unadjusted web responses, while the measurement effect is obtained as the

difference between the web responses adjusted with w_k and the unadjusted PAPI responses. In [Table 9](#) (see Appendix, Section 8) the majority of the estimated effects are significantly different from zero ($\alpha = 0.05$), based on the bootstrap simulation (see details in Subsection 3.2). What can be underlined from [Table 9](#) is that for many of the items of the selected indicators both the effects are relevant.

For the variable “Reading books in the last 12 months” the results of the estimation for selection and measurement effects obtained through the application of the instrumental approach ([Table 6](#)) and the propensity score ([Table 9](#)) are similar. The two results show the existence of both a selection and a measurement effect in the same direction. However, for the other variable “Use of PC” ([Table 7](#) and [Table 9](#)) the results are not similar. A possible explanation is that this variable is strongly influenced by the propensity of web response: the calibration variables used in the first approach to make the samples comparable could be not sufficiently effective, while the propensity score based on auxiliary variables could be more effective in defining comparable groups of respondent units.

6. Third Step: Adjusting Mode Effect in the MM Sample

6.1. Weighting Methods

In this section, some methods for adjusting mode effect are applied: the weighting methods, as propensity score and calibration, are used to correct the selection effect.

Generally, the weighting methods assume that the selection effect is ignorable and the measurement error due to the mix of techniques is negligible or at least constant over time, so as not to affect the estimates of variation. The assumption of the invariance over time of measurement error in repeated sequential MM surveys is not very sustainable, because the composition of the respondents by mode can change over the years, leading to variations in the total measurement error.

To avoid the misinterpretation of variations in the composition of respondent samples as variations in the estimates, a calibration procedure that takes into account fixed levels of mode proportions is used. This method is proposed by [Buelens and Van den Brakel \(2014\)](#) and aims to keep the measurement error constant over the survey occasions. The calibration procedure simultaneously performs with respect to both auxiliary variables, that correct the selection effect, and to fixed levels of proportions of response by mode, that stabilize the total measurement error.

Alternative weighting adjustment methods (standard calibration and propensity score subclassification adjustment) are implemented and compared with the previous one.

6.2. Results

[Table 8](#) shows the comparison of the estimates of the distribution of “Reading books” that derive from the application of different methods.

These methods are based on calibration procedures that involve distributions of the same socio-demographic totals (age class, sex, educational level) at geographical area level, but that differ for other aspects of the procedure: (1) calibration on only socio-demographics; (2) calibration on socio-demographics and observed fixed levels of mode proportions by six municipal typologies; (3) calibration on socio-demographics and

Table 8. Estimate of the distribution of "Reading books in the last 12 months" with different methods.

Variable	Category	Estimate (%)				
		SM estimate	Meth. 1	Meth. 2	Meth. 3	Meth. 4
Reading books (last 12 months)	No	57.81	59.92	59.00	58.66	59.92
	Yes	39.68	36.51	37.43	37.73	36.33
	Nonresponse	2.49	3.58	3.56	3.61	3.75

hypothesized fixed levels (assuming for future occasions an increase of the web response rate in the South and Islands) of mode proportions by six municipal typologies; and (4) calibration on socio-demographics with sampling weights corrected for the web selection effect through correction factors deriving from propensity score (see Subsection 5.2).

What emerges from the table is that the two calibrations including the constraints with respect to the fixed level of mode proportions (methods 2 and 3) determine a difference in the estimate of about 1%.

7. Concluding Remarks

The set of the analyses presented can be considered as a possible framework of steps, usable by researchers of NSIs to carry out an assessment of mode effect in similar situations. They try to cover all the different approaches applicable in this specific survey context, even though they do not claim to be exhaustive.

The analyses carried out highlight several issues linked to the introduction of the mixed-mode in a social survey. The results show that in MM surveys, the bias due to the total nonresponse is reduced. However, it is still cumbersome to get an overall evaluation of the total measurement error based on different conflicting factors, such as the response process and the mode choice. In fact, it is not easy to understand if the different effects are correctly disentangled and estimated.

The analyses conducted on the experimental context of the ADL survey highlight that the MM design helps to better catch the overall population and all the subpopulations that compose it, than the single-mode design. Anyway, the positive impact of MM in terms of obtaining a less selective response does not necessarily improve the accuracy of the estimates. The quality of some of the produced estimates seems affected by a measurement effect.

The results of the analyses conducted on many target variables in the survey, only partially reported in this article, can provide useful advice in planning a further edition of the ADL survey, in order to benefit from the coverage improvement derived from mixing the techniques. In fact, appropriate actions to prevent measurement errors in the survey design phase can make it possible to apply the calibration procedure proposed by [Buelens and Van den Brakel \(2014\)](#) to control for the estimate changes in the time series due to the variation in the composition of web and PAPI respondents (selection effect).

The application of all the methods presented is subject to two important conditions: (1) the availability and the quality of the auxiliary information, which should be mode insensitive and properly explain the selection effect; (2) the validity of the hypotheses underlying the methods, which must be verified by the researcher as far as possible.

In conclusion, the mode effect analyses in a MM survey imply an underlying effort that is hardly compatible with the usual resources and the timing of a statistical process: such

efforts are challenging and demanding and often there is no time to make them. In general, in order to limit the measurement effect, which is the main drawback of the mixed-mode, in advance as much as possible, accurate planning at the data collection phase is advisable.

8. Appendix

Table 9. Selection and measurement effects estimated through propensity score for target variables. Confidence intervals at 95% in brackets.

Variable	Category	Frequencies			Effects		
		Web	Web adjusted	PAPI	Selection	Measurement	
Reading books (last 12 months)	No	0.451 (0.433, 0.469)	0.485 (0.473, 0.497)	0.618 (0.606, 0.630)	0.034 (0.013, 0.055)	-0.132 (-0.154, -0.110)	
	Yes	0.508 (0.494, 0.522)	0.432 (0.425, 0.439)	0.347 (0.335, 0.359)	-0.075 (-0.095, -0.055)	0.085 (0.070, 0.100)	
	Nonresponse	0.047 (0.040, 0.054)	0.043 (0.042, 0.044)	0.035 (0.029, 0.041)	0.002 (-0.005, 0.009)	0.007 (0.000, 0.014)	
Internet access	No	0.163 (0.148, 0.178)	0.197 (0.190, 0.204)	0.361 (0.347, 0.375)	-0.045 (-0.060, -0.030)	-0.120 (-0.137, -0.103)	
	Yes	0.804 (0.782, 0.826)	0.765 (0.753, 0.777)	0.612 (0.597, 0.627)	0.046 (0.014, 0.078)	0.110 (0.085, 0.135)	
	Nonresponse	0.033 (0.026, 0.040)	0.038 (0.037, 0.039)	0.027 (0.020, 0.034)	-0.008 (-0.015, -0.001)	0.016 (0.009, 0.023)	
Use of personal computer	Yes, in the last 3 months	0.623 (0.606, 0.640)	0.577 (0.568, 0.586)	0.423 (0.410, 0.436)	0.055 (0.030, 0.080)	0.109 (0.091, 0.127)	
	Yes, from 3 months to 1 year ago	0.032 (0.028, 0.036)	0.034 (0.033, 0.035)	0.025 (0.021, 0.029)	-0.002 (-0.006, 0.002)	0.010 (0.006, 0.014)	
	Yes, more than 1 year ago	0.059 (0.054, 0.064)	0.058 (0.057, 0.059)	0.049 (0.044, 0.054)	0.001 (-0.005, 0.007)	0.010 (0.004, 0.016)	

Table 9. Continued

Variable	Category		Frequencies		Effects	
	Web	Web adjusted	PAPI	Selection	Measurement	
Use of internet	Nonresponse	0.037 (0.031, 0.043)	0.041 (0.040, 0.042)	0.032 (0.026, 0.038)	-0.005 (-0.011, 0.001)	0.011 (0.005, 0.017)
	Yes, in the last 3 months	0.698 (680, 0.716)	0.662 (0.651, 0.673)	0.548 (0.536, 0.560)	0.042 (0.015, 0.069)	0.077 (0.056, 0.098)
	Yes, from 3 months and 1 year ago	0.026 (0.022, 0.030)	0.026 (0.026, 0.026)	0.020 (0.016, 0.024)	-0.001 (-0.005, 0.003)	0.008 (0.004, 0.012)
	Yes, more than 1 year ago	0.049 (0.044, 0.054)	0.049 (0.048, 0.050)	0.030 (0.025, 0.035)	0.000 (-0.006, 0.006)	0.018 (0.013, 0.023)
	Never	0.190 (0.177, 0.203)	0.221 (0.214, 0.228)	0.372 (0.362, 0.382)	-0.040 (-0.054, -0.028)	-0.109 (-0.124, -0.094)
Life satisfaction	Nonresponse	0.037 (0.031, 0.043)	0.042 (0.041, 0.043)	0.031 (0.025, 0.037)	-0.006 (-0.013, 0.001)	0.013 (0.006, 0.020)
	0-2	0.022 (0.018, 0.025)	0.023 (0.023, 0.023)	0.019 (0.015, 0.023)	-0.001 (-0.004, 0.002)	0.006 (0.003, 0.009)
	3-5	0.126 (0.116, 0.136)	0.138 (0.135, 0.141)	0.149 (0.139, 0.159)	-0.015 (-0.025, -0.005)	0.009 (-0.001, 0.019)
	6-7	0.417 (0.401, 0.433)	0.417 (0.410, 0.424)	0.423 (0.410, 0.436)	-0.005 (-0.024, 0.014)	0.002 (-0.015, 0.019)
	8-10	0.392 (0.379, 0.405)	0.374 (0.367, 0.381)	0.374 (0.361, 0.387)	0.017 (0.001, 0.033)	-0.024 (-0.039, 0.009)
Trust in others	Nonresponse	0.043 (0.037, 0.049)	0.047 (0.046, 0.048)	0.034 (0.028, 0.040)	-0.006 (-0.012, 0.000)	0.016 (0.010, 0.022)
	In the majority of people	0.244 (0.234, 0.254)	0.223 (0.220, 0.226)	0.171 (0.160, 0.182)	0.022 (0.010, 0.034)	0.025 (0.015, 0.035)
	You have to be careful	0.713 (0.693, 0.733)	0.730 (0.716, 0.744)	0.796 (0.784, 0.808)	-0.026 (-0.052, 0.000)	-0.033 (-0.058, -0.008)
	Nonresponse	0.043 (0.037, 0.049)	0.047 (0.046, 0.048)	0.033 (0.027, 0.039)	-0.006 (-0.012, 0.000)	0.018 (0.012, 0.024)

Table 9. Continued

Variable	Category		Frequencies		Effects	
	Web	Web adjusted	PAPI	Selection	Measurement	
Frequency of seeing friends	Everyday	0.086 (0.077, 0.095)	0.097 (0.093, 0.101)	0.200 (0.192, 0.208)	-0.015 (-0.025, -0.005)	-0.086 (-0.096, -0.076)
	Sometimes a week	0.254 (0.243, 0.265)	0.250 (0.245, 0.255)	0.267 (0.257, 0.277)	0.004 (-0.009, 0.017)	-0.019 (-0.032, -0.006)
Once a week	0.212 (0.201, 0.233)	0.208 (0.204, 0.212)	0.188 (0.178, 0.198)	0.003 (-0.009, 0.015)	0.017 (0.005, 0.029)	
	Sometimes a month	0.219 (0.209, 0.229)	0.211 (0.208, 0.214)	0.174 (0.165, 0.183)	0.009 (-0.003, 0.021)	0.028 (0.017, 0.039)
Sometimes a year	0.146 (0.137, 0.155)	0.143 (0.141, 0.145)	0.085 (0.077, 0.093)	0.002 (-0.007, 0.011)	0.058 (0.049, 0.067)	
	Never	0.044 (0.438, 0.450)	0.048 (0.047, 0.049)	0.056 (0.051, 0.061)	-0.005 (-0.011, 0.001)	-0.006 (-0.012, 0.000)
No friends	0.017 (0.014, 0.020)	0.019 (0.019, 0.019)	0.019 (0.016, 0.022)	-0.003 (-0.006, 0.000)	-0.001 (-0.004, 0.002)	
	Nonresponse	0.022 (0.018, 0.026)	0.024 (0.024, 0.024)	0.011 (0.007, 0.015)	-0.002 (-0.006, 0.002)	0.015 (0.011, 0.019)

9. References

- Austin, P.C., and D.S. Small. 2014. "The use of bootstrapping when using propensity-score matching without replacement: a simulation study." *Statistics in Medicine* 33(24) : 4306–4319. Wiley Online Library. DOI: <https://doi.org/10.1002/sim.6276>.
- Ballabio, S., A. Carra, S. Casacci, D. Ferrazza, F. Verrecchia, A. Vitalini, and L.C. Viviano. 2018. "Local decisions and new guidelines of the Official Statistics." In Proceedings of Q2018 European Conference on Quality in Official Statistics, June 2018, Cracovia, Poland.
- Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of nonresponse in household surveys*. Wiley Series in Survey Methodology. Hoboken, New Jersey: John Wiley & Sons, Inc. DOI: <https://doi.org/10.1002/9780470891056>.
- Buelens, B., and J.A. van den Brakel. 2014. "Measurement error calibration in mixed-mode Sample Surveys." *Sociological methods and Research* 44(3) : 391–426. DOI: <https://doi.org/10.1177/0049124114532444>.
- Buelens, B., J.A. van den Brakel, and B. Schouten. 2018. "Current methodologies to deal with mode effects and mode bias in multi-mode designs." *MIMOD Deliverable 1 – WP2*. Available at: <https://www.istat.it/en/research-activity/international-research-activity/essnet-and-grants> (accessed May 2021).
- De Leeuw, E.D. 2005. "To mix or not to mix data collection modes in surveys." *Journal of Official Statistics* 21(2) : 233–255. Available at: <https://www.scb.se/contentassets/ca21efb41fee47d293bbe5bf7be7fb3/to-mix-or-not-to-mix-data-collection-modes-in-surveys.pdf> (accessed May 2021).
- Efron, B. 1979. "Bootstrap methods: Another look at the jackknife." *The Annals of Statistics* 7(1) : 1–26. DOI: <https://doi.org/10.1214/aos/1176344552>.
- Efron, B., and R. Tibshirani. 1986. "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy." *Statistical Science* 1(1) : 54–75. DOI: <https://doi.org/10.1214/ss/1177013815>.
- Eurostat. 2014. *ESS Guidelines for the Implementation of the ESS Quality and Performance Indicators (QPI)*. Luxembourg ESTAT / D4/LA D (2014). Available at: <https://ec.europa.eu/eurostat/documents/64157/4373903/02-ESS-Quality-and-performance-Indicators-2014.pdf/5c996003-b770-4a7c-9c2f-bf733e6b1f31> (accessed May 2021).
- Garofalo, G. 2014. "Il Progetto ARCHIMEDE obiettivi e risultati sperimentali." *Istat Working Paper*. Available at: <https://www.istat.it/it/files/2014/11/IWP-n.-9-2014.pdf> (accessed February 2015).
- Hox, J., E.D. de Leeuw, and T. Klausch. 2015. "Mixed Mode Research: Issues in Design and Analysis." In *Total Survey Error in Practice*, edited by P.P. Biemer et al.: 511–530. New York: John Wiley & Sons. DOI: <http://dx.doi.org/10.1002/9781119041702.ch23>.
- ISTAT (2019). "Aspetti metodologici dell'indagine. Aspetti della vita quotidiana Periodo di riferimento: anno 2017". Rome: ISTAT. Available at: <https://www.istat.it/microdata/download.php?id=import/fs/pub/wwwarmida/60/2017/01/Nota.pdf> (accessed May 2021).
- Martin, P., and P. Lynn. 2011. "The effects of mixed mode survey designs on simple and complex analyses." *ISER Working Paper Series* 2011–28. Available at: <https://ideas.repec.org/p/ese/iserwp/2011-28.html> (accessed May 2021).

- Rosenbaum, P.R., and D.B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55. DOI: <https://doi.org/10.1093/biomet/70.1.41>.
- Rust, K.F., and J.N.K. Rao. 1996. "Variance estimation for complex surveys using replication techniques." In *Statistical methods in medical research* 5(3) : 283–310. DOI: <https://doi.org/10.1177%2F096228029600500305>.
- Schouten, B.F., F. Cobben, and J. Bethlehem. 2009. "Indicators for the representativeness of survey response." In *Survey Methodology* 35(1) : 101–113. Available at: Indicators for the representativeness of survey response (statcan.gc.ca).
- Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Representativity of Response." *Journal of Official Statistics* 27: 231–253. Available at: <https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/indicators-for-monitoring-and-improving-representativeness-of-response.pdf> (accessed May 2021).
- Tourangeau, R., and T. Yan. 2007. "Sensitive questions in surveys." *Psychological Bulletin* 133(5) : 859–883. DOI: <https://doi.org/10.1037/0033-2909.133.5.859>.
- Vandenplas, C., G. Loosveldt, and J.T.A. Vannieuwenhuyze. 2016. "Assessing the use of mode preference as a covariate for the estimation of measurement effects between modes. A sequential mixed mode experiment." *Methods, data, analyses (mda)* 10(2) : 119–142. DOI: <https://doi.org/10.12758/mda.2016.011>.
- Vannieuwenhuyze, J.T A., G. Loosveldt, and G. Molenberghs. 2010. "A Method for Evaluating Mode Effects in Mixed-mode Surveys." *Public Opinion Quarterly* 74(5) : 1027–1045. DOI: <https://doi.org/10.1093/poq/nfq059>.
- Wolter, K.M. 2007. *Introduction to variance estimation*. Statistical for Social and Behavioral Science. New York: Springer Science & Business Media, LLC.

Received June 2019

Revised February 2020

Accepted October 2020

Measuring the Accuracy of Aggregates Computed from a Statistical Register

Giorgio Alleva¹, Piero Demetrio Falorsi², Francesca Petrarca¹, and Paolo Righi²

The Italian National Statistical Institute (Istat) is currently engaged in a modernization programme that foresees a significant revision of the methods traditionally used for the production of official statistics. The main concept behind this transformation is the use of the Integrated System Statistical Registers, created by a massive integration of administrative archives and survey data. In this article, we focus on how to measure the accuracy of register estimates of a population total from measurements calculated at the unit level. We propose the global mean squared error (GMSE) as a statistical quantity suitable for measuring accuracy in the context of the production of official statistics. It can be defined to explicitly consider the main sources of uncertainty that may affect registers. The article suggests a feasible calculation strategy for the GMSE that allows National Statistical Institutes to build algorithms that can promptly be applied for each user request, thus improving the relevance, transparency and confidence of official statistics. Through a simulation study, we verified the efficacy of the proposed strategy.

Key words: Integration; anticipated variance; linearization; mean square error; total survey error.

1. Background and Scope of the Article

The article focuses on how to measure the accuracy of population totals obtained from register data starting from uncertainty measures computed at the unit level. We consider the context where the users may autonomously define their totals, having direct access to the register microdata. In general, their statistics are unplanned and unpredictable in advance. The estimates of population totals are fundamental for knowing the dimension of quantitative variables or the level of diffusion of qualitative variables in a population. They represent the dominant part of the output that the different users produce from data and the common target parameters of the National Statistical Institute (NSIs). Standard linearization techniques (Särndal et al. 1992; Wolter 1986) allow extending the approach proposed herein for measuring the accuracy of non-linear statistics (such as correlations or regression parameters, and so on of a distribution) computable from the register microdata.

The background of the research activity described in this article is the modernization programme that the Italian National Statistical Institute (Istat 2016) launched some years

¹ Sapienza University of Rome, Via del Castro Laurenziano 9, 00161 Rome, Italy. Emails: giorgio.alleva@uniroma1.it and francesca.petrarca@uniroma1.it

² Italian National Institute of Statistics (Istat), Via Cesare Balbo, 16 – 00184 Rome, Italy. Emails: piero.falorsi@gmail.com and parighi@istat.it

ago. The main concept underlying this transformation is the use of the Integrated System of Statistical Registers (ISSR) as the basis for the production of all official statistics. This transformation represents a strategic challenge: it proposes abandoning the paradigm of statistical inference based on sample surveys that has been used for the past 75 years and moving on to a mixed data source paradigm for the future (Citro 2014; Alleva 2017). The ISSR is the result of a massive integration of administrative archives with survey data.

According to the statistical quality framework followed by Statistics Canada (2009), also reported by Wallgren and Wallgren (2014), the term survey includes the following components: (1) a census, which attempts to collect data from all members of a population; (2) a sample survey, in which data are collected from a (usually random) sample of population members; (3) a collection of data from administrative records, in which data are derived from 2 records originally kept for non-statistical purposes; and (4) a derived statistical activity, in which data are estimated, modelled, or otherwise derived from existing statistical data sources. Each of the previous components introduces different sources of uncertainty that should be considered both for predicting the target variables at the individual level and for the register aggregate. For instance, component (1) introduces the possibility of coverage errors, which we can address with specific statistical models. Component (2) includes sampling errors, and components (3) and (4) comprise the uncertainty derived by models adopted for building predictions at the individual level.

To construct ISSRs as the single informative infrastructure for the production of official statistics starting from a microdata level, different statistical techniques have been adopted. Many of these techniques result in computing predictions at the unit level. The register values remain the output of statistical processes subject to statistical uncertainty for both units and variables. The main strategic choice is whether to make the use of ISSR limited and to allow the dissemination of only planned outputs with a certified accuracy or to make the system more flexible, allowing different users, to produce their own statistics from the ISSR. We propose here to opt for the second option, which makes the Institute more relevant for its users, but exposes the NSI to the threat of inappropriate use of the register data by unaware users. Indeed, users who have access to the microdata could conceivably produce their estimates fully unaware of any problems associated with the quality of their register statistics. The European Statistical System (ESS) is aware of the importance of producing new measures of accuracy for multi-source statistics such as those produced by statistical registers (Eurostat 2019). In this article, we suggest a computational strategy for facilitating flexible and correct use of register data by enabling users to quickly estimate global mean squared error (GMSE) on their own. In Section 2, we give the notation and introduce the measure of accuracy we propose to adopt for the register aggregates. Then, in Section 3, to facilitate comprehension, we introduce a simplified statistical framework in which the register is not affected by coverage errors. Section 4 describes the calculus of the GMSE for the simplified statistical framework. Section 5 illustrates the main computational challenges. Section 6 deals with coverage errors. Sections 7 and 8 show the first results of a simulation study and provide preliminary conclusions with some initial reflections on how to develop a feasible validation approach. The derivations of the main results are available in Appendix (Section 9) and in Appendices 2, 3 and 4 published as online supplementary materials for this article.

