Journal of Official Statistics
vol. 31, i. 2 (2015)

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 149–153, http://dx.doi.org/10.1515/JOS-2015-0011

Preface

1. Emerging Trends and Challenges for Official Statistics

As always, official statistics faces a wide range of challenges, and many of those are associated with trends that have been long in the making. On the one hand, challenges are linked to developments that are mostly under the control of statistical institutes; with ever more complex sample designs, paired with an ambition to compute and present statistics and indicators of increasing complexity, even point estimation is not without its challenges. When precision estimates are to be calculated under such circumstances, practitioners often resort to replication methods.

On the other hand, there are trends that are only partially influenced by statistical institutes, such as the exploration of new data sources. While it could be debated whether administrative data is indeed a new source, or whether it rather represents a renaissance, returning to the presampling era of producing official statistics, increased use of such data is a current trend, driven by various factors, such as cost and response burden reduction, and including (sometimes) legal obligations of a statistical office to use all available data before resorting to (sample) surveys. A possible exception to these steadily advancing trends is Big Data, which constitute a rapidly emerging type of source, of great apparent potential, but as of yet with few actual official statistics applications. The well-known problem of the nonrepresentativity of Big Data, and the difficulty of linking sample units with the observational units of Big Data, still constitute major obstacles to their usefulness for official statistics. However, a promising application, where the impact of this could be seen as less severe, is that of Small Area Estimation (SAE).

Regardless of the source used (sample survey, census, administrative or Big Data), it is well known to official statisticians that merely "providing the figures" without any context may end up doing a disservice to policymakers and, in the worst case, result in inappropriately founded policy decisions. There is a clear trend towards a more reflective approach, with an emphasis not only on producing high-quality statistics, but also on rendering explicit details on exactly how this is being achieved. More precisely, there is a need to understand how high the quality is and where there is room for improvement.

2. New Techniques and Technologies Tackling the New Challenges

The present activities of the official statistics research community reflect (and sometimes shape) current trends and attempts to tackle challenges such as those outlined in Section 1. One of the fora exploring the state of the art regarding official statistics research is New Techniques and Technologies for Statistics (NTTS), which is an international scientific

conference series. The NTTS conferences, which are organised by Eurostat, the statistical office of the European Union, cover new techniques and methods for official statistics and the impact of new technologies on statistical collection, production, and dissemination systems.

The purpose of the NTTS conferences is both to allow the presentation of results from currently ongoing research and innovation projects in official statistics and to stimulate and facilitate the preparation of new innovative projects related to research in statistics within the European Framework Programmes for Research and Development.

From 1992-2001, the first four NTTS conferences were organised triennially. After an eight-year gap, the series resumed in 2009, and has been organised biennially in Brussels ever since. With NTTS being one of the major European scientific conferences with a specific focus on official statistics, the Scientific Committee of NTTS 2013 and the JOS Editorial Board agreed to explore the potential for a special issue of JOS based solely on papers from NTTS 2013 (www.NTTS2013.eu). Following a call for papers, a large number of contributions were received, screened and peer reviewed, and the present JOS special issue represents the positive outcome of this undertaking.

While the articles cover a variety of domains, they could be said to represent three major areas of new developments, upon which we have based this issue's structure. Section 3 presents two of the articles which tackle challenges linked to the traditional source of sample surveys; they both address estimation under complex sampling designs, but in very different ways.

The largest group of articles of this special issue deals with various aspects of new data sources. In Section 4, we present four articles (two on administrative data and two on Big Data) which fall into this category.

Finally, this special issue also includes three articles that deal with post-survey topics which we denote as metadata issues. Section 5 is devoted to these articles which include aspects of quality, respondent attitudes, and statistical disclosure control.

No JOS article is complete without some concluding remarks and future outlook, which we present in Section 6.

3. Estimation

As illustrated by Oguz Alper and Berger in the first article of our special issue, an analytical (in this case regression-based) approach to variance estimation is still possible, even when one is faced with the triple challenge of a complex survey design, a complex function, and longitudinal effects.

Moving on from this in-depth treatment of a specific application, in the second article Zardetto proposes a general estimation system that easily manages complex sample survey estimators and their precision assessment. This illustrates a number of emerging trends: the move to Free and Open Source Software taking place in many statistical offices and supported by the ESS Vision 2020 (http://ec.europa.eu/eurostat/documents/10186/756730/ESS-Vision-2020.pdf/), as well as a systems approach, with general applications being developed for the benefit of the entire statistical system, according to a shared services philosophy.