2. Notation and Proposed Measure of Accuracy

Let U be the unknown target population of interest, including $N_{(U)}$ statistical units. Let U_d be a statistical *domain of interest*, which is a subset of U with $N_{(U_d)}$ units. The target parameter of interest, Y_{U_d} , is the total of the variable y within the domain U_d :

$$Y_{U_d} = \sum_{k \in U_d} y_k, \tag{1}$$

where y_k is the true value of the variable y for unit k .

Let R be a statistical register, including $N_{(R)}$ statistical units: ideally, each statistical unit in U should be represented by a *corresponding* unit in R .

Furthermore, let R_d be a subset of R of size $N_{(R_d)}$, which represents the target domain U_d . Let \hat{y}_k be the value recorded in the register that predicts the value y_k . These *predicted* values can be computed according to different statistical models or algorithms. For estimating Y_{U_d} , the users can simply sum the predicted \hat{y}_k values over R_d :

$$\hat{Y}_{R_d} = \sum_{k \in R_d} \hat{y}_k. \tag{2}$$

\hat{Y}_{R_d} is a register-based statistic as in [Wallgren and Wallgren \(2014\)](#) and is the result of an estimation process. We may define the accuracy of \hat{Y}_{R_d} based on the difference between this statistic and the actual value, Y_{R_d} . The accuracy depends on various factors, such as the coverage error of the register and the measurement errors of predictions.

The data structure of the population U and the statistical register, R , are illustrated in [Table 1](#), where the right part represents the population and the left part represents the statistical register. In our table \mathbf{x}_k denotes a vector of l , auxiliary variables available in R for each unit k . Note that the true y_k values are rarely available in the register. The last columns on both parts of the table (the right and the left) are dichotomous membership variables indicating whether the unit is included in domain d . The true values of these

Table 1. Data structure in population U and in statistical register R .

Population U			Statistical register R			
Identifier of the population unit true unknown	True y Value	True membership variable (0,1) of the domain d	Code in R	Predicted value	Auxiliary variables	Register membership variable of the domain d
			1	\hat{y}_1	\mathbf{x}_1	1
1	y_1	1	\vdots	\vdots	\vdots	0
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
k	y_k	0	k	\hat{y}_k	\mathbf{x}_k	1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	1
\vdots	\vdots	\vdots	$N_{(R)}$	$\hat{Y}_{N_{(R)}}$	$\mathbf{x}_{N_{(R)}}$	\vdots
$N_{(U)}$	$y_{N_{(U)}}$	1				

Over-coverage

Under-coverage

variables in the population may coincide with what is reported in the register. The areas with a grey background in the table highlight the over/under-covered units in R .

Depending on the specific objective, the target variable, denoted with the symbol y , and the auxiliary variables, indicated with the symbol x , can be represented by information provided by each of the survey components illustrated in Section 1. To better clarify the different roles, the auxiliary variables are known, and the target variables are those predicted at the unit level. The auxiliary variables are, for example, the sex and age of people in the population register or tax paid by firms in the business register.

In a total survey error approach (Biemer 2010), the mean squared error (MSE) represents the simplest way to measure the accuracy of registered-based statistics. It is expressed as the expected squared difference between the estimator and the true unknown population value:

$$MSE(\hat{Y}_{R_d}) = E(\hat{Y}_{R_d} - Y_{U_d})^2 = V(\hat{Y}_{R_d}) + [Bias(\hat{Y}_{R_d})]^2, \quad (3)$$

where $E(\cdot)$ denotes the operator of expectation and $V(\cdot)$ indicates the operator of the variance. Each specific approach to inference focuses on different sources of variability and bias in the definition of the MSE; these are related to what is treated as fixed or random in the specific inferential approach. For instance, *design-based* (Cochran 1977) or *model-assisted* approaches (Särndal et al. 1992) treat the population values y_k as unknown constants, and the sample selected, with the sample design P , is the only source of randomness; therefore, they develop their expectations considering only the variability of the sampling design. The *model-based* approach (Vaillant 2009; Chambers and Clark 2015) considers the sample as *fixed* and the y_k values as random variables generated according to the model, M ; thus, they develop the MSE considering only the variability embedded in the model. The expectation developed from the model generating the data will be denoted as E_M , and the expectation calculated from the variability of the sampling design will be indicated as E_P . The same notation will be adopted for the operator of variance, thus defining the operators V_M and V_P .

Here, we propose to develop the MSE, taking into consideration all the random components involved in the inferential process for building the predictions. We can simply do this by defining the operator of expectation in Equation (3) as a concatenation of elementary expectation operators, each of which considers a specific random component. Following the proposal of Wolter (1986), who introduces the concept of *global variance*, the measure we propose could be denominated as a *global MSE* (GMSE). The GMSE can be expressed as

$$GMSE(\hat{Y}_{R_d}) = E_P E_M (\hat{Y}_{R_d} - Y_{U_d})^2. \quad (4)$$

When planning the sampling design, GMSE is also known as anticipated variance (Isaki and Fuller 1982; Särndal et al. 1992; Nedyalkova and Tillé 2008; Nirel and Glickman 2009; Falorsi and Righi 2015). Here, the measure is not limited to the sampling context and incorporates additional sources of variability and bias. For instance, the nonresponse by defining GMSE as:

$$GMSE(\hat{Y}_{R_d}) = E_P E_M E_{NR} (\hat{Y}_{R_d} - Y_{U_d})^2, \quad (5)$$

in which E_{NR} indicates the expectation under the models adopted for imputing the nonresponse in survey data. Continuing the illustration, let us consider the case in which we collect the y variable from a census affected by nonresponse. In this case, we can define the GMSE as

$$GMSE(\hat{Y}_{R_d}) = E_{NR}(\hat{Y}_{R_d} - Y_{U_d})^2.$$

The GMSE could be accepted as a measure of precision by the main professional families of methodologists within the NSIs: at least, those who base their inference only on statistical models and those who use the statistical models as a support for inference that continues to be based essentially on sampling design. The global measure has a number of advantageous qualities, including the following: generality, stability over time and robustness in the case of model failures. GMSE is simple to use and communicate to users. It is based on the first and second moments of the random distributions of the specific source of uncertainty. Its calculus does not imply full knowledge of the underlying distributions.

We observe that well-known approaches for estimating GMSE are based on replication methods (Scholtus 2019). However, these techniques are highly time-consuming: the replicates have to produce the whole process generating \hat{Y}_{R_d} . Therefore, in the context of the massive and continuous production of official statistics by NSIs, replication methods do not seem to be a feasible solution.

To facilitate informed and correct use of a registry, once the user defines the target total and the specific domain of interest (R_d), it would be useful to build a dynamic data vector, for example, $\hat{\sigma}_{dy,k}^2$ for $k = 1, \dots, N_{(R_d)}$, so that the GMSE estimate is:

$$GMSE(\hat{Y}_{R_d}) = \sum_{k \in R_d} \hat{\sigma}_{dy,k}^2. \tag{6}$$

The quantities $\hat{\sigma}_{dy,k}^2$ are unit and domain specific. However, we will see in Section 5 that the amount of information to be stored for their calculation is limited for each unpredictable user request. The dependency on the domain is limited to a few useful variables in the register. Therefore, in our proposal, we do not suggest storing the $\hat{\sigma}_{dy,k}^2$ values, but we do recommend stockpiling some intermediate values, not domain specific, from which the $\hat{\sigma}_{dy,k}^2$ may be easily calculated, thus making the proposed solution applicable in the real contexts of NSI informative infrastructures (this will become clearer in Section 5 below). We also note that the definition of the computational formulae for determining quantities $\hat{\sigma}_{dy,k}^2$ represents a relevant result since it enables the NSI to build algorithms that can be applied promptly for each user request even those not planned in advance. To produce the proposed assurance of accuracy at the micro-level (given in Equation 6), we propose a computational strategy based on the following two primary approximations:

1. the linearization of estimator \hat{Y}_{R_d} with respect to each specific source of variability considered in GMSE. The validity of this assumption will be proven as true in the typical asymptotic contexts that are used in these cases (see Appendix A3 in online supplementary materials);

2. the adoption of a form of calculus of the sampling variances based only on first inclusion probabilities.

The computational strategy is made simpler by the well-known result of [Kendall and Stuart \(1976, 196\)](#), according to which we can express the GMSE as a sum of conditional values that is more manageable for the calculus. We repeatedly use this result in Section 4 and in Appendix (Section 9). Finally, we note that although some believe that the use of approximations could weaken the methodological proposal, we find ourselves today in a context where NSIs do not calculate accuracy at all. Furthermore, approximate solutions are usual, for example, in the case of sample variances where closed forms are not available.

3. A Simplified Statistical Framework

Here, we suppose that *coverage error* is negligible, which implies $R_d \equiv U_d$. We also assume that the $\mathbf{x}_k = (x_{k1}, \dots, x_{ki}, \dots, x_{kj})'$ values are not subject to the measurement error. The main sources of variability considered here are the model, M , generating the data y and the sampling design, P . The model and the sampling design are always developed under the non-informative assumption of the current survey sampling activity. The model formulation and fitting are independent of the sampling design and vice versa.

3.1. Model Uncertainty

Regarding the first source, we can suppose that each y_k value is to be expressed as the sum of two components:

$$y_k = \tilde{y}_k + e_k, \quad (7)$$

where $\tilde{y}_k = E_M(y_k)$ is the theoretical value according to which the value of y is generated from a given statistical model, M , for unit k , and e_k denotes the random error with model expectations given by

$$E_M(e_k) = 0, E_M(e_k^2) = V_M(e_k) = \sigma_{y,k}^2, E_M(e_k e_\ell) = \sigma_{y,k,\ell}, \quad (8)$$

where e_ℓ indicates the random error of unit ℓ .

The structure of a model expectation is quite general and may be easily applied to different statistical models. For instance, consider a well-known model with domain random effects, adopted as a small area estimation technique, $y_k = \tilde{y}_k + \varepsilon_k + z_d$ for $k \in U_d$, in which ε_k is random noise and z_d is a random domain effect; we may then reformulate the model expectation structure of this model, according to Equation 8, by defining $e_k = \varepsilon_k + z_d$ for $k \in U_d$.

3.2. Sampling Uncertainty

A generalized framework for defining sampling designs, illustrated in detail in [Falorsi and Righi \(2015\)](#), assumes a sample S of fixed size n selected from R , in accordance with sample design P with inclusion probabilities $\pi = (\pi_1, \dots, \pi_k, \dots, \pi_{N(R)})'$. Many practical sampling designs define domains that are planned sub-populations so that the

sample sizes have been fixed before selecting the sample. Denote by $R_{[h]}(h = 1, \dots, H)$ the planned domain of size $N_{(R_{[h]})} = \sum_{k \in R} d_{k(h)}$ where $d_{k(h)} = 1$ if $k \in R_h$ and $d_{k(h)} = 0$ otherwise. Fixed size sampling designs are those satisfying $\sum_{k \in S} \mathbf{d}_k = \mathbf{n}$, where $\mathbf{d}_k = (d_{k(1)}, \dots, d_{k(h)}, \dots, d_{k(H)})'$ and $\mathbf{n} = (n_1, \dots, n_h, \dots, n_H)'$ is the vector of integer numbers defining the sample sizes fixed at the design stage, with $\sum_{k \in S} d_{k(h)} \pi_k = n$. In our setting, planned domains can overlap; therefore, unit k may have more than one value $d_{k(h)} = 1$ (for $h = 1, \dots, H$). Several customary fixed size sampling designs invite particular consideration. A well-known example is the stratified simple random sampling without replacement (SSRSWOR) design, where strata are the planned domains and each \mathbf{d}_k vector has $H - 1$ elements equal to zero and one element equal to 1, which implies that each unit j belongs to one and only one planned domain. The total Y estimated with the Horvitz-Thompson estimator is $\hat{Y}_{HT} = \sum_{k \in S} y_k (1/\pi_k)$. We suppose that the $N_{(R)} \times H$ matrix $\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_j, \dots, \mathbf{d}_{N_{(R)}})'$ is non-singular. According to this general sampling design framework, Deville and Tillé (2005) proposed an approximated expression of the variance for \hat{Y}_{HT} based on the Poisson sampling theory given by

$$V_P(\hat{Y}_{HT}) \cong \left[\frac{N_{(R)}}{(N_{(R)} - H)} \right] \sum_{k \in R} \left(\frac{1}{\pi_k} - 1 \right) \eta_k^2 \tag{9}$$

$$\text{where } \eta_k = y_k - \pi_k \mathbf{d}'_k \left[\sum_{j \in R} \mathbf{d}_j \mathbf{d}'_j \pi_j (1 - \pi_j) \right]^{-1} \sum_{j \in R} \pi_j \left(\frac{1}{\pi_j} - 1 \right) \mathbf{d}_j y_j. \tag{10}$$

Equation 9 resembles the variance expression of the Horvitz-Thompson estimator under a Poisson sampling design, but it uses the residuals, η_k , instead of the original value, y_k . In practice, when $H = 1$, this is the variance approximation of the conditional Poisson sampling (Deville and Tillé 2005). The above approximation works well when the number of domains H remains small compared to the overall population size $N_{(R)}$. A conservative approximation of Equation (10) may be obtained by substituting η_k with the y_k values.

3.3. Predictions

Let us suppose that \tilde{y}_k can be expressed as a function $f(\cdot)$

$$\tilde{y}_k = f(\mathbf{x}_k; \boldsymbol{\vartheta}), \tag{11}$$

in which $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_i, \dots, \vartheta_I)'$ is a vector of I unknown parameters.

Let

$$\hat{y}_k = f(\mathbf{x}_k; \hat{\boldsymbol{\vartheta}}), \tag{12}$$

be the register predictions, where $\hat{\boldsymbol{\vartheta}} = (\hat{\vartheta}_1, \dots, \hat{\vartheta}_i, \dots, \hat{\vartheta}_I)'$ represents the estimate of $\boldsymbol{\vartheta}$ based on observation of the values y_k on the sample S . The resulting estimator \hat{Y}_{R_d} , given by the Equation 2 belongs to the class of *projection estimators* that can be developed either under the *model-assisted approach* (Kim and Rao 2012) or the *model-based approach* (Chambers and Clark 2015; FAO, part.5, 2014).

Remark. The predictions \hat{y}_k can be built in different ways, thus defining different forms of Equation 2, as follows:

$$\hat{y}_k = \begin{cases} y_k & \text{if } k \in S \cap R \\ f(\mathbf{x}_k; \hat{\boldsymbol{\theta}}) & \text{if } k \in \bar{S} \cap R \end{cases} \quad (a), \quad \hat{y}_k = \begin{cases} y_k & \text{if } k \in S \cap R \\ f(\mathbf{x}_k; \hat{\boldsymbol{\theta}}) + \hat{\varepsilon}_k & \text{if } k \in \bar{S} \cap R \end{cases} \quad (b)$$

$\hat{\varepsilon}_k$ is a residual that can be selected either from the residuals estimated in sample S or from the estimated distribution of the y values (Chen and Haziza 2017). With predictions built as in expression (a), the use of the resulting estimator is more common when using the standard prediction approach for inference. Form (b) is appropriate in cases where register values are used for calculating indicators, such as quantiles or correlations in which the variability at the unit level is relevant. Furthermore, when y is categorical, each specific value of the \hat{y}_k values in R can be set equal to one of the standard modalities of the y variable.

4. The Calculus or the GMSE in the Simplified Statistical Framework

4.1. Decomposition of GMSE

In the observational setting described in Section 3, there are two random vectors of $N_{(R)}$ units: $\mathbf{y} = (y_1, \dots, y_k, \dots, y_{N_{(R)}})'$ and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_k, \dots, \lambda_{N_{(R)}})'$, which is the vector of sample membership indicators with $\lambda_k = 1$ if $k \in S$ and $\lambda_k = 0$ otherwise. We suppose that the estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is model unbiased, which means $E_M(\hat{\boldsymbol{\theta}}|\boldsymbol{\lambda}) = \boldsymbol{\theta}$, where $E_M(\hat{\boldsymbol{\theta}}|\boldsymbol{\lambda})$ denotes the model expectation conditioned on the sample realized value of the vector $\boldsymbol{\lambda}$. Thus, GMSE may be expressed as (see Appendix, Section 9)

$$GMSE(\hat{Y}_{R_d}) = \begin{cases} E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] - V_M(Y_{U_d}) & \text{if } E_P(\hat{Y}_{R_d}|\mathbf{y}) \neq Y_{U_d} \quad (13a) \\ E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] - V_M(Y_{U_d}) - 2Cov_M[E_P(\hat{Y}_{R_d}|\mathbf{y}), Y_{U_d}] & \text{if } E_P(\hat{Y}_{R_d}|\mathbf{y}) \neq Y_{U_d} \quad (13b) \end{cases}$$

where $E_P(\hat{Y}_{R_d}|\mathbf{y})$ denotes the sampling expectation conditioned on the realized value of the vector \mathbf{y} . As seen from Equation (13a), if the estimator \hat{Y}_{R_d} is design unbiased for the aggregate Y_{U_d} , then GMSE will neutralize variability attributing it to the pure model variability of population parameter Y_{U_d} . The conditions for fulfilling design unbiasedness are given in Section 3 of Kim and Rao (2012).

4.2. Calculus of the Dominant Component of GMSE

The dominant component of $GMSE(\hat{Y}_{R_d})$ is $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$. Indeed, the term $V_M(Y_{U_d})$ contributes negatively to the expression Equation (13a), whereas it enters a positive component of the difference $(V_M(Y_{U_d}) - 2Cov_M[E_P(\hat{Y}_{R_d}|\mathbf{y}), Y_{U_d}])$ in the Equation (13b). We also note that the term $-2Cov_M[E_P(\hat{Y}_{R_d}|\mathbf{y}), Y_{U_d}]$ is of the same order of magnitude as the component $V_M(Y_{U_d})$ and represents the model covariance between the two population totals Y_{U_d} and $E_P(\hat{Y}_{R_d}|\mathbf{y})$. The overall difference $(V_M(Y_{U_d}) - 2Cov_M[E_P(\hat{Y}_{R_d}|\mathbf{y}), Y_{U_d}])$ becomes negative when $E_P(\hat{Y}_{R_d}|\mathbf{y}) > \frac{1}{2}Y_{U_d}$ with $Y_{U_d} > 0$. In most of the empirical situations that are encountered, the difference tends to be negligible or negative.

In the following discussion, we will present some asymptotic behaviors of the variables of interest. We refer to the results shown in the literature on the subject (see, for example, Isaki and Fuller 1982; Wolter 1985; Särndal et al. 1992; Deville 1999; Breidt and Opsomer 2017). Here, we limit ourselves to recalling the general framework of our assumptions (as according to Deville 1999): assuming that (1) the size N of the population and the size n of the sample tend to infinity; (2) $N^{-1}Y$ has a finite limit, where Y is the total of the variable y , N is the size of a sequence of populations of increasing size; (3) $N^{-1}(\hat{Y} - Y)$ with \hat{Y} the estimator of the total converges in probability to zero; and (4) $n^{-1/2}N^{-1}(\hat{Y} - Y)$ tends to a multi-dimensional normal distribution, observing the central limit theorem. A consequence of these assumptions is that the terms that are $O_p(n^{-1/2})$ in the decomposition of $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$ can be considered small, and the product of two such small quantities can be deemed negligible.

We focus now on the predominant component $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$. First, we note that the estimate \hat{Y}_{R_d} may be seen as an overall function of three random components, \mathbf{y} , $\boldsymbol{\lambda}$ and $\hat{\mathbf{t}}$, in which the quantities \mathbf{x}_k and $\gamma_{d,k}$ (for $k \in R$) are considered known and not random, with $\gamma_{d,k} = 1$ if $k \in R_d$ and $\gamma_{d,k} = 0$ otherwise. We can express the register-based statistic

$$\hat{Y}_{R_d} = \hat{Y}_{R_d}(\hat{\mathbf{t}}; \mathbf{y}; \boldsymbol{\lambda} | \mathbf{X}, \boldsymbol{\gamma}_d, \boldsymbol{\vartheta})$$

as a function of the three random components, given as the fixed auxiliary variables, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_k, \dots, \mathbf{x}_{N(R)})'$, domain membership variables, $\boldsymbol{\gamma}_d = (\gamma_{d,1}, \dots, \gamma_{d,k}, \dots, \gamma_{d,N(R)})'$, and super population parameter, $\boldsymbol{\vartheta}$.

We arrive at the final computable expression of $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$ through a three-step linearization (see Appendix A4 in online supplementary materials). We carry out the first linearization on the random quantities $\hat{\mathbf{t}}$ and the second and third linearization steps on the random vectors \mathbf{y} , and $\boldsymbol{\lambda}$. We calculate each that are first derived, taking into consideration the specific source of variability for the given conditional set-up.

1. *First linearization step.* The estimator \hat{Y}_{R_d} is linearized with respect to the vector $\hat{\mathbf{t}}$ where the derivatives are computed at the model expected value $\boldsymbol{\vartheta}$, thus obtaining:

$$\hat{Y}_{R_d} = E_M(\hat{Y}_{R_d}|\boldsymbol{\lambda}, \boldsymbol{\vartheta}) + \sum_{k \in R} \sum_{i=1}^I \gamma_{d,k} f_{ki}(\hat{t}_i - \vartheta_i) + r_1, \tag{14}$$

where

$$f_{ki} = \left. \frac{\partial f(\mathbf{x}_k; \hat{\mathbf{t}})}{\partial \hat{t}_i} \right|_{\hat{t}_i = \vartheta_i} : k = 1, \dots, N_{(R_d)}; i = 1, \dots, I, \text{ being } r_1 = O_p(1/\sqrt{n}), \tag{15}$$

is a remainder of minor order. Discarding the remainder, the model variance, $V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})$, which represents the core part of $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$, is given by:

$$\begin{aligned} V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda}) &\cong V_M \left[\sum_{k \in R} \sum_{i=1}^I \gamma_{d,k} f_{ki}(\hat{t}_i - \vartheta_i) | \boldsymbol{\lambda} \right] = V_M \left[\boldsymbol{\gamma}'_d \mathbf{F}(\hat{\mathbf{t}} - \boldsymbol{\vartheta}) | \boldsymbol{\lambda} \right] \\ &= \boldsymbol{\gamma}'_d \mathbf{F} [V_M(\hat{\mathbf{t}}|\boldsymbol{\lambda})] \mathbf{F}' \boldsymbol{\gamma}_d, \end{aligned} \tag{16}$$

where $\mathbf{F} = [f_{ki}]$ is a $N_{(R)} \times I$ matrix.

2. *Second linearization step.* The term

$$\sum_{k \in R} \sum_{i=1}^I \gamma_{d,kf_{ki}}(\hat{t}_i - \vartheta_i) = \boldsymbol{\gamma}'_d \mathbf{F}(\hat{\mathbf{t}} - \boldsymbol{\vartheta})$$

in Equation (16) is linearized with respect to the \mathbf{y} variables, where the derivatives are computed at $\hat{\mathbf{t}} = \boldsymbol{\vartheta}$ and $\mathbf{y} = \tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_k, \dots, \tilde{y}_{N(R)})'$ keeping the $\boldsymbol{\lambda}$ vector fixed. The mathematical explanations are detailed in Appendix A2 in the online supplementary materials. Here, in this section, we limit ourselves to providing and describing the main and essential results. Let

$$\sum_{j \in R} \mathbf{g}_j(\hat{\mathbf{t}}; \mathbf{y}; \boldsymbol{\lambda}) = \mathbf{0}_I, \tag{17}$$

be the system of GEE, generalized estimating Equations (Ziegler 2015) for estimating the vector $\hat{\mathbf{t}}$ in which

$$\mathbf{g}_j(\hat{\mathbf{t}}; \mathbf{y}; \boldsymbol{\lambda}) = [g_{j1}(\hat{\mathbf{t}}; \mathbf{y}; \boldsymbol{\lambda}), \dots, g_{ji}(\hat{\mathbf{t}}; \mathbf{y}; \boldsymbol{\lambda}), \dots, g_{jI}(\hat{\mathbf{t}}; \mathbf{y}; \boldsymbol{\lambda})]', \tag{17b}$$

is the I vector of the score functions $g(\hat{\mathbf{t}}; \mathbf{y}; \boldsymbol{\lambda})$ for the parameter \hat{t}_i ($i = 1, \dots, I$) of unit j , where $\mathbf{0}_I$ is a vector of I zeroes. Adopting the linear approximation first proposed by Binder and Patak (1994) and then, among others, by Chambers and Clark (2015, 123–125), we have:

$$(\hat{\mathbf{t}} - \boldsymbol{\vartheta}) \cong \mathbf{A}_{\boldsymbol{\vartheta}}^{-1} \sum_{j \in R} \mathbf{g}_j(\boldsymbol{\vartheta}; \mathbf{y}; \boldsymbol{\lambda}), \tag{18}$$

being

$$\mathbf{A}_{\boldsymbol{\vartheta}} = \left[a_{i\ell|\boldsymbol{\vartheta}} = \frac{\partial \sum_{j \in R} g_{ji}(\hat{\mathbf{t}}; \mathbf{y}; \boldsymbol{\lambda})}{\partial \hat{t}_\ell} \Bigg|_{\hat{\mathbf{t}} = \boldsymbol{\vartheta}} \right],$$

a $(I \times I)$ matrix (in which $i, \ell = 1, \dots, I$), and where $\mathbf{g}_j(\boldsymbol{\vartheta}; \mathbf{y}; \boldsymbol{\lambda}) = \{g_{ji}(\boldsymbol{\vartheta}; \mathbf{y}; \boldsymbol{\lambda}); i = 1, \dots, I\}$ is defined in Equation (17b) by the substitution of $\hat{\mathbf{t}}$ with $\boldsymbol{\vartheta}$. Thus, according to the Equation (A3) of Appendix A2 in online supplementary materials, we have the following approximation, which holds for $n \gg I$:

$$\begin{aligned} V_M(\hat{Y}_{R_d} | \boldsymbol{\lambda}) &\cong V_M \left[\sum_{k \in R} \sum_{j \in R} \sum_{i=1}^I \gamma_{d,kf_{ki}} u_{j|i\lambda} y_j | \boldsymbol{\lambda} \right] \\ &= V_M \left(\boldsymbol{\gamma}'_d \mathbf{F} \sum_{j \in R} \mathbf{u}_{j|\lambda} y_j \Big| \boldsymbol{\lambda} \right) \\ &= \boldsymbol{\gamma}'_d \mathbf{F} \sum_{j \in R} \left[\mathbf{u}_{j|\lambda} \mathbf{u}'_{j|\lambda} \sigma_{y_j}^2 + \sum_{\ell \neq j} \mathbf{u}_{j|\lambda} \mathbf{u}'_{\ell|\lambda} \sigma_{y_j \ell} \right] \mathbf{F}' \boldsymbol{\gamma}_d, \end{aligned} \tag{19}$$

where $\mathbf{u}_{j|\lambda} = \frac{-\partial [\mathbf{A}_{\boldsymbol{\vartheta}}^{-1} \mathbf{g}_j(\boldsymbol{\vartheta}; \mathbf{y}; \boldsymbol{\lambda})]}{\partial y_j} \Big|_{\mathbf{y} = \tilde{\mathbf{y}}} (u_{j1|\lambda}, \dots, u_{j1|\lambda}, \dots, u_{j1|\lambda})'$.

For the general linear model in which $\tilde{y}_k = \mathbf{x}'_k \boldsymbol{\vartheta}$, the matrix $\mathbf{A}_\boldsymbol{\vartheta}$ is an $I \times I$ identity matrix.

3. *Third linearization step.* According to the approach proposed by Graf (2015) and by Vallée and Tillé (2019), the terms $\mathbf{u}_{j|\lambda}$, included in Equation (19), are linearized with respect to the sampling indicators $\boldsymbol{\lambda}$ around the sample design expected value $\boldsymbol{\pi}$. We have,

$$\mathbf{u}_{j|\lambda} \cong \mathbf{u}_{j,\pi} + \partial \mathbf{u}_j (\lambda_j - \pi_j), \tag{20}$$

where $\mathbf{u}_{j,\pi}$ is obtained as $\mathbf{u}_{j|\lambda}$ by substituting the values of λ_j with the corresponding expected values π_j and

$$\partial \mathbf{u}_j = \left. \frac{\partial \mathbf{u}_{j|\lambda}}{\partial \lambda_j} \right|_{\lambda=\pi} = (\partial u_{j1}, \dots, \partial u_{ji}, \dots, \partial u_{jI})'.$$

Then, we have

$$E_P(\mathbf{u}_{j|\lambda} \mathbf{u}'_{j|\lambda}) = \mathbf{u}_{j,\pi} \mathbf{u}'_{j,\pi} + \partial \mathbf{u}_j \partial \mathbf{u}'_j \pi_j (1 - \pi_j). \tag{21}$$

$$E_P(\mathbf{u}_{j|\lambda} \mathbf{u}'_{\ell|\lambda}) = \mathbf{u}_{j,\pi} \mathbf{u}'_{\ell,\pi} + \partial \mathbf{u}_j \partial \mathbf{u}'_{\ell} (\pi_{j\ell} - \pi_j \pi_{\ell}), \tag{22}$$

where $\pi_{j\ell}$ is the joint inclusion probability of units j and ℓ .

Then, considering the sampling expected values Equations (21) and (22) into Equation (19), we have

$$\begin{aligned} E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] &\cong \boldsymbol{\gamma}'_d \mathbf{F} \{E_P[V_M(\hat{\mathbf{y}}|\boldsymbol{\lambda})]\} \mathbf{F}' \boldsymbol{\gamma}_d \\ &= \boldsymbol{\gamma}'_d \mathbf{F} \left\{ \sum_{j \in R} \left[\left(\mathbf{u}_{j,\pi} \mathbf{u}'_{j,\pi} + \partial \mathbf{u}_j \partial \mathbf{u}'_j \pi_j (1 - \pi_j) \right) \sigma_{y_j}^2 + \right. \right. \\ &\quad \left. \left. \sum_{\ell \neq j} \left(\mathbf{u}_{j,\pi} \mathbf{u}'_{\ell,\pi} + \partial \mathbf{u}_j \partial \mathbf{u}'_{\ell} (\pi_{j\ell} - \pi_j \pi_{\ell}) \right) \sigma_{y_{j\ell}} \right] \right\} \mathbf{F}' \boldsymbol{\gamma}_d. \end{aligned} \tag{23}$$

The above expression cannot be computed for many usual sampling designs since the joint inclusion probabilities, $\pi_{j\ell}$, are unknown. Starting from result of Equation (9), we propose an upward approximation of Equation (23) based on the first-order inclusion probabilities, which we recommend for the calculus of $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$:

$$\begin{aligned} &E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] \\ &\leq \boldsymbol{\gamma}'_d \mathbf{F} \sum_{j \in R} \left[\left(\mathbf{u}_{j,\pi} \mathbf{u}'_{j,\pi} + \partial \mathbf{u}_j \partial \mathbf{u}'_j \pi_j (1 - \pi_j) \right) \sigma_{y_j}^2 + \left(\sum_{\ell \neq j} \mathbf{u}_{j,\pi} \mathbf{u}'_{\ell,\pi} \sigma_{y_{j\ell}} \right) \right] \mathbf{F}' \boldsymbol{\gamma}_d. \end{aligned} \tag{24}$$

In Appendix A3 in the online supplementary materials, we give a lower bound of $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$. Equation (24) is defined using unit-level elements: $\mathbf{u}_{j,\pi}$, $\partial \mathbf{u}_j$, π_j , $\sigma_{y_j}^2$, $\sigma_{y_{j\ell}}$ which are not domain specific.

4.3. Calculus of the Other Components of GMSE

According to the model setting (3.2), the component $V_M(Y_{U_d})$ is given by

$$V_M(Y_{U_d}) = \sum_{j \in R_d} \left(\sigma_{y,j}^2 + \sum_{\ell \neq j} \sigma_{y,j\ell} \right). \tag{25}$$

For the calculus of $Cov_M[E_P(\hat{Y}_{R_d}|\mathbf{y}), Y_{U_d}]$, we adopted a linearization consisting of two steps, where in the first step, the estimator \hat{Y}_{R_d} is linearized with respect to the vector $\hat{\mathbf{t}}$ and the first step is computed at the sampling design expected value $\mathbf{t} = E_P(\hat{\mathbf{t}}|\boldsymbol{\lambda})$. In the second step, adopting the same approach as Binder and Patak (1994), the estimating expressions of \mathbf{t} are linearized around $\tilde{\mathbf{y}}$ and $\boldsymbol{\vartheta}$, thus obtaining:

$$Cov_M[E_P(\hat{Y}_{R_d}|\mathbf{y}), Y_{U_d}] \cong \sum_{k \in R} \sum_{j \in R} \sum_{i=1}^I \gamma_{d,k} f_{ki} u_{ji,\pi} \left(\sigma_{y,j}^2 + \sum_{(\ell \neq j) \cap (\ell \in R_d)} \gamma_{d,\ell} \sigma_{y,j\ell} \right). \tag{26}$$

4.4. Plug-In Estimate of the GMSE

The plug-in estimate of GMSE may be computed by inserting the estimates of $\hat{\mathbf{t}}$, \hat{y}_k ($k = 1, \dots, N_{(R)}$), $\hat{\sigma}_{y,j}^2$ and $\hat{\sigma}_{y,j\ell}$ ($j, \ell = 1, \dots, N_{(R)}$) in place of the unknown parameters $\boldsymbol{\vartheta}$, \tilde{y}_k ($k = 1, \dots, N_{(R)}$), $\sigma_{y,j}^2$ and $\sigma_{y,j\ell}$ ($j, \ell = 1, \dots, N_{(R)}$) in the expressions of the different components of the GMSE. According to Ziegler (2015, point 5, 121), these plug-in estimates are strongly consistent estimators of the different components of the variance.

4.5. Example with the Logistic Model

Consider a logistic model where $y_k = 1$ with probability $\tilde{y}_k = f(\mathbf{x}_k; \boldsymbol{\vartheta})$ and $y_k = 0$ with probability $1 - \tilde{y}_k$, where $f(\mathbf{x}_k; \boldsymbol{\vartheta}) = e^{\mathbf{x}'_k \boldsymbol{\vartheta}} / (1 + e^{\mathbf{x}'_k \boldsymbol{\vartheta}})$. Under the model-based approach, the estimating equations for the calculus of the GMSE using the first-order linear approximations, derived by the log-likelihood, are

$$\sum_{j \in R} \mathbf{g}_j(\hat{\mathbf{t}}; \mathbf{y}; \boldsymbol{\lambda}) = \sum_{j \in R} \mathbf{x}_j \left[y_j - \frac{e^{\mathbf{x}'_j \hat{\mathbf{t}}}}{1 + e^{\mathbf{x}'_j \hat{\mathbf{t}}}} \right] \lambda_j.$$

The matrix $\mathbf{A}_{\boldsymbol{\vartheta}}$ is given by:

$$\mathbf{A}_{\boldsymbol{\vartheta}} = - \sum_{j \in R} \left[\frac{\mathbf{x}_j \mathbf{x}'_j e^{\mathbf{x}'_j \boldsymbol{\vartheta}}}{(1 + e^{\mathbf{x}'_j \boldsymbol{\vartheta}})^2} \right] \lambda_j$$

and the vectors $\mathbf{u}_{j|\lambda}$, $\mathbf{u}_{j,\pi}$ and $\partial \mathbf{u}_j$ are expressed as $\mathbf{u}_{j|\lambda} = -\mathbf{A}_{\boldsymbol{\vartheta}}^{-1} \mathbf{x}_j \lambda_j$, $\mathbf{u}_{j,\pi} = -\mathbf{A}_{\boldsymbol{\vartheta}}^{-1} \mathbf{x}_j \pi_j$, $\partial \mathbf{u}_j = -\mathbf{A}_{\boldsymbol{\vartheta}}^{-1} \mathbf{x}_j$.

5. Tips on Computational Aspects

GMSE may be expressed as the sum of elementary unit variances, $\sigma_{dy,k}^2$, over the register domain units:

$$\text{GMSE}(\hat{Y}_{R_d}) \cong \sum_{k \in R_d} \sigma_{dy,k}^2. \tag{27}$$

Considering jointly the Equations (13a) or (13b), (24), (25) and (26), we have that the quantities $\sigma_{dy,k}^2$ are given by

$$\sigma_{dy,k}^2 = \begin{cases} \sigma_{A_{dy,k}}^2 - \sigma_{B_{dy,k}}^2 & \text{if } E_P(\hat{Y}_{R_d} | \mathbf{y}) = Y_{U_d} \\ \sigma_{A_{dy,k}}^2 + \sigma_{B_{dy,k}}^2 - 2\sigma_{C_{dy,k}}^2 & \text{if } E_P(\hat{Y}_{R_d} | \mathbf{y}) \neq Y_{U_d} \end{cases} \tag{28a}$$

$$\tag{28b}$$

where $\sigma_{A_{dy,k}}^2$, $\sigma_{B_{dy,k}}^2$ and $\sigma_{C_{dy,k}}^2$ are the elementary unit variances from the sum of which(over the register domain units) the three components of GMSE are obtained (namely, $E_P[V_M(\hat{Y}_{R_d} | \boldsymbol{\lambda})]$, $V_M(Y_{U_d})$ and $\text{Cov}_M[E_P(\hat{Y}_{R_d} | \mathbf{y}), Y_{U_d}]$).

With simple algebra from Equations (24), (25) and (26), we have:

$$\sigma_{A_{dy,k}}^2 = \sum_{(k' \neq k) \cap (k' \in R_d)} \sum_{i=1}^I \sum_{i'=1}^I f_{ki} f_{k'i'} v_{y,ii'}, \tag{29}$$

$$\sigma_{B_{dy,k}}^2 = \sigma_{yk}^2 + \sum_{(k' \neq k) \cap (k' \in R_d)} \sigma_{y,kk'}, \tag{30}$$

$$\sigma_{C_{dy,k}}^2 = \sum_{j \in R} \sum_{i=1}^I f_{ki} u_{ki, \pi} \left(\sigma_{yj}^2 + \sum_{(k' \neq k) \cap (k' \in R_d)} \sigma_{y,kk'} \right), \tag{31}$$

in which

$$v_{y,ii'} = \sum_{j \in R} \left(\sigma_y^2, u_{ji, \pi} u_{ji', \pi} + \sum_{\ell \neq j} u_{ji, \pi} u_{\ell i', \pi} \sigma_{y,j\ell} \right) + \partial u_{ji} \partial u_{ji'} \sigma_{y,j}^2 \pi_j (1 - \pi_j). \tag{32}$$

From Equation (24), we see that the sum over the domain units of $\sigma_{A_{dy,k}}^2$ is an upward approximation of the component of $E_P[V_M(\hat{Y}_{R_d} | \boldsymbol{\lambda})]$

$$E_P[V_M(\hat{Y}_{R_d} | \boldsymbol{\lambda})] \leq \sum_{k \in R_d} \sigma_{A_{dy,k}}^2 = \sum_{k \in R_d} \sum_{k' \in R_d} \sum_{i=1}^I \sum_{i'=1}^I f_{ki} f_{k'i'} v_{y,ii'}. \tag{33}$$

The estimates $\hat{\sigma}_{dy,k}^2$ (introduced in Equation 6) of the elementary unit variances $\sigma_{dy,k}^2$ may be estimated using the usual plug-in technique

$$\hat{\sigma}_{dy,k}^2 = \begin{cases} \hat{\sigma}_{A_{dy,k}}^2 - \hat{\sigma}_{B_{dy,k}}^2 & \text{if } E_P(\hat{Y}_{R_d} | \mathbf{y}) = Y_{U_d} \\ \hat{\sigma}_{A_{dy,k}}^2 + \hat{\sigma}_{B_{dy,k}}^2 - 2\hat{\sigma}_{C_{dy,k}}^2 & \text{if } E_P(\hat{Y}_{R_d} | \mathbf{y}) \neq Y_{U_d} \end{cases} \tag{34a}$$

$$\tag{34b}$$

by substituting the expressions of $\hat{\sigma}_{A_{dy,k}}^2$, $\hat{\sigma}_{B_{dy,k}}^2$ and $\hat{\sigma}_{C_{dy,k}}^2$ the estimates $\hat{\boldsymbol{\tau}}$, \hat{y}_k ($k = 1, \dots, N_{(R)}$), $\hat{\sigma}_{y,j}^2$ and $\hat{\sigma}_{y,j\ell}$ ($j, \ell = 1, \dots, N_{(R)}$), in place of the unknown parameters $\boldsymbol{\tau}$, y_k ($k = 1, \dots, N_{(R)}$), $\sigma_{y,j}^2$ and $\sigma_{y,j\ell}$ ($j, \ell = 1, \dots, N_{(R)}$).

From the above expressions, the main results are as follows.

1. *Domain characterization.* The quantities $v_{yit'}$, f_{ki} , $u_{ji,\pi}$, ∂u_{ji} , σ_{yj}^2 , and $\sigma_{y,j\ell}$ (and their corresponding plug-in estimates $\hat{v}_{y,ii'}$, \hat{f}_{ki} , $\hat{u}_{ji,\pi}$, $\hat{\sigma}_{yj}^2$ and $\hat{\sigma}_{y,j\ell}$) are not domain specific. $\sigma_{Ady,k}^2$, $\sigma_{Bdy,k}^2$ and $\sigma_{Cdy,k}^2$ are domain specific since they are defined as a sum over R_d .
2. *Space for the storage.* A small amount of space is needed for storing the $(I \times I)$ matrix $\hat{\mathbf{v}}_y = \{\hat{v}_{yit'}; i, i' = 1, \dots, I\}$ whereas storing the matrix $\hat{\mathbf{F}} = \{\hat{f}_{ki}; k = 1, \dots, N_{(R)}; i = 1, \dots, I\}$ requires a large volume of space. The quantities $\hat{u}_{ji,\pi}$, $\hat{\sigma}_{yj}^2$, and $\hat{\sigma}_{y,j\ell}$ require a large volume of space. They are directly involved only in the calculus of the subdominant parts of GMSE and generally result in providing a negative contribution to this quantity.
3. *Computational complexity.* Regarding the predominant component, the calculus of the matrix $\hat{\mathbf{v}}_y$ involves the estimation of the parameters $\hat{\mathbf{t}}$, $\hat{\sigma}_{y,j}^2$ and $\hat{\sigma}_{y,j\ell}^2$ and the calculus of the vectors $\hat{\mathbf{u}}_{j,\pi}$ and $\partial \hat{\mathbf{u}}_j$ including different steps of linearization. On the other hand, the calculus of $\hat{\mathbf{F}}$ will indeed seem simple once the estimate of the parameter $\hat{\mathbf{t}}$ has been obtained, and the calculus of the values \hat{f}_{ki} may be obtained directly by just applying the specific *analytical expression* to the \mathbf{x}_k vector. On the fly, the calculus of the sub dominant components of GMSE may be cumbersome mainly because of the domain dependency on the sum $\sum_{(k' \neq k) \cap (k' \in R_d)} \sigma_{y,kk'}$; nevertheless, if we neglect these terms, we introduce only as light upward approximation of the GMSE.
4. *Stability over time.* The quantities $\hat{\mathbf{v}}_y$ and the functional form of $\hat{\mathbf{F}}$ are relatively table over time. The functional form of $\hat{\mathbf{F}}$ does not change unless the method of prediction is modified. The matrix $\hat{\mathbf{v}}_y$ is essentially a function of (A) the sample design properties (and does not depend on the specific sample selection), which change only rarely when ever the survey is restructured, and (B) the structure of the model variances and covariances $\hat{\sigma}_{y,j}^2$ and $\hat{\sigma}_{y,j\ell}^2$, which change rarely and only in the case where there is a structural break in the y value sand in the model for their generation.

The above expressions and the results of the empirical experiment (in Section 7) offer some suggestions on how to develop a feasible and robust computational strategy. First, we note that the subdominant components give a negative contribution to GMSE and that this tends to be negligible for unplanned domains in which $E_P(\hat{Y}_{R_d} | \mathbf{y}) \neq Y_{U_d}$. This finding invites the examination of two different alternatives: one for the planned domain (for which $E_P(\hat{Y}_{R_d} | \mathbf{y}) = Y_{U_d}$) and the other for unplanned domains.