4. New Data Sources in Official Statistics

4.1. Administrative Data

Administrative data have many uses, but coincidentally, both of the articles related to administrative data presented here concern imputation. Again, however, just as for the estimation articles of Section 3, this section's articles offer the same type of variety on the application/system scale. On the one hand, Fleishman, Gubman, and Tur-Sinai discuss a very specific application of imputation, where dwelling prices are used as proxies for the socioeconomic level of an area. On the other hand, Schnetzer and coauthors present a set of general measures to evaluate imputation quality, and embed this assessment in a broader quality framework.

4.2. Big Data

Given the major recent developments in the field of Big Data, with for instance the High-Level Group for the Modernisation of Statistical Production and Services (HLG) sponsoring the project *The Role of Big Data in the Modernisation of Statistical Production* (HLG 2015; this project was the subject of a highly attended satellite workshop at NTTS 2015) and the ESS Big Data Action Plan and Roadmap 1.0 (Eurostat 2014), it is interesting to recall that at the time of NTTS 2013, Big Data in official statistics was still in its infancy. Apart from the keynote address by Robert M. Groves, the presentations dealing with Big Data were few and far between. One of the notable exceptions was the paper of Daas, Puts, Beulens, and van den Hurk, which also forms part of this special issue, in which the exploration of both opportunities and challenges associated with the application of Big Data for official statistics is discussed. Refreshingly, the article not only expresses itself in generalities, but also presents concrete applications to Dutch traffic loop data and Dutch social media messages.

Under the Fay-Herriot model (Fay and Herriot 1979), Big Data can be used as covariates at area level, without the need for linking at unit level, a requirement for which Big Data frequently prove insufficient, in spite of their size. As many Big Data sources (GPS, GSM, satellite images, traffic loops) have a geographical dimension, this is a promising application. Moreover, SAE provides an excellent framework for using Big Data (albeit mainly in the role of area-level covariates), as the tools developed for other types of arealevel covariates are applicable to Big Data as well. In the special issue article by Giusti and coauthors, mobility data are combined with survey data to explore the potential of SAE in this regard.

5. Metadata

The article by Schnetzer and coauthors discussed in Section 4.1 does in fact deal with measuring (imputation) quality rather than proposing methods for the production of official statistics; in other words, there is a certain overlap between this section and the previous one.

5.1. Respondent Attitude

In the seventh article of this special issue, Torres van Grinsven and Snijkers use social media data sources to gauge the sentiment of business survey respondents, with a view to improving the statistical bureau's understanding of this group of respondents. This understanding could then lead to better strategies for increasing business survey response rates.

While the authors use a novel data source, they do not do so for the direct purpose of producing official statistics. Interestingly, the actual number of social media messages relevant to the analysis of Torres van Grinsven and Snijkers is rather small (a few hundred). Nonetheless, this could in a sense be said to be a Big Data application, because unless there had been an enormous database to begin with (three million new entries each day) to scan, the few messages related to business survey respondents could not have been extracted. Thus, the high granularity of Big Data (rendering it possible to extract information even on a very small subpopulation) is what has rendered this particular analysis possible.

5.2. Statistical Disclosure Control

With an increased demand from policymakers and researchers for specialised frequency tables, many statistical agencies are considering the development of web-based software platforms where users can generate tables of interest.

However, this requires a particular type of metadata to be generated (and then immediately used) "on the fly": the disclosure risk has to be assessed and, if needed, a statistical disclosure control method has to be applied prior to the near realtime release of the table whilst preserving the utility of generated tables to the users to the greatest possible extent. The perennial struggle to strike a balance between utility to data users and data subjects' privacy is thus further complicated by the time pressure introduced by dynamic "on-the-fly" table generation. To remedy this, Shlomo, Antal, and Elliot propose, in the eighth article of the special issue, a new disclosure risk measure and a data utility measure that can be defined at the level of the generated output table and calculated "on the fly".

5.3. Comprehensive Metadata

In the final article, Signore, Brancato, and Scanu argue for going beyond the production chain, and integrating all relevant metadata into one system. Most notably, this extension includes business-related metadata concerning planning, execution, and assessment. The business metadata are needed for the planning phase concerns, e.g., strategic goals and high-level decisions, programme planning (including multiannual plans), and plans for training as well as methodological and IT support. In the execution phase, service-level agreements with stakeholders (administrative data suppliers, outsourcing contractors) are at the core of the business-related metadata. Finally, the authors mention the various business-related metadata available in the assessment phase (user satisfaction surveys, statistics on data access, spontaneous feedback, staff satisfaction surveys, etc.).