For the **planned case** where domains are well known in advance and limited in number, the Equation (13a) is used for the computation. In particular:

1. The matrix $\hat{\mathbf{v}}_y$ is computed and stored.
2. The matrix $\hat{\mathbf{F}}$ is not stored. The values \hat{f}_{ki} are computed on the fly on the basis of its functional form, which links these values directly to the auxiliary variables \mathbf{x}_k and to the parameters $\hat{\mathbf{t}}$.
3. With regard to the computation of matrix $\hat{\mathbf{F}}$, the only two objects that are permanently stored are the vector of the parameter $\hat{\mathbf{t}}$ and the functional forms that permit the computation of the \hat{f}_{ki} values.
4. The elementary unit variances $\hat{\sigma}_{Bdy,k}^2$ are computed and stored.

For the **unplanned domains** (Equation (13b)), generally unpredictable in their number, the quantity $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$ is used as an upward approximation of the GMSE. This can be easily computed on the fly on the basis of the stored material as defined in points 1 and 3 listed above.

6. The Coverage Errors

The register may be affected by coverage errors, which implies $N_{(R_d)} \neq N_{(U_d)}$, where $N_{(R_d)}$ is the number of units in domain d listed in the register, and $N_{(U_d)}$ is the domain population size.

An unbiased estimate $\hat{N}_{umb,(U_d)}$ may be obtained directly from the register with the Extended Dual System Estimator (Nirel and Glickman 2009; Pfeffermann 2015) by summing up the predicted values of a particular w variable over the register units:

$$\hat{N}_{umb,(U_d)} = \sum_{k \in R_d} \hat{w}_k, \tag{35}$$

in which

$$\hat{w}_k = \frac{\hat{P}(k \in U|k \in R)}{\hat{P}(k \in R|k \in U)} \tag{36}$$

represents the model (roughly) unbiased prediction of

$$w_k = \frac{P(k \in U|k \in R)}{P(k \in R|k \in U)} \tag{37}$$

where $\hat{P}(k \in U|k \in R)$ is the model’s unbiased estimate of the conditional probability $P(k \in U|k \in R)$ such that unit k included in the register belongs to the population and $\hat{P}(k \in R|k \in U)$ is the model’s unbiased estimate of the conditional probability $P(k \in R|k \in U)$ such that unit k belonging to the population is included in the register. This approach has been adopted for the Italian Population Base Register by integrating the register data with the Census *Population Coverage Survey* carried out each year as a component of the Italian Permanent Census Survey System (Righi et al. 2021). The GMSE of $\hat{N}_{umb,(U_d)}$ may be expressed as and specified as described in Subsection 4.2 by defining the predictions and model’s expected values of the w variables (instead of the y variables).

Moreover, we note that in the case in which the register is affected by coverage errors, a weighted estimator of the total of a generic y variable in the domain U_d , Y_{U_d} can conveniently be expressed as:

$$\hat{Y}_{U_d} = \sum_{k \in R_d} \hat{y}_k \hat{w}_k.$$

The GMSE of this estimator can be obtained by considering its linear approximation:

$$\text{GMSE}(\hat{Y}_{U_d}) \cong \text{GMSE} \left(\sum_{k \in R_d} \hat{y}_k w_k + \sum_{k \in R_d} \tilde{y}_k \hat{w}_k \right). \tag{38}$$

We omit further technical developments, which can be easily derived according to the procedure given in Section 4.

7. Experimental Study

This experimental study, which is based on real data, compares the empirical GMSE of a Monte Carlo simulation with the approximate GMSE obtained from Taylor approximations.

The data set for the empirical study is an administrative archive that contains information regarding the population of 21,782 Sapienza University of Rome (Italy) alumni who graduated between March 1, 2008 and February 28, 2009 and who signed a job contract in the three years following graduation (Alleva and Petrarca 2013; Gruppo UNI. CO 2015; Petrarca 2014, a, b). The study focuses on the disciplinary sectors of engineering, sciences, literature, economics and statistics, psychology, chemistry and pharmacy, and architecture. The data set has 7,085 units. However, the simulations conducted on other subsets of disciplinary sectors confirm the results shown here.

The target y variable is the number of days worked during the three years after graduation. The vector of auxiliary variables for a unit is $\mathbf{x}_k = (x_{k1}, x_{k2}, x_{k3}, x_{k4}, x_{k5}, x_{k6}, x_{k7}, x_{k8})'$, where $x_{k1} = 1$; x_{k2} : gender of a graduate; x_{k3} : age at the time of graduation; x_{k4} : graduation on time (yes/no); x_{k5} : graduation from a second-cycle programme (yes/no); x_{k6} : number of days that a graduate has waited before obtaining a permanent contract; x_{k7} : number of days that a graduate has waited before obtaining a contract with a highly qualified position (ISCO 1-ISCO 2); and x_{k8} : number of days that a graduate has waited before obtaining a contract with an actual duration of more than or equal to eight months.

7.1. Standard Simulation and Linear Approximation

We generated 1,000 populations of 7,085 units. For each population, a vector of the target variables y_k was generated as described in Section 3 by taking the sum of two components: $y_k = E_M(y_k) + e_k = \tilde{y}_k + e_k$, where \tilde{y}_k is the vector of the fitted values obtained from a linear regression model attuned to the super-population, and e_k is generated with a normal distribution with mean 0 and variance equal to the variance of the y_k in the real data set ($\sigma^2 = 0.1159733 \cdot 10^7$). For each population, 1,000 samples of $n = 500$ units were selected utilizing a simple random sample design without replacement. The two processes that generate populations and samples allow us to simulate the model and sampling uncertainty. For each sample we obtained from a simple linear regression model, the estimated regression coefficients $\hat{\mathbf{t}}$ formulate the $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{t}}$ vector to be utilized for constructing predictions. The sum of these values restricted to the domain built the \hat{Y}_{R_d} estimates.

The evaluation of our information by linear approximation only requires us to know the matrix of the auxiliary variables and then implement the calculation of the formulae given in the text. We had two types of domains: the internal domains for which the domain membership variable belongs to the vector of auxiliary variables and the external domains otherwise. Note that in the case of internal domains, we apply a generalized regression estimator (Särndal et al. 1992). The size of each domain is presented in Table 2.

7.2. Results

First, we discuss the simulation results concerning GMSE for the two large domains: *gender* and *scientific group* (obtained by summing *sciences*, *chemistry* and *pharmacy*,

Table 2. Internal and external domains with their size.

Internal domain	Population size	External domain	Population size
Gender_female	4,281	Scientific group	3,368
Gender_male	2,804	Others	3,717

economics and statistics and engineering) and the domain others (containing architecture, literature and psychology).

In Tables 3 and 4, we report the values of GMSE and its components for the large internal and external domains, respectively. For the sake of brevity, the tables display only the results for gender female and scientific group. Similar results were produced for the other two domains. Part A of Table 3 reports the experimental results of the $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$ in the case of an internal domain (gender female), whereas part B of Table 3 shows the results for an external domain (scientific group) achieved using the Monte Carlo simulation and linear approximations, hereinafter referred to as “empirical” and “linearized” expectations, respectively.

For the empirical expectations of the GMSE, we considered Equations (4) and (13a) for the internal domains. In the case of the external domain, the evaluation of GMSE is based on Equation (13b) because in this case, the projection estimator is biased, $E_P[(\hat{Y}_{R_d}|\mathbf{y})] \neq Y_{U_d}$. Equations (3.2) and (3.7) of Table 3 show that the dominant part of GMSE comes from $E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})]$.

For the linearized Equations (3.3–7.5 of Table 3-part A; 3.8–3.10 of Table 3-part B), we give the estimates for the dominant contribution of GMSE computed from the definitions of Equation (23) applied to the linear regression model. As expected, these evaluations are sufficiently in agreement with one another.

Table 3. GMSE and its components.

A: Case of an internal domain (Gender_female)	
Empirical	
$V_M(Y_{U_d}) = 426.6$	
$GMSE(\hat{Y}_{R_d}) = E_P E_M(\hat{Y}_{R_d} - Y_{U_d})^2 = 6,649.0$	(3.1)
$GMSE(\hat{Y}_{R_d}) \cong E_P[V_M(\hat{Y}_{R_d} \boldsymbol{\lambda})] - V_M(Y_{U_d}) = 7,069.9 - 426.6 = 6,643.3$	(3.2)
Linearized	
$V_M(Y_{U_d}) = \boldsymbol{\gamma}'_d \sum_{y_j} \boldsymbol{\gamma}_d = 496.5$	(3.3)
$E_P[V_M(\hat{\mathbf{t}} \boldsymbol{\lambda})] = \sum_{j \in R} [E_P(\mathbf{u}_{j \lambda} \mathbf{u}'_{j \lambda}) \sigma_{y_j}^2] = 7,021.6$	(3.4)
$GMSE(\hat{Y}_{R_d}) = E_P[V_M(\hat{Y}_{R_d} \boldsymbol{\lambda})] - V_M(Y_{U_d}) = 7,021.6 - 496.5 = 6,525.1$	(3.5)
A: Case of an external domain (Scientific_group)	
Empirical	
$V_M(Y_{U_d}) = 380.2$	
$GMSE(\hat{Y}_{R_d}) = E_P E_M(\hat{Y}_{R_d} - Y_{U_d})^2 = 2,996.4$	(3.6)
$GMSE(\hat{Y}_{R_d}) \cong E_P[V_M(\hat{Y}_{R_d} \boldsymbol{\lambda})] + V_M(Y_{U_d}) - 2Cov_M[E_P(\hat{Y}_{R_d} \mathbf{y}), Y_{U_d}]$ $= 3,045.8 + 380.2 - 429.5 = 2,996.5$	(3.7)
Linearized	
$V_M(Y_{U_d}) = \boldsymbol{\gamma}'_d \sum_{y_j} \boldsymbol{\gamma}_d = 390.6$	(3.8)
$E_P[V_M(\hat{\mathbf{t}} \boldsymbol{\lambda})] = \sum_{j \in R} [E_P(\mathbf{u}_{j \lambda} \mathbf{u}'_{j \lambda}) \sigma_{y_j}^2] = 2,990.3$	(3.9)
$GMSE(\hat{Y}_{R_d}) = E_P[V_M(\hat{Y}_{R_d} \boldsymbol{\lambda})] + V_M(Y_{U_d}) - 2Cov_M[E_P(\hat{Y}_{R_d} \mathbf{y}), Y_{U_d}]$ $= 2,990.3 + 390.6 - 429.5 = 2,951.4$	(3.10)

The numbers are scaled by a factor of 10^7 .

Table 4. Values of $E_P[V_M(\hat{Y}_{R_d}|\lambda)]$ in the case of simulation and linearization.

Domains	Domain population size	Expected sampling fraction (n_d/N_d)	$E_P[V_M(\hat{Y}_{R_d} \lambda)]$		Difference* (A-B)	Relative difference (A-B)/A
			Empirical* (A)	Linearized* (B)		
Architecture	905	0.128	214.2	210.8	3.36	0.016
Chemistry and Pharmacy	400	0.056	85.8	83.2	2.56	0.030
Economics and Statistics	1,349	0.190	472.6	458.3	14.23	0.030
Engineering	1,259	0.178	590.6	580.5	10.13	0.017
Literature	1,975	0.279	1,039.4	1,020.7	18.70	0.018
Psychology	837	0.118	248.0	237.9	10.15	0.041
Sciences	360	0.051	38.8	37.5	1.26	0.033
Scientific group	3,368	0.475	3,045.8	2,990.3	55.50	0.018

*The numbers are scaled by a factor of 10^7 .

Our attention now focuses on $E_P[V_M(\hat{Y}_{R_d}|\lambda)]$. Experiment serves to investigate the external disciplinary sector domains, allowing us to compute the expectation $E_P[V_M(\hat{Y}_{R_d}|\lambda)]$ for small and large domains. Table 4 reports the values of $E_P[V_M(\hat{Y}_{R_d}|\lambda)]$ in the case of simulation (column A) and linearization (column B). The differences between the two estimates are positive and small, ranging between 16 and 41%.

Figures 1a and 1b show the data of Table 4, where the circled points correspond to the values of external domains, whereas the diamond points correspond to the values for the

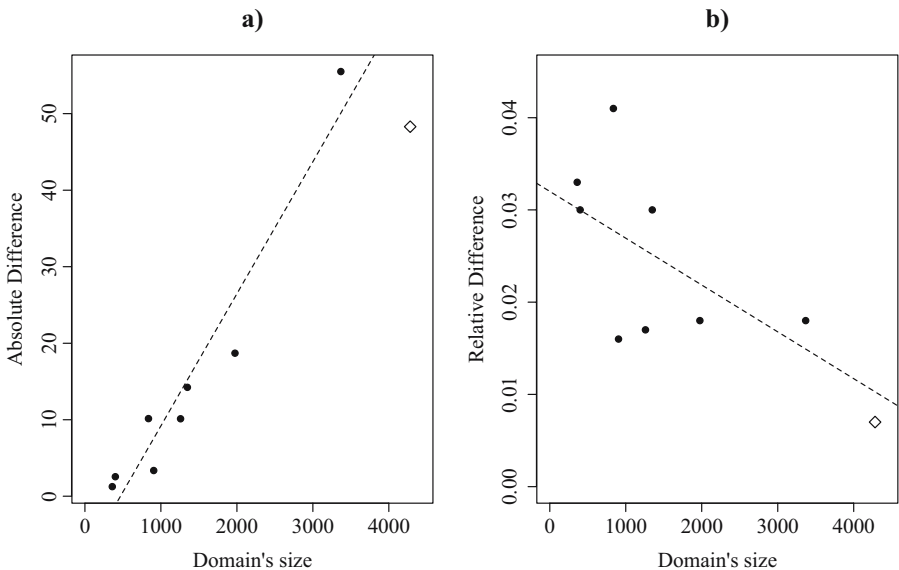


Fig. 1. Absolute (a) and relative (b) differences in $E_P[V_M(\hat{Y}_{R_d}|\lambda)]$ between simulation and linearization according to the size of the domain.

*The point \diamond refers to the internal domain gender_female

internal domain *gender female*, also shown here for completeness. The trend lines were drawn without considering the values of the internal domain.

As shown in Figure 1a, the difference between the empirical and linearized expectations indicates a positive relationship with the population size of the external domains. On the other hand, Figure 1b, shows that the relative difference increases when the domain sample size decreases. These findings confirm that the linearization method produces a downward approximation directly related to the sample size.

8. Conclusions

In this article, we have proposed the *global mean square error* as an appropriate measure to assess the accuracy of register aggregates. This measure has some relevant qualities: generality, stability over time and robustness in the case of model failure. It is easy to use and communicate to users and could be accepted as a measure of accuracy by the main professional families of methodologists within the National Statistical Institutes.

Our approach is based on only the first and second moments of the random distributions of the specific source of uncertainty. Its calculus does not imply full knowledge of the underlying distributions.

In addition, we suggested an immediate GMSE calculation strategy for any unexpected user request by simply aggregating domain-dependent variances estimated at the unit level. The amount of information to store for this calculation is limited, and the domain dependency is limited to a few useful variables. The calculation strategy suggested here is a powerful advantage of our proposal, as it allows NSIs to build algorithms that can be applied *instantly* to any user request, thus improving the relevance, transparency and confidence of official statistics.

The simulation conducted confirms the accuracy of the different GMSE decompositions proposed in Section 4 as model-assisted projection estimators, whether they are design-unbiased or biased. Furthermore, the very small discrepancies between empirical and linearized expectations suggest that the proposed approximation method can be undertaken as a valid computational strategy. We emphasize once again that the linearized variance is calculated using unit-level elements: $\mathbf{u}_{j,\pi}$, $\partial \mathbf{u}_{j,\pi}$, π_j , $\sigma_{y,j}^2$ and is suitable for the calculation of an accuracy measurement of the statistics based on registers.

In addition to the more extensive validation studies to be launched to confirm the benefits and robustness of the empirical results, the main further steps to be taken in the research outlined here are the definition of a validation strategy and targeted extensions with regard to both other sources of uncertainty and parameters other than means and totals that cannot be expressed as simple linear functions of the register predictions.

Although significant validation may not be feasible at this stage, we suggest that NSIs plan, on a regular (e.g., annual) basis, experimental studies conducted for specific domains where the GMSE values obtained as proposed in this article are compared with corresponding values obtained from other approaches. We suggest considering replication methods that repeat the whole process of calculating the register predictions. These experiments may show inconsistencies in the experimental evidence on which further empirical and theoretical investigations should be considered. Furthermore, they could validate whether the asymptotic properties adopted here hold.

Beyond the coverage error considered here, other extensions need to be developed. In our view, those on which research should be prioritized are linkage errors and predictions based on machine learning algorithms. For both cases, it should be analyzed whether the main tools adopted here, the linearization and decomposition of the GMSE into simpler conditional components, could be successfully applied. Other simple and straight forward extensions could be obtained by considering standard linearization techniques to measure the accuracy of non- linear parameters derived from register microdata, such as correlations, regression parameters or quantiles.

Finally, we stress that another aspect to consider in facilitating the strategy feasibility and its wider applicability is to implement software tools that make it easy to calculate GMSE from the microdata of the register and from the functional form used to build the forecasts.

9. Appendix

9.1. Demonstration of the Equations (13a) and (13b)

To derive Equation (13a), we add and subtract the overall mean, $E_P E_M(\hat{Y}_{R_d}) = \tilde{Y}_{R_d}$ in the expression of GMSE. We have

$$\begin{aligned} GMSE(\hat{Y}_{R_d}) &= E_P E_M(\hat{Y}_{R_d} - E_P E_M(\hat{Y}_{R_d}) + E_P E_M(\hat{Y}_{R_d}) - Y_{U_d})^2 \\ &= E_P E_M(\hat{Y}_{R_d} - E_P E_M(\hat{Y}_{R_d}))^2 + E_P E_M(E_P E_M(\hat{Y}_{R_d}) - Y_{U_d})^2 \\ &\quad + 2E_P E_M[(\hat{Y}_{R_d} - E_P E_M(\hat{Y}_{R_d}))(E_P E_M(\hat{Y}_{R_d}) - Y_{U_d})] \\ &= E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] - V_M(Y_{U_d}), \end{aligned}$$

since, from [Kendall and Stuart \(1976, 196\)](#), it is

$$E_P E_M(\hat{Y}_{R_d} - E_P E_M(\hat{Y}_{R_d}))^2 = E_P[V_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] + V_P[E_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})],$$

and $E_M(\hat{\mathbf{t}}|\boldsymbol{\lambda}) = \boldsymbol{\vartheta}$, then

$$V_P[E_M(\hat{Y}_{R_d}|\boldsymbol{\lambda})] \cong V_P\left[\sum_{k \in R_d} f[\mathbf{x}_k; E_M(\hat{\mathbf{t}}|\boldsymbol{\lambda})]\right] = V_P\left[\sum_{k \in R_d} f(\mathbf{x}_k; \boldsymbol{\vartheta})\right] = 0$$

$$E_P E_M(\tilde{Y}_{R_d} - Y_{U_d})^2 = E_M(\tilde{Y}_{R_d} - Y_{U_d})^2 = V_M(Y_{U_d}) \text{ and}$$

$$E_P E_M[(\hat{Y}_{R_d} - E_P E_M(\hat{Y}_{R_d}))(E_P E_M(\hat{Y}_{R_d}) - Y_{U_d})]$$

$$= \left[(E_P E_M(\hat{Y}_{R_d}))^2 - E_M(Y_{U_d}^2) \right] = -V_M(Y_{U_d}).$$

To demonstrate Equation (13b), consider that

$$\begin{aligned} E_M E_P(\hat{Y}_{R_d} - E_P E_M(\hat{Y}_{R_d}))(E_P E_M(\hat{Y}_{R_d}) - Y_{U_d}) \\ = -E_M[(E_P(\hat{Y}_{R_d}|\mathbf{y})Y_{U_d})] + E_P E_M(\hat{Y}_{R_d}) = -Cov_M(E_P(\hat{Y}_{R_d}|\mathbf{y}), Y_{U_d}). \end{aligned}$$

10. References

- Alleva, G., and F. Petrarca. 2013. "New indicators for investigating the Integration of Sapienza graduates into the labor market." Working papers n. 120/2013 del Dipartimento Memotef, ISSN 2239-608X.
- Alleva, G. 2017. "The new role of sample surveys in official statistics." ITACOSM 2017, The 5th Italian Conference on Survey Methodology, June 14, 2017, Bologna, Italy. Available at: https://www.istat.it/it/files//2015/10/Alleva_ITACOSM_14062017.pdf (accessed May 2021).
- Biemer, P.P. 2010. "Total Survey Error Design, implementation, and evaluation." *Public Opinion Quarterly* 4(5) : 817–848. DOI: <https://doi.org/10.1093/poq/nfq058>.
- Binder, D.A., and Z. Patak. 1994. "Use of estimating functions for estimation from complex surveys." *Journal of the American Statistical Association* 89: 1035–1043. DOI: <https://doi.org/10.1080/01621459.1994.10476839>.
- Breidt, F.J., and J.D. Opsomer. 2017. "Model-Assisted Survey Estimation with Modern Prediction Techniques." *Statistical Science* 32(2) : 190–205. DOI: <https://doi.org/10.1214/16-STS589>.
- Chambers, R.L., and R.G. Clark. 2015. "An Introduction to Model-Based Sampling with Applications." *Oxford Statistical Science*. 37. DOI: <https://doi.org/10.1093/acprof:oso/9780198566625.001.0001>.
- Chen, S., and D. Haziza. 2017. "Multiply robust imputation procedures for the treatment of item nonresponse in surveys." *Biometrika* 102: 439–453. DOI: <https://doi.org/10.1007/s40300-017-0128-9>.
- Citro, C.F. 2014. "From multiple modes for surveys to multiple data sources for estimates." *Survey Methodology*. Statistics Canada. Available at: <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14128-eng.pdf?st=emZzAE9> (accessed May 2021).
- Cochran, W.G. 1977. *Sampling techniques*, (Third edition). New York: Wiley. Available at: https://glad.geog.umd.edu/Potapov/_Library/Cochran_1977_Sampling_Techniques_Third_Edition.pdf (accessed May 2021).
- Deville, J.-C. 1999. "Variance estimation for complex statistics and estimators: Linearization and residual techniques." *Survey Methodology* 25(2) : 193–203. Statistics Canada. Available at: <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1999002/article/4882-eng.pdf?st=qNANeGtP> (accessed May 2021).
- Deville, J.-C., and Y. Tillé. 2005. "Variance approximation under balanced sampling." *Journal of Statistical Planning and Inference* 128(2) : 569–591. Available at: <https://core.ac.uk/download/pdf/43673958.pdf> (accessed May 2021).
- Eurostat. 2019. Available at: https://ec.europa.eu/eurostat/cros/content/essnet-quality-mul-tisource-statistics-komuso_en (accessed May 2021).
- Falorsi, P.D., P. Lavallée, and P. Righi. 2019. "Cost Optimal Sampling for the Integrated Observation of Different Populations." *Survey Methodology*. Available at: <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019003/article/00004-eng.pdf?st=qZIAruhQ> (accessed May 2021).

- FAO. 2014. *Technical Report on the Integrated Survey Framework*, Technical Report Series GO-02- 2014. Available at: http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf (accessed May 2021).
- Graf, M. 2015. *A Simplified Approach to Linearization Variance for Surveys*. Technical Report, Institut De Statistique, Université de Neuchâtel. Available at: <https://doc.rero.ch/record/324723/files/00002754.pdf> (accessed May 2012).
- Gruppo UNI.CO. 2015. *La Domanda di Lavoro per i laureati*. I risultati dell'integrazione tra gli archivi amministrativi dell'Università Sapienza di Roma e del Ministero del Lavoro e delle Politiche Sociali, Edizioni Nuova Cultura- Roma. ISBN 9788868124816. DOI: <https://doi.org/10.4458/4816>.
- Isaki, C., and W.A. Fuller. 1982. "Survey design under the regression superpopulation model." *Journal of the American Statistical Association* 77: 89–96.
- ISCO 1-ISCO 2. Available at: https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/@publ/documents/publication/wcms_172572.pdf
- Istat. 2016. *Istat's Modernisation Programme*. Available at: https://www.istat.it/en/files/2011/04/IstatsModernisationProgramme_EN.pdf (accessed May 2021).
- Kendall, M.G, and A. Stuart. 1976. *The Advanced Theory of Statistics: Design and analysis, and time- series*. Hafner.
- Kim, J.K., and J.N.K. Rao. 2012. "Combining data from two independent surveys: a model-assisted approach." *Biometrika* 99(1) : 85–100. DOI: <https://doi.org/10.1093/biomet/asr063>.
- Nedyalkova, D., and Y. Tillé. 2008. "Optimal sampling and estimation strategies under the linear model." *Biometrika* 95: 521–537. DOI: <https://doi.org/10.1093/biomet/asn027>.
- Nirel, R., and H. Glickman. 2009. "Chapter 21 – Sample Surveys and Censuses." In *Handbook of Statistics*, edited by C.R. Rao.: Elsevier.
- Petrarca, F. 2014a. "Non-metric PLS path modeling: Integration into the labour market of Sapienza graduates." In *Advances in latent variables. Studies in theoretical and applied statistics*: 159–170. Berlin: Springer. DOI: https://doi.org/10.1007/10104_2014_16
- Petrarca, F. 2014b. *Assessing Sapienza University alumni job careers: Enhanced partial least squares latent variable path models for the analysis of the UNI.CO administrative archive*. PhD diss., Dipartimento di Economia dell'Università degli studi Roma Tre. Available at: <http://hdl.handle.net/2307/4167> (accessed May 2021).
- Pfeffermann, D. 2015. "Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture." *Journal of Survey Statistics and Methodology* 3(4) : 425–483. DOI: <https://doi.org/10.1093/jssam/smv035>.
- Righi, P., P.D. Falorsi, S. Daddi, E. Fiorello, P. Massoli, and M.D. Terribili. 2021. "Optimal sampling for the Population Coverage Survey of the new Italian Register Based Census." *Journal of Official Statistics* (September 2021)
- Särndal, C.E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer-Verlag.
- Scholtus, S. 2019. "A bootstrap method for estimators based on combined administrative and survey data." In *NTTS Conference 2019*. Brussels, Belgium. Available at: <https://slidetodoc.com/download.php?id=4386>.

- Statistics Canada. 2009. *Statistics Canada Quality Guidelines*, (6th edition). Available at: <https://www150.statcan.gc.ca/n1/en/pub/12-539-x/12-539-x2019001-eng.pdf?st=y2AqFuiY>.
- Wallgren, A., and B. Wallgren. 2014. *Register-based Statistics: statistical methods for administrative data*, (2nd Edition). Chichester: Wiley.
- Wolter, K.M. 1985. *Introduction to Variance Estimation*. New York: Springer.
- Wolter, K.M. 1986. “Some Coverage Error Models for Census Data.” *Journal of the American Statistical Association* 81: 338–346. DOI: <https://doi.org/10.2307/2289222>.
- Vaillant, R. 2009. “Model based predictions of finite population totals.” In *Chapter 23 in Handbook of statistics 29: Design, Methods and Applications*, edited by P. Pfefferman and C.R. Rao. Amsterdam: North Holland.
- Vallée, A.A., and Y. Tillé. 2019. “Linearisation for Variance Estimation by Means of Sampling Indicators: Application to Non-response.” *International Statistical Review* 0(0) : 1–21. DOI: <https://doi.org/10.1111/insr.12313>.
- Ziegler, A. 2015. *Generalized Estimating Expressions*, Springer and Verlag. Lecture Notes in statistics.

Received May 2019

Revised August 2020

Accepted January 2021

A Hybrid Technique for the Multiple Imputation of Survey Data

Humera Razzak¹ and Christian Heumann¹

Most of the background variables in MICS (Multiple Indicator Cluster Surveys) are categorical with many categories. Like many other survey data, the MICS 2014 women's data suffers from a large number of missing values. Additionally, complex dependencies may be existent among a large number of categorical variables in such surveys. The most commonly used parametric multiple imputation (MI) approaches based on log linear models or chained Equations (MICE) become problematic in these situations and often the implemented algorithms fail. On the other hand, nonparametric MI techniques based on Bayesian latent class models worked very well if only categorical variables are considered. This article describes how chained equations MI for continuous variables can be made dependent on categorical variables which have been imputed beforehand by using latent class models. Root mean square errors (RMSEs) and coverage rates of 95% confidence intervals (CI) for generalized linear models (GLM's) with binary response are estimated in a simulation study and a comparison is made among proposed and various existing MI methods. The proposed method outperforms the MICE algorithms in most of the cases with less computational time. The results obtained by the simulation study are supported by a real data example.

Key words: Complex dependencies; MICE; multiple indicator cluster surveys.

1. Introduction

Information on many variables (e.g., mortality, nutrition, child health and women reproductive health, and so on) is collected in different large-scale surveys. Multiple Indicator Cluster Surveys or MICS is such a large-scale complex survey which provides opportunities to fill data gaps for monitoring the health situation of children and women in under developed countries. Information based on background variables of the indicators in MICS are very important for data analysis, and for policy making (Corsi et al. 2017). However, the problem of missing data is inevitable in such studies. For example, the data set of individual women from MICS 2014 Punjab, which has been used in the survey data example latter, has between 14 to 95 percentages of data missing on more than 200 women's background characteristics (e.g., demographics, age, education, motherhood and recent births and so on). Whereas, only few variables are completely observed. Respondents feeling shy to answer sexual activity related questions etc. may result in missing data problem.

¹ Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstraße. 33, D80539 Munich, Germany. Emails: ali.arif89@yahoo.com, chris@stat.uni-muenchen.de.

Acknowledgments: The authors thank the editor and two referees for their suggestions, which greatly helped improve the article.

1.1. Statement of the Problem

The representativeness of the sample can be reduced and inferences about the population can be distorted due to missing values. Moreover, ignoring missing data can lead to a bias of unknown direction and magnitude in the estimated parameters. Therefore, it is critical to impute the data, which usually provides more accurate inference compared to ad-hoc methods (e.g., complete case (CC) analysis or single imputation) even when variables are missing at random (MAR) (Abdella and Marwala 2005; Little and Rubin 2002). Multiple imputation (MI) was originally introduced to handle nonresponse in public use data files or shared databases (Rubin 1987).

Usually, household survey data sets contain mixed type variables (i.e., both continuous and categorical variables with many categories). Additionally, complex dependencies may be existent among a large number of categorical variables in such surveys. Often the number of continuous variables is less as compared to categorical variables in such studies (Chandra et al. 2005; Gulliford et al. 1999). The most commonly used parametric MI approaches based on log linear models or chained Equations (MICE) (Raghunathan et al. 2001; Van Buuren 2007) become problematic in these situations and often the implemented algorithms fail. On the other hand, nonparametric MI techniques based on Bayesian latent class models have been seen working very well if only categorical variables are considered. Therefore, one is limited in the choice of MI methods, that is, either; one has to sacrifice continuous variables in the analysis (or categorize them) or has to sacrifice categorical variables if many categorical variables are involved. Due to certain limitations, various MI approaches cannot be used together without correct modifications. These limitations create serious problems for researchers to obtain complete data sets with mixed type variables. Our aim is many fold. Firstly, to develop methods for imputing mixed type data from large scale complex surveys. Secondly, to avoid difficulties of complicated models in high dimensions. Thirdly, to combine existing techniques to handle incomplete large scale complex data sets. Lastly, to gain computational efficiency.

An easy to implement hybrid MI technique is proposed in this article which describes how different model-based imputation approaches can be blended together to impute mixed type data having complex dependency structures. The remainder of this article is organized as follows. In Section 2, we describe methodological background of existing approaches to deal missing data problem and their shortcomings in high dimensions. In Section 3, combining rules for imputed data sets are described. The hybrid algorithm is described in Section 4. Section 5 compares the performance of different imputation methods in simulation studies. In Section 6, the proposed method is applied to a survey data set. Results are discussed in Section 7. Concluding remarks are given at the end.

2. Methodological Background

Missing data problem arises, for example, due to item nonresponse (INR) or entry errors and so on. Besides INR, general reasons for the missing data sets may include system failures. There are three missing data mechanisms. Missing values in any data can be missing completely at random (MCAR), or missing at random (MAR), or missing not at random (MNAR) (Rubin 1987; Little and Rubin 2002). In MCAR, the probability of missing data on a variable is not associated to itself and or other measured variables. In

MAR, the probability of missing depends on other, observed, variables. Finally, data are MNAR if the probability of missing depends on the variable value itself. Practically all methods implemented in software assume MAR. MNAR is called “non-ignorable” if the parameters driving the missing data process and the parameters driving the data generating process are distinct (or independent in a Bayesian analysis), but this is not further considered in the article. Exact missing data mechanisms are often unknown when dealing with large scale data sets. Therefore, most of the time, certain assumptions are made accordingly. Li et al. (2012) address some problems with missing large data. Little’s MCAR test proposed by Little (1988) is used commonly for testing missing data being MCAR.

2.1. Review of Existing Approaches

The CC analysis sacrifices all units where at least the value of one variable is missing. Such methods are still very popular in psychological research (Schlomer et al. 2010). In a regression setting, if the probability that a covariate is missing, depends on the response variable, the CC analysis (listwise deletion) can lead to biased estimates (Little and Rubin 2002). The CC method also results in a loss of power, which can make the analysis inefficient (Little and Rubin 2002). Despite of being the worst available methods (Wilkinson and Task Force on Statistical Inference 1999), CC is still the most applied technique due to the simplicity and availability as default options in statistical software packages (Van Ginkel 2007). The hot-deck method is another approach and belongs to the family of a single-imputation approach. This method replaces missing values with values from a “similar” responding unit (Andridge and Little 2010) and the empirical distribution obtained is used to draw the imputed values. In the case that the entire sample of respondents is being used as a single donor pool, this method produces consistent and unbiased estimates for missing completely at random (MCAR) data (Rubin 1976; Little and Rubin 2002). This method uses covariate information, avoids strong parametric assumptions and requires no careful modelling to develop selection criteria for imputing a value because it does not have any parametric model (Schafer and Graham 2002). However, the problem with this method is that it lacks the clear criteria to guide the selection of the donor set of complete cases (Pérez et al. 2002). Bayesian bootstrap (Rubin 1987) is a useful alternative when standard hot-deck becomes unsuitable to impute in the presence of a large number of variables (Andridge and Little 2017). Other proposed methods for missing data use various statistical methods including self-organizing maps (SOM) (Kohonen 1995; Oja and Kaski 1999), k-nearest neighbour (kNN) (Batista and Monard 2003), multi-layer perceptron (Sharpe and Solly 1995), recurrent neural networks (Bengio and Gingras 1995). Auto-associative neural network imputations, with genetic algorithms are proposed by Pyle (1999), Narayanan et al. (2002), Chung and Merat (1996), Marseguerra and Zoia (2005) and Marwala and Chakraverty (2006) also implement some of the well-known methods used for handling missing data. Multi-task learning approaches are some other techniques based on machine learning methods (Ankaiah and Ravi 2011).. According to the studies of Horton and Kleinman (2007), Honaker et al. (2011), Royston and White (2011) and Van Buuren and Groothuis-Oudshoorn (2011), over the last three decades a wide range of variety and settings of MI techniques has been

introduced for catering missing data problem in different research areas (Abdella and Marwala 2005; Honaker et al. 2011; Little and Rubin 2002; Schafer and Graham 2002). MI, likelihood based analysis, and weighting approaches are alternatives to listwise and pairwise deletion methods. These methods usually make the assumption that the missing data is missing at random (MAR), hence making the estimates unbiased, consistent, and asymptotically normal (Allison 2002; Barnard and Meng 1999; Roth 1994; Schafer and Graham, 2002) if that assumption holds.

Model-based imputation is currently considered the most popular method of addressing missing data problems. The true complete-data distribution and the missing-data mechanism form the basis of the imputation model which can be explicit or implicit by nature (Rubin 1987). Indicators in survey data sets are mostly categorical. Schafer (1997) describes that MI with log-linear models can be used to generate imputed values for such indicators by capturing the associations in the joint distribution. A severe restriction is that the number of variables must in general be small (Vermunt et al. 2008). A brief description of model-based imputation methods and their shortcomings in high dimensions are given the following sections.

2.2. *The Fully Conditional Specification (FCS)*

The fully conditional specification (FCS) (Van Buuren 2007), also known as MI by chained Equations (MICE) (Raghunathan et al. 2001; Van Buuren 2007) is a model-based imputation method. Missing values are sequentially imputed by estimating a series of univariate conditional models. Normal regressions and logistic or multinomial logistic regressions are used for continuous and categorical dependent variables, respectively. Alternatively, a method called predictive mean matching (PMM) can be used. Newer implementations also allow classification and regression trees (CART). MICE is an iterative method and imputes missing values variable by variable. It uses the current regression estimates for the response variable, where the response variable in this context is the actual target variable in the iterative process for which missing values are imputed. Draws from the posterior predictive distribution of the unobserved data given the observed data can be used to impute missing values. M imputed data sets are created by repeating this process (Rubin 1987). By conducting the analysis on each of these data sets, the resulting M point and M variance estimates are then combined by a set of rules (Rubin 1987). MICE assumes that equivalent, or at least nearly as good, draws for the joint distribution of the variables can be approximated by the sequential draws from the univariate conditional models. Missing values in continuous variables are often treated using a multivariate normal MI. These models are often robust to departure from normality by nature (Graham and Schafer 1999; Schafer 1997). There are three main limitations or difficulties in the implementation of MICE. First, the possible lack of compatibility among the set of univariate conditional regression models and the joint distribution of the variables being imputed (Arnold and Press 1989; Gelman and Speed 1993). Although an algorithm is proposed which selects the sequence of regression models such that they are assumed to be a good fit for the data, but it is very complicated to establish exact conditions for convergence (Zhu and Raghunathan 2016). Second, the risk of overlooking higher order interactions arises when MICE includes only the main effects in the univariate conditional regression models, although using CART

may resolve this problem. Third, the procedure is very time consuming when higher-order interactions are included parametrically in the model (Vermunt et al. 2008). Since categorical variables are internally represented as dummy variables which could easily double the actual number of predictors, the implementation of the FCS MI by chained equations algorithm becomes extremely slow or difficult in the presence of categorical variables with missing values. The R package “mice” by Van Buuren and Groothuis-Oudshoorn (2011) implements MI by chained equations.

2.3. Fully Bayesian Joint Modelling (JM) and Latent Class (LC) Imputation

The joint modelling (JM) MI involves specifying a multivariate distribution for the data and draws imputations from their conditional distributions by Markov Chain Monte Carlo (MCMC) methods. Latent Class MI is a JM technique which imputes missing categorical data in the presence of a large number of observed categorical variables (Vermunt et al. 2008). There are two main advantages of the Latent Class (LC) analysis. Firstly, the variation in observed variables can be explained by using a small number of latent classes. Secondly, it is directly applicable in the presence of many categorical variables. These advantages makes LC analysis a natural choice to apply in the missing categorical data setting. Vermunt et al. (2008) show that LC MI outperforms log-linear imputation however, only limited assessments of the method in terms of missing data mechanism scenarios are available. To understand the JM approach multiply impute large categorical data sets, it is important to understand few details regarding how mixture models are used for density estimation and MI.

The distribution of categorical data can be described by a mixture model known as Latent Class Model or LCM (Lazarsfeld 1950). Mixture models are considered as flexible tools which model the association structure of a set of variables (their joint density) by utilizing a finite mixture of simpler densities (McLachlan and Peel 2000). The probability of having a specific response pattern is defined by each mixture component in a Latent Class Analysis (LCA). A weighted average of the class-specific densities generates the estimated overall density. As described by Lazarsfeld (1950), the scores of different items are independent of each other within latent classes due to local independence assumptions in LCA. A brief introduction to the mathematical form of a LCM as a tool for density estimation is given in the following lines.

Let y_{ij} be the score of the i_{th} person on the j_{th} categorical item belonging to an $n \times J$ data-matrix Y ($i = 1, \dots, n, j = 1, \dots, J$), y_i the J -dimensional vector with all scores of person i , and x_i a discrete (unobserved) latent variable with K categories. In the LCM, the joint density $P(y_i; \boldsymbol{\pi})$ has the following form:

$$\begin{aligned}
 P(y_i; \boldsymbol{\pi}) &= \sum_{k=1}^K P(x_i = k; \boldsymbol{\pi}_x) P(y_i|x_i = k; \boldsymbol{\pi}_y) \\
 &= \sum_{k=1}^K P(x_i = k; \boldsymbol{\pi}_x) \prod_{j=1}^J P(y_{ij}|x_i = k; \boldsymbol{\pi}_{yj})
 \end{aligned}
 \tag{1}$$

where $\boldsymbol{\pi} = (\boldsymbol{\pi}_x, \boldsymbol{\pi}_y)$ is a set of LCM parameters which can be partitioned into two parts. The first part contains the latent class proportions ($\boldsymbol{\pi}_x$) and the second contains class-specific item response probabilities ($\boldsymbol{\pi}_y$). A separate set of parameters for each of the J items ($\boldsymbol{\pi}_{yj}$) is assigned to the second part. Due to the fact that a mixture distribution is used, a weighted

sum of the K class-specific multinomial densities $P(y_i|x_i = k; \pi_y)$ generates the overall density. In this generation, the latent proportions are used as weights. From (1) it can be seen that the product over the J independent multinomial distributions (conditional on the k -th latent class) makes use of the local independence assumption. The first, second, and higher-order moments of the J response variables can be captured in LC models by setting the number of latent classes large enough (McLachlan and Peel 2000). The generated higher-order moments are actually the univariate margins, bivariate associations, and higher-order interactions when dealing with categorical variables (Vermunt et al. 2008). The unit's posterior class membership probabilities, that is, the probability that a unit belongs to the k -th class given the observed data pattern y_i , is the quantity of interest when using LC models. According to the theorem of Bayes we can define this quantity as follows:

$$P(x_i = k|y_i; \boldsymbol{\pi}) = \frac{P(x_i = k; \boldsymbol{\pi}_x)P(y_i|x_i = k; \boldsymbol{\pi}_y)}{P(y_i; \boldsymbol{\pi})}. \quad (2)$$

A fully Bayesian JM approach is proposed by Si and Reiter (2013) which is called Dirichlet process mixture of products of multinomial distributions (DPMPM). This approach uses nonparametric Bayesian versions of Latent Class Models (LCM) (Lazarsfeld 1950) to multiply impute high-dimensional categorical data (Vermunt et al. 2008). DPMPM approach has two stages. In stage one, a mixture of independent multinomial distributions is modelled for a contingency table of the categorical variables. In the second stage, the mixture distributions are estimated non-parametrically with Dirichlet process prior distributions. Arbitrarily complex dependencies can be described by such mixtures of multinomials. Since the computation of these dependencies is practical and generally easy, they can serve as an effective general purpose MI engine. These models have been successfully used to impute missing values in up to 80 categorical variables (Si and Reiter 2013). Murray and Reiter (2016) have also worked on combining Dirichlet process mixtures of multinomial and multivariate normal distributions for categorical and continuous variables, but this approach involves complicated models to create the dependence structure between the continuous and the categorical variables. The DPMPM approach by Dunson and Xing (2009) is described as:

1. Assume that each individual i belongs to exactly one of $K < \infty$ latent classes,
2. For $i = 1, \dots, n$, let $x_i \in \{1, \dots, k\}$ indicate the class of individual i , and let $\pi_k = P(x_i = k)$. Assume further, that $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_\infty\}$ is the same for all individuals. Within any class, we suppose that each of the j variables independently follows a class-specific multinomial distribution i.e., for any value $y_j \in \{1, \dots, d_j\}$ let $\mathbb{Y}_{kj}^{(j)} = P(y_{ij} = y_j|x_i = k)$.

Here d_j is the the total number of categories for the variable j .