6. Advancing the State of the Art of Official Statistics Research

There are several indispensable ingredients in a special issue such as this one. First and foremost, there must be a large body of high-quality articles. Therefore, we thank all the authors who have kindly submitted their conference papers for our consideration. In addition, we would like to thank the referees for taking the time to review all manuscripts, sometimes several versions, and providing constructive comments, leading to considerable improvements in relation to the original manuscripts. Well over 70 referees have contributed to this special issue.

With conference papers as the point of departure for this endeavour, there is a clear challenge in transforming these papers into articles to be included into a scientific journal such as JOS. The purpose and audience are sometimes very different, and mere project presentations, which might be perfectly legitimate at a conference, need to undergo considerable reworking if they are to make it into a scientific journal. On the side of the NTTS Scientific Committee, a step in this direction has been taken through the more stringent requirements for NTTS 2015 conference abstracts. The articles that were ultimately included in this special issue of JOS follow the remit of the JOS editorial board for research results about "survey methods, quality, applications, policy issues and other aspects of production of official statistics" and each article contains relevant sections including an overview of the state of the art, the contribution of the article, a critical discussion of alternatives, a summary discussion with conclusions and future areas of research.

Through this special issue, as with the NTTS series of conferences, we hope to contribute to, and disseminate, some of the richness of the state of the art in official statistics research.

Martin Karlberg
Guest Editor

Silvia Biffignandi
Piet J.H. Daas
Anders Holmberg
Beat Hulliger
Pascal Jacques
Risto Lehtonen
Ralf T. Münnich
Natalie Shlomo
Roxane Silberman
Ineke Stoop
Guest Associate Editors

7. References

Eurostat. 2014. "ESS Big Data Action Plan and Roadmap 1.0." Approved by the 22nd Meeting of the European Statistical System Committee. Available at: http://www.cros-portal.eu/content/ess-big-data-action-plan-and-roadmap-10 (accessed May 7, 2015).

Fay, R. and R. Herriot. 1979. "Estimates of Income for Small Places: an Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association* 74: 269–277.

HLG. 2015. "The Role of Big Data in the Modernisation of Statistical Production." Available at: http://www1.unece.org/stat/platform/display/bigdata/2014+Project (accessed 23 March 2015).

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 155–175, http://dx.doi.org/10.1515/JOS-2015-0012

Variance Estimation of Change in Poverty Rates: an Application to the Turkish EU-SILC Survey

Melike Oguz Alper[1] *and Yves G. Berger*[1]

Interpreting changes between point estimates at different waves may be misleading if we do not take the sampling variation into account. It is therefore necessary to estimate the standard error of these changes in order to judge whether or not the observed changes are statistically significant. This involves the estimation of temporal correlations between cross-sectional estimates, because correlations play an important role in estimating the variance of a change in the cross-sectional estimates. Standard estimators for correlations cannot be used because of the rotation used in most panel surveys, such as the European Union Statistics on Income and Living Conditions (EU-SILC) surveys. Furthermore, as poverty indicators are complex functions of the data, they require special treatment when estimating their variance. For example, poverty rates depend on poverty thresholds which are estimated from medians. We propose using a multivariate linear regression approach to estimate correlations by taking into account the variability of the poverty threshold. We apply the approach proposed to the Turkish EU-SILC survey data.

Key words: Linearisation; multivariate regression; stratification; unequal inclusion probabilities.

1. Introduction

In order to monitor progress towards agreed policy goals, particularly in the context of the Europe 2020 strategy, there is an interest in evaluating the evolution of social indicators. In order to interpret changes between indicators at different waves, it is important to estimate the standard error of these changes, so that we can judge whether or not observed changes are statistically significant. The poverty rate is an important policy indicator, especially within the context of the Europe 2020 strategy. This rate is defined as the proportion of people with an equivalised total net income below 60 percent of the national median income (Eurostat 2003, 2). This indicator is calculated from the European Union Statistics on Income and Living Conditions (EU-SILC) surveys (Eurostat 2012) which collect yearly information on income, poverty, social exclusion and living conditions from approximately 300,000 households across Europe. The poverty rate is a complex statistic, unlike population totals or means, since it is based on a poverty threshold computed from the median of the income distribution. Hence, there exist two