Mathematically expressing the finite mixture model as:

$$y_{ij}|x_i, \mathbb{Y} \sim \text{Multinomial}(\mathbb{Y}_{x_i d_j}^{(j)}, \dots, \mathbb{Y}_{x_i d_j}^{(j)}) \text{ for all } i \text{ and } j \quad (3)$$

$$x_i|\boldsymbol{\pi} \sim \text{Multinomial}(\pi_1, \dots, \pi_\infty) \text{ for all } i. \quad (4)$$

For prior distributions on \mathbb{Y} and $\boldsymbol{\pi}$, we have

$$\pi_k = V_k \left(\prod_{l < k} 1 - V_l \right) \text{ For } k = 1, \dots, \infty$$

$$V_k \overset{\sim}{\text{iid}} \text{Beta}(1, \alpha)$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$$

$$\forall_{kj} \sim \text{Dirichlet}(a_{j1}, \dots, a_{jd_j})$$

Every V_k is drawn from a Beta distribution with parameters $(1, \alpha)$, where, α , the concentration parameter of the process, vary according to a Gamma distribution with parameters (a_α, b_α) . The conditional responses (and their prior) have multinomial densities with Dirichlet priors.

The R ([R Core Team, 2018](#)) package “NPBayesImputeCat” by [Quanli et al. \(2018\)](#) is a tool for non-parametric Bayesian JM MI, but the implementation of this package is restricted to categorical variables.

3. Rubin’s Inference

In order to incorporate the uncertainty introduced by missing data and the imputations into the inferences, the estimates for quantities of interest obtained by analyzing each completed data set are combined by utilizing rules proposed by [Rubin \(1987\)](#). Let Q be any quantity of interest (e.g., a population proportion or a probability or a regression coefficient). For $m = 1, \dots, M$, let $q^{(m)}$ and $u^{(m)}$ be respectively the point estimates of Q in the m -th imputed data set with variance estimates $q^{(m)}$. Valid inferences for a scalar Q by combining the $q^{(m)}$ and $u^{(m)}$ according to [Rubin \(1987\)](#) are obtained as follows:

$$\bar{q}_M = \sum_{m=1}^M \frac{q^{(m)}}{M} \tag{5}$$

$$b_M = \sum_{m=1}^M \frac{(q^{(m)} - \bar{q}_M)^2}{M - 1} \tag{6}$$

$$\bar{u}_M = \sum_{m=1}^M \frac{u^{(m)}}{M} \tag{7}$$

\bar{q}_M can be used to estimate Q and the variance of \bar{q}_M can be estimated by

$$T_M = \left(1 + \frac{1}{M}\right)b_M + \bar{u}_M \tag{8}$$

with degrees of freedom $\nu_M = (M - 1) \left(1 + \frac{\bar{u}_M}{\left(\left(1 + \frac{1}{M}\right)b_M\right)^2}\right)$. (9)

Confidence intervals can be constructed using standard multiple imputation confidence interval construction rules, which approximately follows a t-distribution. (For more detail see [Rubin 1996](#); [Barnard and Meng 1999](#); [Reiter et al. 2006](#); [Harel and Zhou 2007](#)).

4. Proposed Hybrid Architecture

Since the application of the package “NPBayesImputeCat” (Quanli et al. 2018) is limited to only categorical variables, the incomplete data set is proposed to be portioned into two sets, one consisting of categorical variables ($Miss_{cat}$), (which MICE may not be able to impute due to reasons described in the introduction) and other consisting of continuous variables ($Miss_{num}$) where variables may be missing in both sets. A fully Bayesian JM (DPMPM) approach is used to fill in missing values by utilizing the package “NPBayesImputeCat” in $Miss_{cat}$. This results in a complete version (Imp_{cat}) of categorical variables independent of information available in the continuous variables. This complete version (Imp_{cat}) of categorical variables can be used by MICE to construct chained equations based on categorical variables which have already been imputed by the fully Bayesian joint models to now impute the continuous variables. To achieve this, the data set ($Miss_{num}$) is added to the data set (Imp_{cat}) and MICE is run. This provides one completely imputed data set where the imputations of the continuous variables obtained by FCS using chained equations depend on the information available in the imputed categorical variables. This process is repeated M times to obtain multiple imputed data sets using different algorithms offered by the R package “mice” (Van Buuren and Groothuis-Oudshoorn, 2011) along with some prior specifications and a number of mixture components used in the R package “NPBayesImputeCat” (Quanli et al. 2018). Algorithm 1 explains the proposed hybrid architecture in detail.

Algorithm 1. Proposed hybrid architecture

Require: P $n \times p$ matrix with incomplete data

1. $Miss_{cat}, Miss_{num} \leftarrow$ Initial division of p variables into factor and numeric subsets.
 2. **for** $z = 1, \dots, Z$ **do**
 3. **for** $m = 1, \dots, M$ **do**
 4. $Imp_{cat_m}^z \leftarrow$ Imputing $Miss_{cat}$ using R package “NPBayesImputeCat”.
 5. $Imp_{cat_m}^z, Miss_{num_m}^z \leftarrow$ Combining $Imp_{cat_m}^z$ and $Miss_{num_m}^z$ to generate partially imputed data set.
 6. $Imp_m^z \leftarrow$ Imputing $Imp_{cat_m}^z, Miss_{num_m}^z$ using R package “mice” i.e., $f(Miss_{num_m}^z | Imp_{cat_m}^z)$
 7. $Imp_m^z \leftarrow$ Final imputed data set.
 8. $\bar{q}_M^z \leftarrow \sum_{m=1}^M \frac{q^{(m)}}{M}$ Pooled point estimates¹.
 9. $b_M^z \leftarrow \sum_{m=1}^M \frac{(q^{(m)} - \bar{q}_M^z)^2}{M-1}$
 10. $\bar{u}_M^z \leftarrow \sum_{m=1}^M \frac{u^{(m)}}{M}$
 11. $T_M^z \leftarrow (1 + \frac{1}{M})b_M^z + \bar{u}_M^z$ Pooled variances².
 12. **end for**
 13. $\bar{q} \leftarrow \sum_{z=1}^Z \frac{\bar{q}_M^z}{Z}$ Average of pooled point estimate³.
 14. $\bar{T} \leftarrow \sum_{z=1}^Z \frac{T_M^z}{Z}$ Average of pooled variance⁴.
-
- end for**

1: \bar{q}_M^z are pooled point estimates over M imputed data sets across z simulations.

2: T_M^z are pooled variances over M imputed data sets across z simulations.

3: \bar{q} is an average of pooled point estimates ($\bar{q}^{(z)}$) across z simulations.

4: \bar{T} is an average of pooled variances ($T^{(z)}$) across z simulations.

5. Simulation Studies

Simulation studies are conducted to examine the impact of MI by our proposed method. The incomplete data is generated as MAR with (known) effect and the number of categorical variables is kept more than the number of continuous variables, aiming to compare strategies in a realistic data situation.

We generate a sample of size $n = \{1,000\}$ for five $(X_1, X_2, X_3, X_4, X_5)$ dimensional correlated random covariates from a multivariate normal distribution MVN. The marginal distributions of X_1, X_2, X_3, X_4, X_5 are normal and we set the mean and variance of each variable to 0 and 0.5 respectively. The correlation structure is given as:

$$R = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix},$$

where $\rho = 0.5$. The following component-wise threshold is used to transform random covariates into binary values.

$$X_{ij} = \begin{cases} 0 & \text{if } X_{ij} \leq 0.5, \\ 1 & \text{if } X_{ij} > 0.5, \end{cases}$$

where $i = 1, \dots, 5$ and $j = 1, \dots, 1,000$.

Outcomes for two continuous covariates X_6 and X_7 are generated from normal distributions (ND) i.e.,

$$X_6 \sim N(\mu_6; \sqrt{2}), \tag{10}$$

$$X_7 \sim N(\mu_7; \sqrt{2}). \tag{11}$$

Where μ_6 and μ_7 are two random vectors of means specified for the ND by using by the following relations.

$$\mu_6 = -0.2X_1 - 0.3X_2 + 0.5X_3 - 0.2X_4 + 0.22X_5, \tag{12}$$

$$\mu_7 = -2 + \mu_6. \tag{13}$$

To be noted here, μ_6 assumes different values for each sample units according the different values of X_{ij} . Similarly μ_7 assumes different values for each sample units according the different values of μ_6 . As a consequence μ_6 and μ_7 are random vectors of $j = 1, \dots, 1,000$ values. We generate X_8 from Bernoulli distributions with probabilities governed by the logistic regression with

$$\begin{aligned} \text{logit } P(X_8) = & -3 + 1.5X_1 - 2.15X_2 + 2.25X_3 + 1.6X_4 - 1.88X_5 + 1.11X_6 \\ & - 0.96X_2X_3 + 2.3X_1X_3 + 0.5X_2X_6 - 2X_5X_6 + 1.21X_1X_5 \\ & - 2.7X_1X_2 + 1.2X_1X_2X_3 + 3X_6X_7. \end{aligned} \tag{14}$$

A covariate dependent binary response y is generated from Bernoulli distributions with probabilities governed by the logistic regression with

$$\text{logit } P(y) = 0.2 - 0.1X_1 - 0.1X_2 - 0.1X_3 + 0.3X_4 - 0.5X_5 + 0.2X_6 - 0.1X_7 - 0.1X_8$$

and $\beta_{true} = (0.2; -0.1; -0.1; -0.1; 0.3; -0.5; 0.2; -0.1; -0.1)$. (15)

We suppose that values in all covariates are missing at random with the following probabilities

$$p = I \frac{e^{(-\tau - X_7)}}{(1 + e^{(-\tau - X_7)})}, \quad (16)$$

where τ is a constant. The probabilities defined above yield about 10% to 15 % of the observations in X_i to be missing (at random) for $\tau = -1.5$ and $\tau = -0.5$ respectively. We repeat the process 1,000 times, each time generating new binary response variables and new missing patterns.

We use three purely MICE based MI methods, namely classification and regression trees (MICE-CART) (Breiman 2001), predictive mean matching (MICE-PMM) (Morris et al. 2014) and the Default (MICE-DEF) which uses logistic models for categorical and PMM for continuous variables. We use two Hybrid Multiple Imputation (HMI) methods, for example H.CART and H.DEF depending on various combinations with MICE algorithms (Default and CART) and different tuning parameters (a_α , b_α ; k). Since Si and Reiter (2013) did not observe noticeable differences in the posterior distributions of θ for higher values of prior specifications, we set relatively small prior specification values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$). Akande et al. (2017) suggest that the latent classes (k) less than 40 can appear sufficiently large based on tuning with initial runs. However, we follow Dunson and Xing (2009) who suggest that large enough k can make the latent class model consistent for any joint probability distribution in case of dependencies among variable. Therefore, we further define H.CART₁, H.CART₂, H.CART₃ and H.DEF₁, H.DEF₂, H.DEF₃ to compare the performance of H.CART and H.DEF methods for three sufficiently large number of latent classes (i.e., $k = 80, 150, 400$) and run each MCMC chain for 1,000 iterations using the first 200 as burn-in. The R (R Core Team, 2018) version 3.0.1 is used to perform all calculations. The packages “mice” (Van Buuren and Groothuis-Oudshoorn, 2011), version 2.17 and “NPBayesImputeCat” (Quanli et al., 2018), version 0.1 are used to perform MICE for continuous data and non-parametric Bayesian MI for categorical variables, respectively. According to Van Buuren (2012) a small number of imputations may be created in the beginning when building the imputation model with an exploratory analysis, and increase M gradually for the final analysis. Therefore, the number of multiple imputations ($M = 10$) is large in order to get better estimates of standard errors. Even a higher number of M would have been desirable but would have led to further increased computing times. However, three sets of $M = 10$ imputed data sets are generated using MICE methods, that is, MICE-PMM, MICE-DEF and MICE-CART. Similarly six sets of ($M = 10$) imputed data sets are generated using H.CART₁, H.CART₂, H.CART₃, H.DEF₁, H.DEF₂ and H.DEF₃. In order to achieve convergence and estimates from simulations in a reasonable time, a Gibbs sampler with 100 Markov-Chain-Monte-Carlo

(MCMC) iterates is used. Two hundred iterations are run to insure convergence and to have the results of the simulations in a reasonable time when using the HMI methods.

5.1. Evaluation Criteria

The quality of MI methods is evaluated based on two error-based measurements i.e., root mean square error (RMSE) and empirical standard error (ESE) (Akande et al. 2017; Armina et al. 2017). RMSE is computed as a combination of the bias and variance of the estimate (Burton et.al. 2006). ESE can be considered to access the between imputation variations. The smaller values for RMSE and ESE indicate better performance (Oba et al. 2003). RMSE and ESE are calculated using the following formulas:

$$\text{Root mean square error (RMSE}_{\bar{q}_m}) = \sqrt{\frac{\sum_{z=1}^Z (\bar{q}_M^z - \beta)^2}{Z}}, \quad (17)$$

$$\text{Empirical standard errors (ESE}_{\bar{q}_m}) = \sqrt{\frac{\sum_{z=1}^Z (\bar{q}_M^z - \bar{q})^2}{Z}}, \quad (18)$$

where \bar{q}_M^z denotes the estimated parameter pooled over M imputed data sets in simulation run number z and β denotes the original parameter. The arithmetic mean of \bar{q}_M^z across all $z = \{1, \dots, Z\}$ simulations is presented by \bar{q} . The value for \bar{q} is calculated as follow:

$$\bar{q} = \sum_{z=1}^Z \frac{\bar{q}_M^z}{Z}. \quad (19)$$

The average value of the standard errors ($\sqrt{T_M}$) across all $z = \{1, \dots, Z\}$ simulations can also be used to measure the performance for the MI methods. It can be calculated as:

$$\sqrt{T} = \sum_{z=1}^Z \frac{\sqrt{T_M^z}}{Z}. \quad (20)$$

The amount of bias, which shows how good our estimator is in estimating the real value, can be calculated by a simple difference that is,

$$\text{Bias} = \text{RMSE} - \text{ESE} \quad (21)$$

5.2. Results and Discussions

As discussed earlier, we used two HMI methods that is, (“H.CART” and “H.DEF”) for comparison with three MICE based MI methods, that is, (“MICE.DEF”, “MICE.CART” and “MICE.PMM”). We generated simulations for two missing rates that is, 10% and 15% of values MAR. Simulated root mean square errors (RMSEs), empirical standard errors (ESEs) and coverage rates of 95% confidence intervals for all regression coefficients are estimated via combining rules described above and a comparison is made among the proposed and various existing MI methods. Tables 1–2 display the coverage rates of 95% confidence intervals (CI) and RMSE (ESE) and estimated bias for the 10% and 15% MAR data sets, across 1,000 simulations respectively. Figures 1–2 and Figures 3–4 show

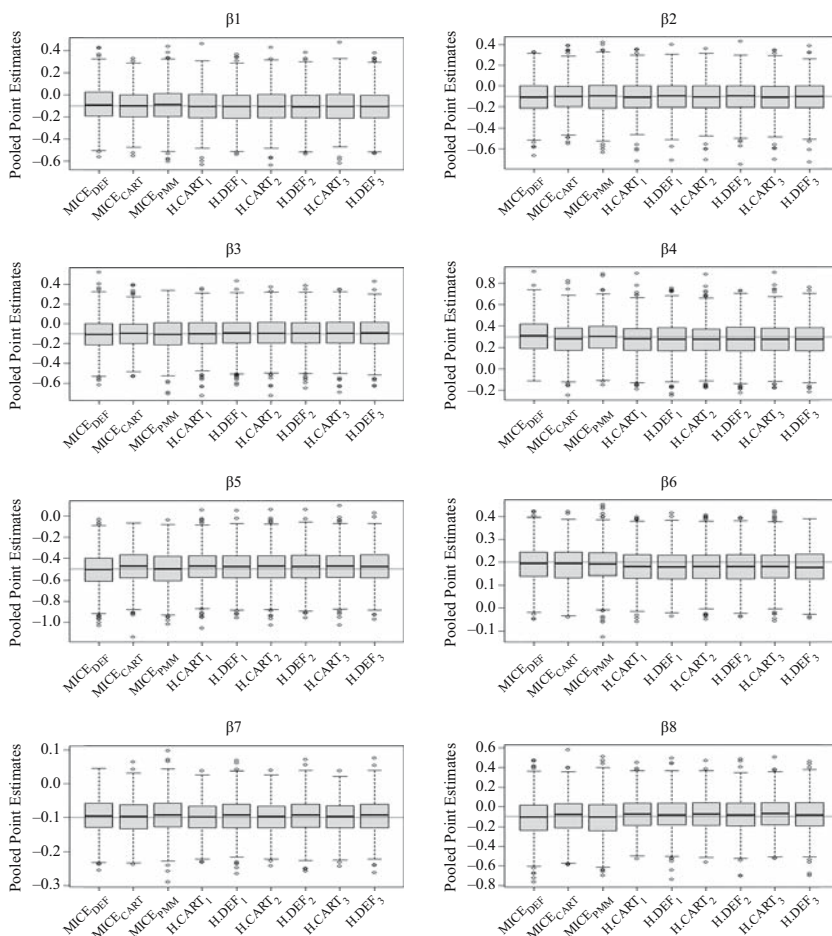


Fig. 1. Simulated data: Boxplots of the pooled point estimates for eight regression coefficients (i.e., β_1, \dots, β_8) under 10% MAR over 1000 simulations. Within each plot the methods are: $MICE_{DEF}$ (default), $MICE_{CART}$ (classification and regression trees), $MICE_{PMM}$ (predictive mean matching), $CART_1$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 80$) and CART), $H.CART_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 150$) and CART), $H.CART_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 400$) and CART), $H.DEF_1$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 80$) and default), $H.DEF_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 150$) and default), $H.DEF_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 400$) and default).

boxplots of the pooled point estimates and standard errors for 10% and 15% MAR data sets, across 1,000 simulations respectively.

Overall, “ $MICE_{DEF}$ ” tends to result in the all mean coverage rates concentrated around 95% for 10% and 15% MAR. Mean coverage rates for “ $MICE_{PMM}$ ” tend to be concentrated around 95% for 10% and fewest high rates (i.e., 97% for X_1 and X_8) under 15% MAR. “ $MICE_{CART}$ ” results in coverage rates above 95% for most of the covariates under 10% MAR. Sometimes it reaches very high rates for categorical covariates (i.e., 98%, 97%, 97% for X_1 , X_3 and X_8 respectively) under 15% MAR, except one binary covariate (X_4) where it

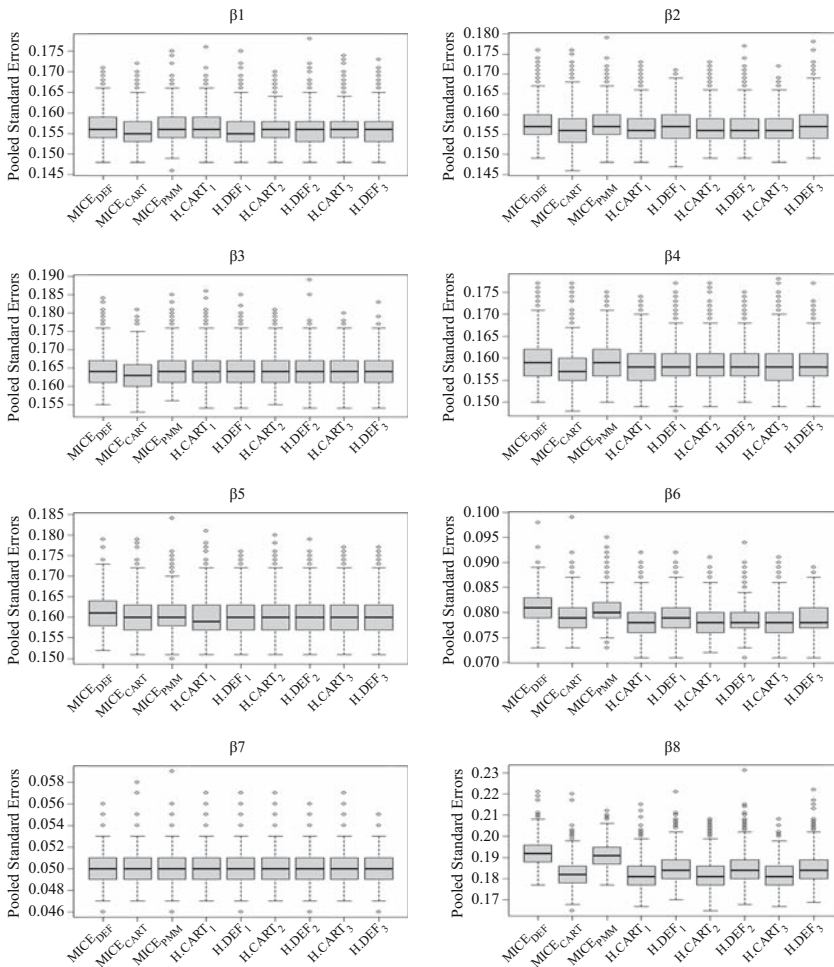


Fig. 2. Simulated data: Boxplots of the pooled standard errors for eight regression coefficients (i.e., β_1, \dots, β_8) under 10% MAR over 1000 simulations. Within each plot the methods are: $MICE_{DEF}$ (default), $MICE_{CART}$ (classification and regression trees), $MICE_{PMM}$ (predictive mean matching), $CART_1$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 80$) and CART), $H.CART_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 150$ and CART), $H.CART_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 400$) and CART), $H.DEF_1$ (combination of and prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 80$) and default), $H.DEF_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 150$) and default), $H.DEF_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 400$) and default).

reaches low rate (i.e., 94%). Mean coverage rates for “H.DEF₁, H.DEF₂, H.DEF₃” tend to be concentrated around 95% for most of the covariates under 10% MAR except a binary covariate X_8 where it reaches high rates under H.DEF₂ and H.DEF₃ (i.e., 97%). Under 15% MAR “H.DEF₁”, “H.DEF₂”, “H.DEF₃” tend to result in the coverage rates above 95% for most of the covariates. Mean coverage rates for the binary covariate X_8 under “H.DEF₂”, “H.DEF₃” are equal to “MICE.PMM” and “MICE.CART” and comparable to that of “H.DEF₁”. This may imply that larger values for k have effect on the overall performance of

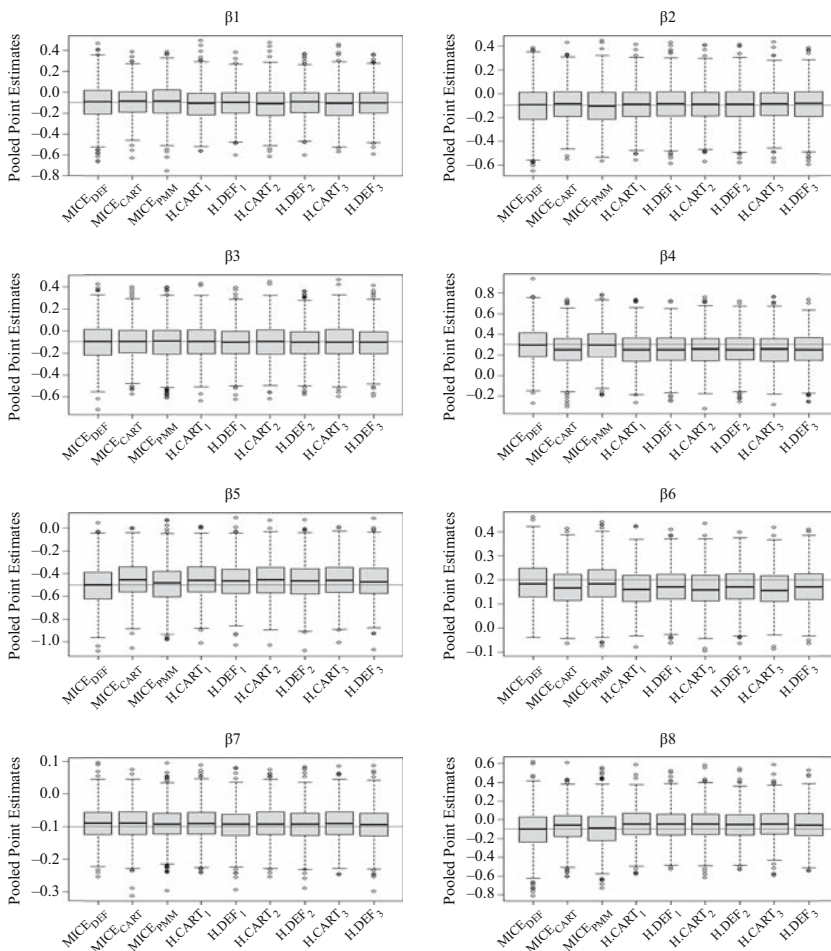


Fig. 3. Simulated data: Boxplots of the pooled point estimates for eight regression coefficients (i.e., β_1, \dots, β_8) under 15% MAR over 1000 simulations. Within each plot the methods are: $MICE_{DEF}$ (default), $MICE_{CART}$ (classification and regression trees), $MICE_{PMM}$ (predictive mean matching), $CART_1$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 80$) and CART), $H.CART_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 150$ and CART), $H.CART_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 400$) and CART), $H.DEF_1$ (combination of and prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 80$) and default), $H.DEF_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 150$) and default), $H.DEF_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 400$) and default).

“H.DEF” MI methods. We observe that all three MI methods based on “H.CART” result in the mean coverage rates for all covariates that are concentrated slightly above 95% under 10% and 15% MAR, but the results are comparable to that of “ $MICE_{CART}$ ” where the longest upper tail, sometimes reaching very high rates (i.e., 15% MAR: $X_1 = 98\%$) (see Table 1). However, we observe no such real differences in the Monte Carlo Errors. This can be due to limited number of simulation runs used. We observe that the estimated mean ESEs and RMSEs for all HMI methods are smaller as compared to “ $MICE_{DEF}$ ” and “ $MICE_{PMM}$ ”

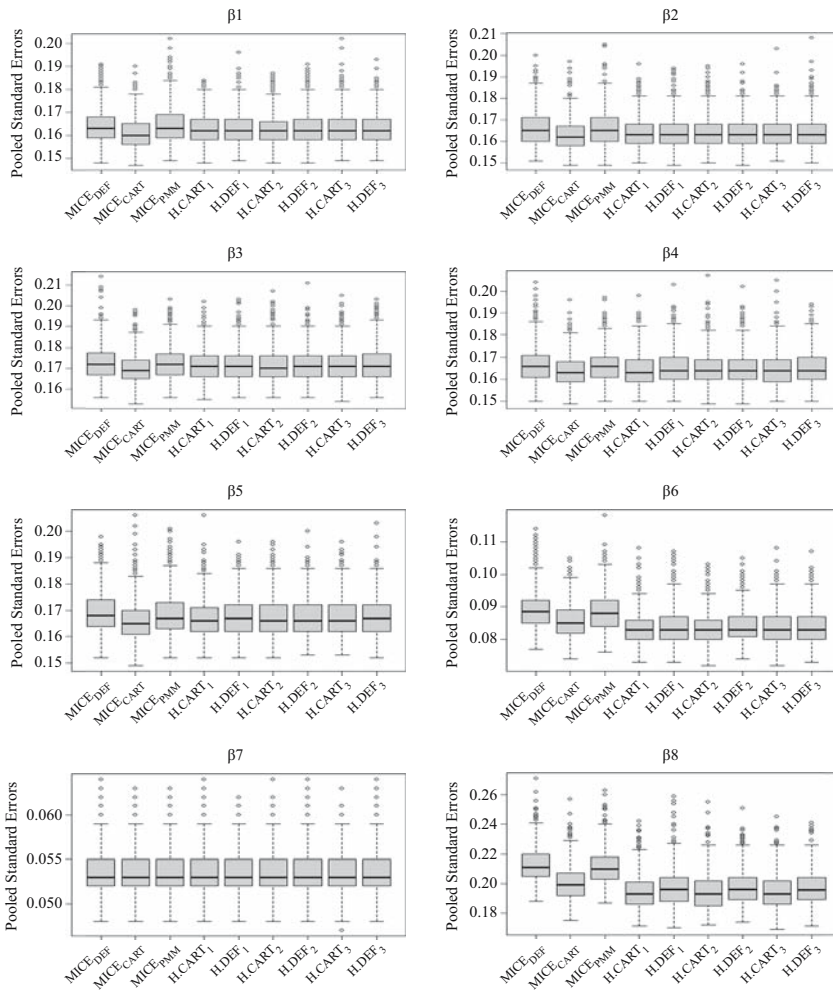


Fig. 4. Simulated data: Boxplots of the pooled standard errors for eight regression coefficients (i.e., β_1, \dots, β_8) under 15% MAR over 1000 simulations. Within each plot the methods are: $MICE_{DEF}$ (default), $MICE_{CART}$ (classification and regression trees), $MICE_{PMM}$ (predictive mean matching), $CART_1$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 80$) and CART), $H.CART_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 150$) and CART), $H.CART_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 400$) and CART), $H.DEF_1$ (combination of and prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 80$) and default), $H.DEF_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 150$) and default), $H.DEF_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25, b_\alpha = 0.25, k = 400$) and default).

under 10% and 15% MAR for most of the covariates. The estimated mean ESEs and RMSEs for “ $MICE_{CART}$ ” are similar to all HMI methods under 10% and 15% MAR for most of the covariates. The amount of bias is also relatively less for the proposed HMI methods (see Table 2). The average point estimates based on proposed HMI methods are close to the corresponding true values in most of the cases (see Figures 1–2). Average standard errors based on proposed HMI methods are also smaller for all cases as compared to three MICE based MI methods (see Figures 3–4).

Table 1. Estimated coverage probabilities for several MI methods under different missing percentages and MAR assumption. All results are based on 10 imputations. Estimates are shown for eight regression coefficients (i.e. β_1, \dots, β_8).

		10% MAR							
Coef.	MICE _{PMM}	MICE _{CART}	MICE _{DEF}	H.DEF ₁	H.CART ₁	H.DEF ₂	H.CART ₂	H.DEF ₃	H.CART ₃
β_1	95	97	95	96	95	96	95	96	96
β_2	95	96	95	96	96	96	96	96	96
β_3	96	97	96	96	97	96	96	96	96
β_4	95	96	96	94	94	95	94	94	95
β_5	95	96	95	95	96	95	97	95	97
β_6	94	96	96	95	96	95	95	94	96
β_7	95	95	95	96	97	95	96	95	96
β_8	96	96	95	96	97	97	96	97	97

		15% MAR							
β_1	97	98	94	97	96	98	96	98	96
β_2	95	96	95	97	97	96	96	96	96
β_3	95	97	95	97	97	97	96	96	97
β_4	95	95	96	96	95	96	96	96	96
β_5	95	94	96	96	96	95	96	95	96
β_6	96	95	96	96	96	96	96	96	96
β_7	95	95	96	95	96	95	96	96	96
β_8	97	97	96	98	96	97	97	97	97

CART₁ is a combination of MICE_{CART}, prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 0.25, k = 80$), H.CART₂ is a combination of MICE_{CART} and prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 0.25, k = 150$), H.CART₃ is a combination of MICE_{CART} and prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 0.25, k = 400$), H.DEF₁ is a combination of MICE_{DEF} and prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 0.25, k = 80$), H.DEF₂ is a combination of MICE_{DEF} and prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 0.25, k = 150$), H.DEF₃ is a combination of MICE_{DEF} and prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 0.25, k = 400$).

Table 2. Estimated coverage probabilities for several MI methods under different missing percentages and MAR assumption. All results are based on 10 imputations. Estimates are shown for eight regression coefficients (i.e. β_1, \dots, β_8). Bold figures indicate the smallest root mean square errors among nine MI methods.

Coef.	10% MAR								
	MICE- _{PMM}	MICE- _{CART}	MICE- _{DEF}	H.DEF ₁	H.CART ₁	H.DEF ₂	H.CART ₂	H.DEF ₃	H.CART ₃
β_1	0.16 (0.16) 0	0.14 (0.14) 0	0.16 (0.16) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0
β_2	0.16 (0.16) 0	0.15 (0.15) 0	0.16 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0
β_3	0.16 (0.16) 0	0.15 (0.15) 0	0.16 (0.16) 0	0.16 (0.16) 0	0.15 (0.15) 0	0.16 (0.16) 0	0.15 (0.15) 0	0.16 (0.16) 0	0.15 (0.15) 0
β_4	0.16 (0.16) 0	0.15 (0.15) 0	0.16 (0.16) 0	0.16 (0.16) 0	0.16 (0.16) 0	0.16 (0.16) 0	0.16 (0.16) 0	0.16 (0.16) 0	0.16 (0.16) 0
β_5	0.16 (0.16) 0	0.16 (0.16) 0	0.16 (0.16) 0	0.16 (0.15) 0	0.15 (0.15) 0	0.16 (0.15) .01	0.15 (0.15) 0	0.16 (0.15) .01	0.15 (0.15) 0
β_6	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0
β_7	0.05 (0.05) 0	0.04 (0.04) 0	0.04 (0.04) 0	0.04 (0.04) 0	0.04 (0.04) 0	0.05 (0.05) 0	0.04 (0.04) 0	0.05 (0.04) .01	0.04 (0.04) 0
β_8	0.19 (0.19) 0	0.17 (0.17) 0	0.19 (0.19) 0	0.17 (0.17) 0	0.17 (0.16) .01	0.17 (0.17) 0	0.17 (0.16) .01	0.17 (0.17) 0	0.17 (0.16) .01

Coef.	15% MAR								
	MICE- _{PMM}	MICE- _{CART}	MICE- _{DEF}	H.DEF ₁	H.CART ₁	H.DEF ₂	H.CART ₂	H.DEF ₃	H.CART ₃
β_1	0.15 (0.15) 0	0.14 (0.14) 0	0.17 (0.17) 0	0.14 (0.14) 0	0.15 (0.15) 0	0.14 (0.14) 0	0.15 (0.15) 0	0.14 (0.14) 0	0.15 (0.15) 0
β_2	0.16 (0.16) 0	0.15 (0.15) 0	0.17 (0.17) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0	0.15 (0.15) 0
β_3	0.17 (0.17) 0	0.15 (0.15) 0	0.17 (0.17) 0	0.15 (0.15) 0	0.16 (0.16) 0	0.15 (0.15) 0	0.16 (0.16) 0	0.16 (0.15) 0	0.16 (0.16) 0
β_4	0.16 (0.16) 0	0.16 (0.16) 0	0.16 (0.16) 0	0.16 (0.15) .01	0.16 (0.16) 0	0.16 (0.15) .01	0.16 (0.16) 0	0.16 (0.15) .01	0.16 (0.16) 0
β_5	0.17 (0.17) 0	0.17 (0.16) 0	0.17 (0.17) 0	0.16 (0.16) 0	0.16 (0.16) 0	0.17 (0.16) .01	0.16 (0.16) 0	0.17 (0.16) .01	0.16 (0.16) 0
β_6	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0	0.08 (0.08) 0
β_7	0.05 (0.05) 0	0.05 (0.05) 0	0.05 (0.05) 0	0.05 (0.05) 0	0.05 (0.05) 0	0.05 (0.05) 0	0.05 (0.05) 0	0.05 (0.05) 0	0.05 (0.05) 0
β_8	0.20 (0.20) 0	0.18 (0.17) 0	0.21 (0.21) 0	0.17 (0.17) 0	0.18 (0.17) .01	0.18 (0.17) .01	0.18 (0.17) .01	0.18 (0.17) .01	0.18 (0.17) .01

CART₁ is a combination of MICE-_{CART}, prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 80$), H.CART₂ is a combination of MICE-_{CART} and prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 150$), H.CART₃ is a combination of MICE-_{CART} and prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 400$), H.DEF₁ is a combination of MICE-_{DEF} and prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 80$), H.DEF₂ is a combination of MICE-_{DEF} and prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 150$), H.DEF₃ is a combination of MICE-_{DEF} and prior specifications values (i.e. $a_\alpha = 0.25, b_\alpha = 400$).

6. Survey Data Example

The Bureau of Statistics Punjab has conducted the Multiple Indicator Cluster Survey (MICS) Punjab, 2014 Pakistan in collaboration with the United Nations Children's Fund (UNICEF). The MICS 2014 women's data comprises more than 200 background characteristics (e.g., demographics, age, education, motherhood and recent births) on 61,286 observations from 36 districts of Punjab. Face to face interviews with household members are conducted to collect data. The documents related to MICS Punjab consisting of the final report, key findings, survey plan, list of indicators and questionnaires can be found on the MICS website ([www.http://bos.gov.pk](http://bos.gov.pk)). The health benefits of breastfeeding are no longer in doubt (WHO 2003). Breastfeeding does not only contribute to the early development of a child but is also crucial for the wellbeing of the mother as well. MICS 2014 women's data can be used to determine the effect of various factors affecting the feeding practices in Punjab. This analysis could be very helpful in decision making policies regarding women and child health.

6.1. Imputation of MICS Background Variables

Since MICS data for women contains data with a possibly complex dependency structure, the application of the package "mice" can become problematic due to various limitations, for example non-convergence of the Gibbs sampler in special cases, large amount of missing values, tedious work required for specification of imputation models and interaction terms in presence of large data bases with hundreds of variables and multicollinearity problems (Van Buuren and Oudshoorn 1999). It was not possible to have a proper comparison of the proposed and existing MI approaches in such cases. Therefore, it was decided to select a subset containing *seven* continuous and *37* categorical variables. The selection of variables is made according to the evidence from demographical and behavioral risk factors effecting inclination towards breastfeeding. Some of the selected categorical variables, that is, district, has lots of categories ($k = 36$), hence keeping the analysis comparable and challenging at the same time. Among these *44* variables, *five* variables have less than *14%* missing values; *15* variables have between *32%* to *68%* missing values; *22* variables have between *80%* to *95%* missing values. Only *two* variables are completely observed. The summary of all categorical and continuous variables in survey data can be seen in the Tables S1–S2 of supplementary file respectively. All variables are included in the imputation model. The reasons of missing observations in MICS data are typical, that is, nonresponse, don't know, not reached etc. For the sake of multiple imputations, all reasons for item nonresponse are treated as MAR.

20 sampling simulations are run and $M = 10$ completed data sets are generated for each MI method. The binary response (Ever Breastfeed) which comprises two categories (Yes / No) is finally modeled using a GLM analysis model depending on four categorical variables (Mother Ever Attended School: two categories, Delivery by C Section: two categories, Satisfaction from Health: two categories, Area: two categories) and two continuous covariates (Age of Mother and Freq. of Mother Reads New). The R package "VIM" (Templ et al. 2012) is utilized to explore the pattern of missing values. Figure 5 displays the proportion of missing values and the missing data pattern for the variables used in the analysis model. The graphics of incomplete categorical variables used in

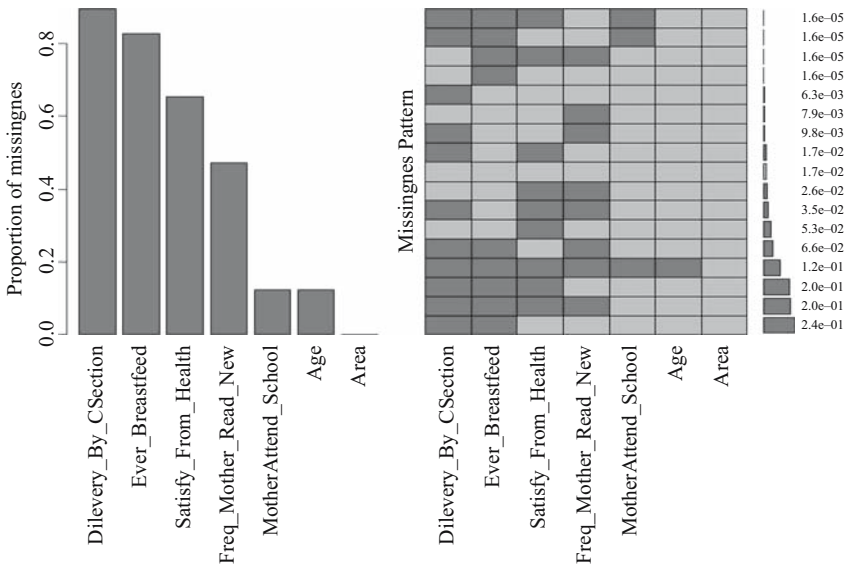


Fig. 5. Survey data: Aggregation graphic for the variables in the analysis model.

survey data are presented in first four aggregate plots of supplementary file (see Figure S1 in online supplemental file). Aggregate plots for the continuous variables used in the survey data can be seen in Figure S2 of supplementary file. Since there are no true values to compare for in the survey data example, we calculated the complete case (CC) estimates for comparison purposes (Table 3). The time taken by each MI method is shown in Table 4. Boxplots of the pooled point estimates and standard errors for survey data are shown in Figures 6 and 7 respectively.

6.2. Results and Discussions

Figure 5 in the survey data example displays the bar plot on the left side which shows the proportions of missing values in the predictors. The categorical predictor “Delivery By C Section” has the highest amount of missing values (i.e., more than 80%) followed by “Ever Breastfeed” (about 80%), “Satisfaction From Health” (about 60%) and “Freq. of Mother

Table 3. Survey data: complete case (CC) estimates

Variables	est	se
Age	0.14	0.06
Mother attended School	-0.59	0.77
Freq. mother reads news	-0.09	0.15
Dilevery by CSecion	0.43	0.25
Satisfaction from health	0.27	0.27
Area	0.16	0.25

The CC analysis uses only the complete cases (n = 4264), “est” and “se” denote the point estimates and standard errors of the coefficients of the GLM, respectively

Table 4. Survey data: Time taken by various MI methods

Method	Time
MICE-CART	4.20d
MICE-PMM	3.52d
MICE-DEF	3.14d
H.DEF ₁	1.70d
H.CART ₁	1.62d
H.DEF ₂	1.68d
H.CART ₂	1.64d
H.DEF ₃	1.82d
H.CART ₃	1.77d

Note: Time = the time to complete 10 multiple imputation by variants of MI across 20 simulations and d = days

Reads New” (about 40%). The amount of missing values is rather small for “Mother Ever Attended School” and “Age” (i.e., less than 20%). The categorical predictor “Area” has no missing values. An aggregation plot on the right side shows all existing combinations of missing (red) and observed (blue) values. The frequencies of different combinations can be seen by a small bar plot on the right side (Templ et al. 2012). The aggregation plot reveals that if missing values occur in the variable “Ever Breastfeed”, they most often also occur in the variables “Satisfaction From Health”, “Freq. of Mother Reads New” and “Delivery By C Section”. We note, that the standard errors for most of the coefficients are relatively smaller to the (absolute) point estimates under all MI methods (see Figures 6–7). We noticed that point estimates in MICE-CART are nearer to the estimates in complete case analysis for most of the cases as compared to the hybrid methods (see Table 3). In the survey data example, the HMI methods tend to show smaller pooled standard errors for most of the co-variates as compared to the MICE methods. We see, that when HMI MI methods are applied to the survey data set, the pooled standard errors are comparatively smaller for all covariates as compared to the “MICE-DEF” MI method and smaller for most the covariates (i.e., “Age”, “Freq. of Mother Reads New”, “Delivery By C Section” and “Area”) as compared to the “MICE-PMM” MI method. The “H.CART” method tends to show smaller pooled standard errors for the covariates (i.e., “Age” and “Delivery by C Section”) as compared to its counterparts. For the rest of the co-variates, the differences are also not so high, which suggests a reasonable performance compared to MICE, see Figures 6–7. The computational burden is significantly reduced for most of the settings using the proposed HMI methods, see Table 4.