[1] University of Southampton, University Road Bldg. 58, Southampton SO17 1BJ, UK, Emails: M.OguzAlper@soton.ac.uk and Y.G.Berger@soton.ac.uk

Acknowledgments: Melike Oguz Alper was funded by the Jean Monnet Scholarship Programme, the European Union, and the Economic and Social Research Council (ESRC), United Kingdom. The authors wish to thank to Turkish Statistical Institute (TurkStat) for providing the datasets used in this article. This work was supported by consulting work for the Net-SILC2 project (Atkinson and Marlier 2010).

sources of variability: one is due to the estimated threshold and the other one comes from the estimated proportion given the estimated threshold (e.g., Berger and Skinner 2003; Verma and Betti 2011).

Several methods to estimate the variance of the poverty rate like resampling and linearisation techniques have been discussed in the literature (e.g., Preston 1995; Deville 1999; Berger and Skinner 2003; Demnati and Rao 2004; Verma and Betti 2005; Osier 2009; Goedemé 2010; Verma and Betti 2011; Muennich and Zins 2011; Osier et al. 2013; Berger and Priam 2015). However, variance of change for the poverty rate has been studied in only limited number of papers (e.g., Betti and Gagliardi 2007; Muennich and Zins 2011; Osier et al. 2013; Berger and Priam 2015). Berger and Priam (2010, 2015) proposed an estimator for the variance of change which takes into account the complexities of the sampling design, such as stratification, unequal probabilities, clustering and rotation (see also Osier et al. 2013). The approach proposed relies neither on the second-order inclusion probabilities nor on the resampling methods, unlike its competitors (Betti and Gagliardi 2007; Wood 2008; Muennich and Zins 2011, 20). It is based on a multivariate linear regression (general linear model) approach that can be easily implemented by any statistical software (Berger and Priam 2015). Berger et al. (2013) show how it can be implemented in SPSS.

The estimator proposed by Berger and Priam (2010, 2015) ignores the sampling variability due to the poverty threshold by treating the poverty rate as a ratio. In Sections 4 and 5, we show how this approach can be adjusted to take into account the sampling variability of the poverty threshold. In Section 6, we compare the approach proposed with the more simple approach proposed by Berger and Priam (2010, 2015; see also Osier et al. 2013) via a series of simulations. In Section 7, we apply the approach proposed to the Turkish EU-SILC survey data. The variance estimator proposed depends on a bandwidth used for the estimation of the density. We also show how sensitive the variance estimates are to the chosen bandwidth parameter by considering different bandwidth parameters.

2. Rotating Sampling Designs

With rotating panel surveys, it is common practice to select new units in order to replace old units that have been in the survey for a specified number of waves (e.g., Gambino and Silva 2009; Kalton 2009). The units sampled on both waves usually represent a large fraction of the first-wave sample. This fraction is called the fraction of the common sample. For example, for the EU-SILC surveys, this fraction is 75 percent. For the Canadian labour force survey and the British labour force survey, this fraction is 80 percent. For the Finnish labour force survey, this fraction is 60 percent. We consider that the sample design is such that the common sample has a fixed number of units. Throughout this article, we assume that the sampling fractions are negligible, that is, $(1 - \pi_{t;i}) \approx 1$, where $\pi_{t;i}$ denote the inclusion probabilities of unit i at wave t.

3. Estimation of Change of a Poverty Rate

Let s_1 and s_2 be the samples selected at Wave 1 and Wave 2 respectively. Suppose, we wish to estimate the absolute net change $\Delta = \theta_2 - \theta_1$ between two population

poverty rates θ_1 and θ_2, from Wave 1 and Wave 2 respectively. Suppose that Δ is estimated by $\hat{\Delta} = \hat{\theta}_2 - \hat{\theta}_1$; where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the cross-sectional estimators of poverty rates defined by

$$\hat{\theta}_1 = \frac{\hat{\tau}_1}{\hat{\tau}_2} = \frac{\sum_{i \in s_1} \delta\{y_{1;i} \leq 0.6\hat{Y}_{1;0.5}\}\pi_{1;i}^{-1}}{\sum_{i \in s_1} \pi_{1;i}^{-1}} \quad \text{and}$$

$$\hat{\theta}_2 = \frac{\hat{\tau}_3}{\hat{\tau}_4} = \frac{\sum_{i \in s_2} \delta\{y_{2;i} \leq 0.6\hat{Y}_{2;0.5}\}\pi_{2;i}^{-1}}{\sum_{i \in s_2} \pi_{2;i}^{-1}},$$

where $y_{t;i}$ is the *net equivalised income* (see Eurostat 2003, 2) of individual i at wave t and $\hat{Y}_{t;0.5}$ is the estimate of the median of the population income distribution at wave t ($t = 1, 2$). The function $\delta\{A\} = 1$ when A is true, and $\delta\{A\} = 0$ otherwise.