7. Concluding Remarks

Investigation of optimal strategies for fitting MI on the classical regression techniques in the presence of a large number of variables is questionable. There is no general agreement on especially to the how many variables should the imputation model have. According to Van Buuren (2012) the number of predictors should be as large as possible for the generally accepted principle for imputation. On the other hand Hardt et al. (2012)

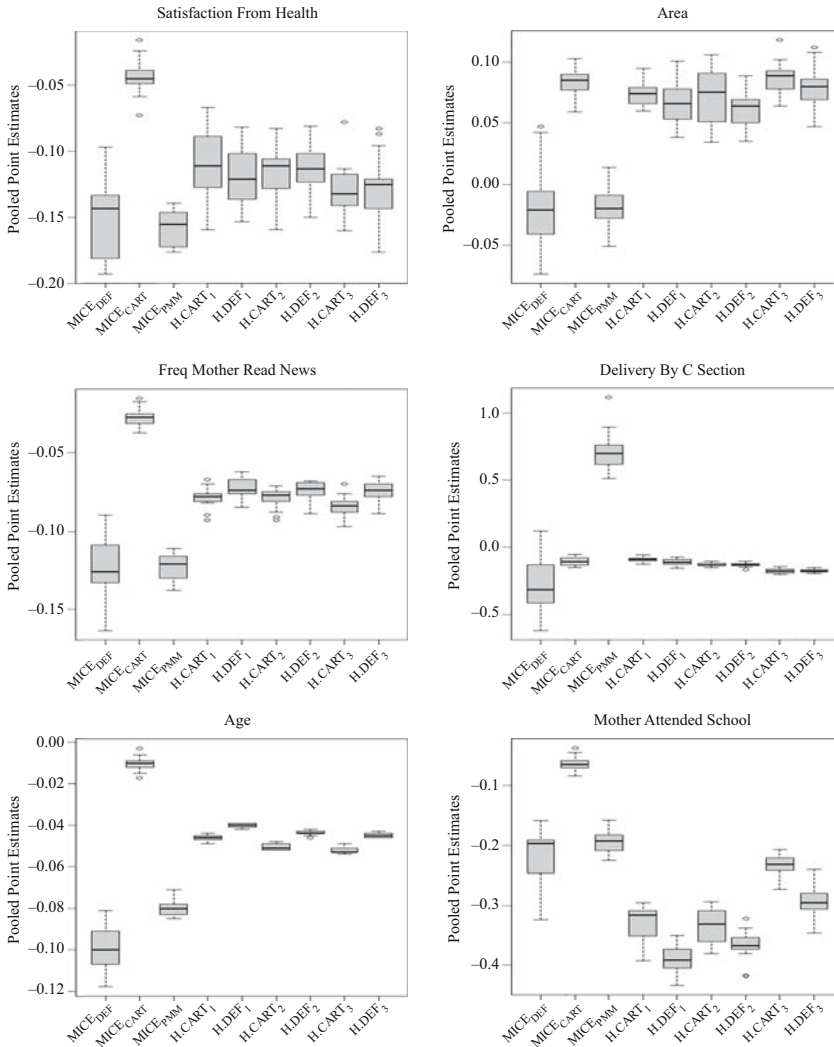


Fig. 6. Survey data: Boxplots of pooled point estimates for variables in the analysis model. Within each plot the methods are: $MICE_{DEF}$ (default), $MICE_{CART}$ (classification and regression trees), $MICE_{PMM}$ (predictive mean matching), $CART_1$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 80$) and CART), $H.CART_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 150$) and CART), $H.CART_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 400$) and CART), $H.DEF_1$ (combination of and prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 80$) and default), $H.DEF_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 150$) and default), $H.DEF_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, $k = 400$) and default).

recommends that the small number of variables will be sufficient to successfully implement MI in the R package “mice”. It is worth noting that the performance of the regression techniques is known to deteriorate as number of variables increases and it is generally not feasible to include all variables in imputation models. Little (2018) focuses on the flexibility of MICE by referring a large list of references to the application of chained equation MI in real applications. As opposed to Little (2018), we claim that high-dimensional real applications in these references are limited. Many of the references

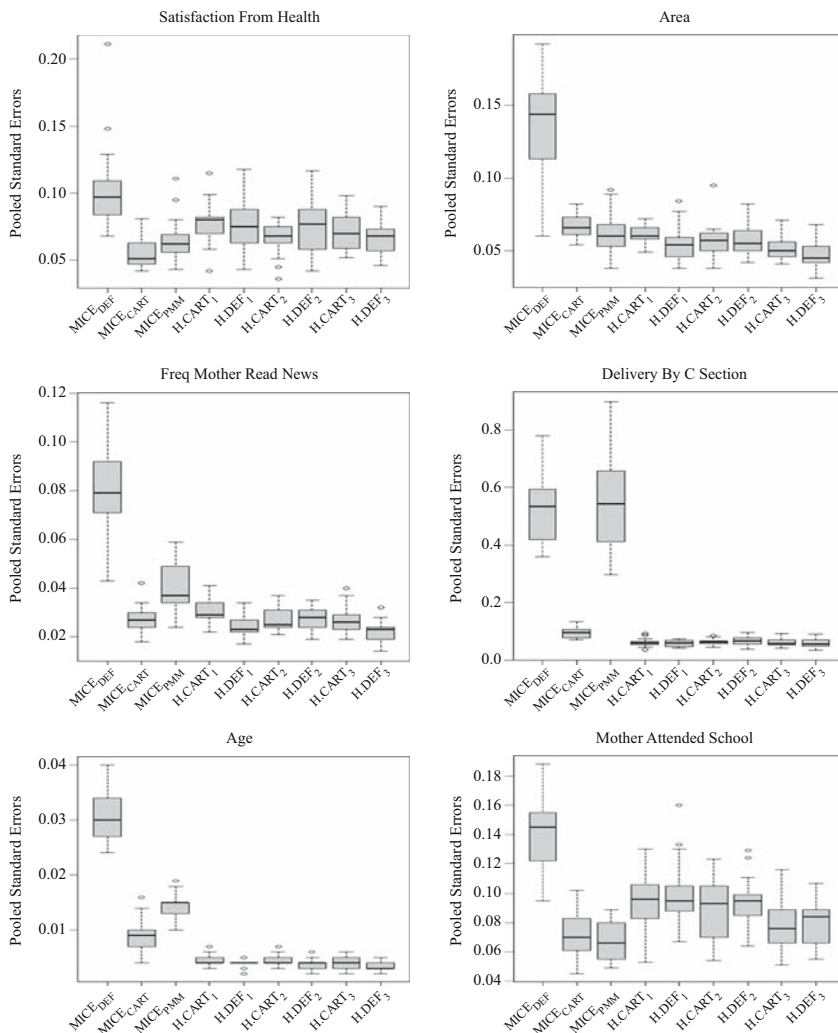


Fig. 7. Survey data: Boxplots of the pooled standard errors for variables in the analysis model. Within each plot the methods are: $MICE_{DEF}$ (default), $MICE_{CART}$ (classification and regression trees), $MICE_{PMM}$ (predictive mean matching), $CART_1$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, k 80) and CART), $H.CART_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, k 150) and CART), $H.CART_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, k 400) and CART), $H.DEF_1$ (combination of and prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, k 80) and default), $H.DEF_2$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, k 150) and default), $H.DEF_3$ (combination of prior specifications values (i.e., $a_\alpha = 0.25$, $b_\alpha = 0.25$, k 400) and default).

applied MICE to epidemiological real data in context of large sample sizes rather than a large number of mixed type variables. Hybrid MI techniques open the door to conduct imputation in the high-dimensional setting by combining various properties of existing MI approaches, the main advantages of the proposed methodology are as follows: (1) it is flexible and can be implemented to mixed type high-dimensional data, (2) it does not rely on heuristic rules of thumb for predictor selection and (3) it is fast.

Various issues concerning the implementation of the hybrid imputation models need further research. For example, CART method resulted similar or improved performance over hybrid models in most our all applications in simulation studies where we have considered moderate rates of missingness. Whereas, for real world applications where we have high missing rates, hybrid models performed relatively better than CART which gives an indication that we may need even higher rates of missingness than we used in our simulations to get improved performance over CART. Moreover for better performance it may be that we need an even larger number of imputations than we used in our application. Also HMI methods eliminate the use of predictor selection beforehand. Further comparisons can be made for data with ordinal nature and more categories with large values of prior specifications. Our proposed method is also computationally inexpensive and requires less time even when performed with a large number of iterations. However, of note, one limitation of the proposed method is that, the information available in the continuous variables is not used for imputing the categorical variables. Therefore, it is too early to make any final conclusion until unless experiments with diversity of settings are conducted.

8. References

- Arnold, B.C., and S.J. Press. 1989. "Compatible Conditional Distributions". *Journal of the American Statistical Association* 84:152–156. DOI: <https://doi.org/10.2307/2289858>.
- Allison P.D. 2002. *Missing Data*. Thousand Oaks, CA: Sage Publications. DOI: <https://dx.doi.org/10.4135/9781412985079>.
- Abdella, M., and T. Marwala, 2005. "The use of genetic algorithms and neural networks to approximate missing data in database". In Proceedings of the IEEE 3rd International Conference on Computational Cybernetics, 2005. 24: 207–212. DOI: <https://doi.org/10.1109/ICCCYB.2005.1511574>.
- Ankaiah, N., and V.Ravi. 2011. "A novel soft computing hybrid for data imputation". In Proceedings of the 7th International Conference on Data Mining (DMIN). Las Vegas, USA. Available at: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.217.7984&rep=rep1&type=pdf>.
- Akande, O., F. Li, and J. Reiter. 2017. "An empirical comparison of multiple imputation methods for categorical data". *The American Statistician* 71: 162–170. DOI: <https://doi.org/10.1080/00031305.2016.1277158>.
- Andridge, R.R., and R.J.A. Little. 2017. "A Review of Hot Deck Imputation for Survey Non-response". *International statistical review* 78(1): 40–64. DOI: <https://doi.org/10.1111/j.1751-5823.2010.00103.x>.
- Armina, R., A.M. Zain, N.A. Ali, and R. Sallehuddin, 2017. "A review on missing value estimation using imputation algorithm". *Journal of Physics: Conference Series* 892(1). DOI: <https://doi.org/10.1088/1742-6596/892/1/012004>.
- Bengio, Y., and F. Gingras. 1995. "Recurrent neural networks for missing or asynchronous data. In Touretzky, D.S., Mozer, M.C. and Hasselmo, M.E. editors". *Advances in Neural Information Processing Systems* 8: 95–401. MIT Press, Cambridge, MA. Available at: <https://proceedings.neurips.cc/paper/1995/file/ffeed84c7cb1ae7bf4ec4bd78275bb98-Paper.pdf>.

- Barnard, J., and X. Meng. 1999. "Applications of multiple imputation in medical studies: From AIDS to NHANES". *Statistical Methods in Medical Research* 8:17–36. DOI: <https://doi.org/10.1177/096228029900800103>.
- Breiman, L. 2001. "Random Forests". *Machine Learning* 45(1): 5–32. DOI: <https://doi.org/10.1023/A:1010933404324>.
- Batista, G., and M.C. Monard. 2003. *Experimental comparison of K-nearest neighbour and mean or mode imputation methods with the internal strategies used by C4.5 and CN2 to treat missing data*. University of Sao Paulo. Available at: <https://www.semanticscholar.org/paper/Experimental-comparison-pf-K-NEAREST-NEIGHBOUR-and-Batista-Monard/35346d559d1bcfdf27acff66267e8f1d67190f23>.
- Burton, A., D. G. Altman, P. Royston, and R.L. Holder. 2006. "The design of simulation studies in medical statistics". *Statistics in Medicine* 25: 4279–4292. DOI: <https://doi.org/10.1002/sim.2673>.
- Chung, D., and F.L. Merat. 1996. Neural network based sensor array signal processing. In: Proc Int Conf Multisens Fusion Integr Intell Syst. Washington, USA: 757–764. DOI: <https://doi.org/10.1109/MFI.1996.572313>.
- Chandra, A., G.M. Martinez, W.D. Mosher, J.C. Abma, and J. Jones. 2005. "Fertility, family planning, and reproductive health of U.S. women: data from the 2002 National Survey of Family Growth". *Vital Health Stat* 23: 1–160. Available at: <https://pubmed.ncbi.nlm.nih.gov/16532609/>
- Corsi, D.J., J.M. Perkins, and S.V. Subramanian. 2017. "Child anthropometry data quality from Demographic and Health Surveys, Multiple Indicator Cluster Surveys, and National Nutrition Surveys in the West Central Africa region: are we comparing apples and oranges?". *Global Health Action*. DOI: <https://doi.org/10.1080/16549716.2017.1328185>.
- Dunson, D.B., and C. Xing. 2009. "Nonparametric Bayes modeling of multivariate categorical data". *Journal of the American Statistical Association* 104: 1042–1051. DOI: <https://doi.org/10.1198/jasa.2009.tm08439>.
- Gelman, A., and T.P. Speed. 1993. "Characterizing a joint probability distribution by conditionals". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 55: 85–188. DOI: <https://doi.org/10.1111/j.2517-6161.1993.tb01477.x>.
- Graham, J.W., and J.L. Schafer. 1999. "On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.)". *Statistical strategies for small sample research*: 1–29.
- Gulliford, M.C., O.C. Ukoumunne, and, S. Chinn. 1999. "Components of Variance and Intra class Correlations for the Design of Community-based Surveys and Intervention Studies: Data from the Health Survey for England". *American Journal of Epidemiology* 149(9): 876–883. DOI: <https://doi.org/10.1.1.565.7897>.
- Harel, O., and X.H. Zhou. 2007. "Multiple imputation: Review of theory, implementation and Software". *Statistics in Medicine* 26: 3057–3077. DOI: <https://doi.org/10.1002/sim.2787>.
- Horton, N.J., and K.P. Kleinman. 2007. "Much ado about nothing: a comparison of missing data methods and software to fit incomplete regression models". *The American Statistician* 61: 79–90. DOI: <https://doi.org/10.1198/000313007X172556>.

- Honaker, J., G. King, and M. Blackwell. 2011. "Amelia II: A program for missing data". *Journal of Statistical Software* 45(7): 1–47. DOI: <https://doi.org/10.18637/jss.v045.i07>.
- Hardt, J., M. Herke, and R. Leonhart. 2012. "Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research". *BMC Medical Research Methodology* 12(1). DOI: <https://doi.org/10.1186/1471-2288-12-184>.
- Kohonen, T. 1995. *Self-Organizing Maps*. Springer. Heidelberg. Available at: <https://www.springer.com/gp/book/9783642976100>.
- Lazarsfeld, P.F. 1950. *The logical and mathematical foundation of latent structure analysis*. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Studies in social psychology in World War II: Vol. 4. Measurement and prediction*. Chap. 10: 362–412. Princeton, NJ: Princeton University Press. Available at: <https://psycnet.apa.org/record/1951-03037-000>.
- Li, F., Y. Yu, and D.B. Rubin. 2012. *Imputing missing data by fully conditional models: some cautionary examples and guidelines*. Duke University Department of Statistical Science Discussion Paper: 11–24. Available at: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.228.7010>.
- Little, R.J.A. 1988. "A Test of Missing Completely at Random for Multivariate Data with Missing Values". *Journal of the American Statistical Association* 83(404): 1198–1202. DOI: <https://doi.org/10.1080/01621459.1988.10478722>.
- Little, R.J. 2018. "On Algorithmic and Modeling Approaches to Imputation in Large Data Sets". *Statistica Sinica*. <http://www3.stat.sinica.edu.tw/statistica/J30N4/J30N401/J30N401.html>
- Little, R.J.A., and D.B. Rubin. 2002. *Statistical analysis with missing data* (2nd edition.). New York: Wiley. Available at: <https://www.wiley.com/en-us/Statistical+Analysis+with+Missing+Data%2C+2nd+Edition-p-9781119013563>.
- McLachlan, G.J., and D. Peel. 2000. *Finite mixture models*. New York: Wiley. DOI: <http://dx.doi.org/10.1002/0471721182>.
- Marseguerra, M., and A. Zoia. 2005. "The autoassociative neural network in signal analysis. II. Application to on-line monitoring of a simulated BWR component". *Annals of Nuclear Energy* 32(11): 1207–1223. DOI: <https://doi.org/10.1016/j.anucene.2005.03.005>.
- Marwala, T., and S. Chakraverty. 2006. "Fault classification in structures with incomplete measured data using auto associative neural networks and genetic algorithm". *Current Science India* 90(4): 542-548. JSTOR. Available at: www.jstor.org/stable/24088946.
- Morris, T.P., R.W. Ian, and R. Patrick. 2014. "Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws". *BMC Medical Research Methodology* 14 (1): 75. DOI: <https://doi.org/10.1186/1471-2288-14-75>.
- Murray, J.S., and J.P. Reiter. 2016. "Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence". *Journal of the American Statistical Association* 111: 1466–1479. DOI: <https://doi.org/10.1080/01621459.2016.1174132>.
- Narayanan, S., J.L.Vian, J. Choi, M. El-Sharkawi, and B.B.Thompson. 2002. *Set constraint discovery: missing sensor data restoration using auto-associative regression*

- machines*. In Proceedings of the international Joint Conference on Neural Networks (IJCNN): 2872–2877. DOI: <https://doi.org/10.1109/IJCNN.2002.1007604>.
- Oja, E., and S. Kaski. 1999. *Kohonen Maps*. Elsevier, Amsterdam. Available at: <https://www.elsevier.com/books/kohonen-maps/oja/978-0-444-50270-4>.
- Oba, S., M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii. 2003. “A Bayesian missing value estimation method for gene expression profile data”. *Bioinformatics* 19: 2088–2096. DOI: <https://doi.org/10.1093/bioinformatics/btg287>.
- Pyle, D. 1999. *Data preparation for data mining*. Morgan Kaufmann Publishers Inc. San Francisco. Available at: <https://dl.acm.org/doi/book/10.5555/299577>.
- Pérez, A., R.J. Dennis, J.F. Gil, M.A. Rondón, and A. López. 2002. “Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia”. *Statistics in Medicine* 21: 3885–3896. DOI: <https://doi.org/10.1002/sim.1391>.
- Quanli, W., M.V. Danial, J.P. Reiter, and H. Jigchen. 2018. *NPBayesImputeCat: Non-Parametric Bayesian Multiple Imputation for Categorical Data*. R package version 0.1, Available at: <https://CRAN.R-project.org/package=NPBayesImputeCat>.
- Rubin, D.B. 1976. “Inference and Missing Data”. *Biometrika* 63: 581–590. DOI: <https://doi.org/10.2307/2335739>.
- Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. Available at: <https://www.wiley.com/en-us/Multiple+Imputation+for+Nonresponse+in+Surveys-p-9780471655749>.
- Roth, P.L. 1994. “Missing data: A conceptual review for applied psychologists”. *Personnel Psychology* 47: 537–560. DOI: <https://doi.org/10.1111/j.1744-6570.1994.tb01736.x>.
- Rubin, D.B. 1996. “Multiple imputation after 18 + years”. *Journal of the American Statistical Association* 91: 473–489. DOI: <https://doi.org/10.1080/01621459.1996.10476908>.
- Raghunathan, T.W., J.M. Lepkowski, J. van Hoewyk, and P.A. Solenbeger. 2001. “Multivariate technique for multiply imputing missing values using a sequence of regression models”. *Survey Methodology* 27: 85–95. Available at: <https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.405.4540>.
- Reiter, J.P., T.E. Raghunathan, and S. Kinney. 2006. “The importance of modeling the survey design in multiple imputation for missing data”. *Survey Methodology* 32: 143–149. Available at: <http://www2.stat.duke.edu/~jerry/Papers/SM06.pdf>.
- Royston, P., and I.R. White. 2011. “Multiple imputation by chained equations (mice): Implementation in Stata”. *Journal of Statistical Software* 45(4): 1–20. DOI: <https://doi.org/10.18637/jss.v045.i04>.
- R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org>.
- Sharpe, P.K., and R.J. Solly. 1995. “Dealing with missing values in neural network-based diagnostic systems”. *Neural Computing and Applications* 3(2): 73–77. DOI: <https://doi.org/10.1007/BF01421959>.
- Schafer, J.L. 1997. *Analysis of incomplete multivariate data*. London: Chapman and Hall. DOI: <https://doi.org/10.1201/9780367803025>

- Schafer, J.L. and J.W. Graham. 2002. "Missing data: Our view of the state of the art". *Psychological methods* 7: 147–177. DOI: <https://doi.org/10.1037/1082-989X.7.2.147>.
- Schlomer, G.L., S. Bauman, and N.A. Card. 2010. "Best Practices for Missing Data Management in Counseling Psychology". *Journal of Counseling Psychology* 57(1): 1–10. DOI: <https://doi.org/10.1037/a0018082>.
- Si, Y., and J.P. Reiter. 2013. "Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys". *Journal of Educational and Behavioral Statistics* 38: 499–521. DOI: <https://doi.org/10.3102/1076998613480394>.
- Templ, M., A. Andreas, K. Alexander, and P. Bernd. 2012. *VIM: Visualization and Imputation of Missing Values*. Available at: <http://cran.r-project.org/web/packages/VIM/VIM.pdf>.
- Van Buuren, S. 2007. "Multiple imputation of discrete and continuous data by fully conditional specification". *Statistical Methods in Medical Research* 16: 219–242. DOI: <https://doi.org/10.1177/0962280206074463>.
- Van Buuren, S. 2012. *Flexible Imputation of Missing Data*, London: Chapman and Hall/CRC. DOI: <https://doi.org/10.1201/b11826>.
- Van Buuren, S., and K. Groothuis-Oudshoorn. 1999. *Flexible multivariate imputation by MICE*. TNO Prevention and Health. Leiden. Available at: <https://stefvanbuuren.name/publications/Flexible%20multivariate%20-%20TNO99054%201999.pdf>.
- Van Buuren, S., and K. Groothuis-Oudshoorn. 2011. "mice: Multivariate imputation by chained equations". *R. Journal of Statistical Software* 45(3): 1–67. DOI: <https://doi.org/10.18637/jss.v045.i03>.
- Van Ginkel, J.R. 2007. *Multiple imputation for incomplete test, questionnaire and survey data*. Ph.D. dissertation. Tilburg University. Department of Methodology and Statistics. Available at: <https://pure.uvt.nl/ws/portalfiles/portal/839209/224433.pdf>.
- Vermunt, J.K., J.R. van Ginkel, L.A. van der Ark, and K. Sijtsma. 2008. "Multiple imputation of incomplete categorical data using latent class analysis". *Sociological Methodology* 38: 369–397. DOI: <https://doi.org/10.1111/j.1467-9531.2008.00202.x>.
- WHO (World Health Organization). 2003. *Community-based Strategies for Breastfeeding Promotion and Support in Developing Countries, 2003*. Dept. of child and adolescent health and development. Geneva. Available at: https://www.who.int/maternal_child_adolescent/documents/9241591218/en/.
- Wilkinson, L., and Task Force on Statistical Inference. 1999. "Statistical methods in psychology journals: Guidelines and explanations". *American Psychologist* 54: 594–604. DOI: <https://doi.org/10.1037/0003-066X.54.8.594>.
- Zhu, J., and T.E. Raghunathan. 2016. "Convergence Properties of a Sequential Regression Multiple Imputation Algorithm". *Journal of the American Statistical Association* 110(511): 1112–1124. DOI: <https://doi.org/10.1080/01621459.2014.948117>.

Received March 2019

Revised July 2020

Accepted December 2020