The design-based variance of the estimator of change $\hat{\Delta}$ is given by

$$\text{var}(\hat{\Delta}) = \text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2) - 2\,\text{corr}(\hat{\theta}_1, \hat{\theta}_2)\sqrt{\text{var}(\hat{\theta}_1)\text{var}(\hat{\theta}_2)}\,. \tag{1}$$

Standard design-based estimators can be used to estimate the cross-sectional variances $\text{var}(\hat{\theta}_1)$ and $\text{var}(\hat{\theta}_2)$ (e.g., Deville 1999). The correlation $\text{corr}(\hat{\theta}_1, \hat{\theta}_2)$ is the most difficult part to estimate as $\hat{\theta}_1$ and $\hat{\theta}_2$ are estimated from different samples because of the rotation. Estimation of the covariance term has been discussed in several papers (Kish 1965, 457–458; Tam 1984; Laniel 1987; Nordberg 2000; Holmes and Skinner 2000; Berger 2004; Qualité and Tillé 2008; Wood 2008; Muennich and Zins 2011).

Berger and Priam (2010, 2015) proposed a multivariate approach to estimate the correlation between functions of totals by incorporating the information related to the whole sample, $s = s_1 \cup s_2$. This approach can be used to estimate the variance of change between poverty rates when we ignore the sampling variability due to the estimated poverty threshold $0.6\hat{Y}_{t;0.5}$, that is, when we treat the poverty rates as simple ratios.

When we treat the threshold as fixed, the change becomes a smooth function of four totals, that is, $\hat{\Delta} = g(\hat{\tau})$, where $\hat{\tau} = (\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3, \hat{\tau}_4)^\top$ is a vector of four estimated totals. Berger and Priam (2010, 2015) showed that using the first-order Taylor approximation, the design-based variance of $\hat{\Delta}$ can be estimated by

$$\widehat{\text{var}}(\hat{\Delta}) = \widehat{\mathbf{grad}}(\hat{\tau})^\top \widehat{\mathbf{var}}(\hat{\tau})\ \widehat{\mathbf{grad}}(\hat{\tau}), \tag{2}$$

where $\widehat{\mathbf{grad}}(\hat{\tau})$ is the gradient of $g(\hat{\tau})$ evaluated at $\hat{\tau}$, that is,

$$\widehat{\mathbf{grad}}(\hat{\tau}) = \frac{\partial g(\hat{\tau})}{\partial \hat{\tau}} = \left(-\frac{1}{\hat{\tau}_2}, -\frac{\hat{\tau}_1}{\hat{\tau}_2^2}, \frac{1}{\hat{\tau}_3}, -\frac{\hat{\tau}_3}{\hat{\tau}_4^2}\right)^\top,$$

and $\widehat{\mathbf{var}}(\hat{\tau})$ is given by

$$\widehat{\mathbf{var}}(\hat{\tau}) = \hat{\boldsymbol{D}}^\top \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{D}},$$

with

$$\hat{\boldsymbol{D}} = \text{diag}\left\{\sqrt{\widehat{\text{var}}(\hat{\tau}_1)\hat{\Sigma}_{11}^{-1}}, \sqrt{\widehat{\text{var}}(\hat{\tau}_2)\hat{\Sigma}_{22}^{-1}}, \sqrt{\widehat{\text{var}}(\hat{\tau}_3)\hat{\Sigma}_{33}^{-1}}, \sqrt{\widehat{\text{var}}(\hat{\tau}_4)\hat{\Sigma}_{44}^{-1}}\right\},$$

where $\hat{\boldsymbol{\Sigma}}$ is the Ordinary Least Square (OLS) estimator of the residual covariance matrix $\boldsymbol{\Sigma}$

of the multivariate linear regression Model in (3) proposed by Berger and Priam (2010, 2015); $\widehat{\text{var}}(\hat{\tau}_k)$ is the design-based variance estimator of the Horvitz and Thompson (1952) estimator of total τ_k, and $\hat{\Sigma}^{-1}_{kk}$ is the k-th diagonal element of $\hat{\boldsymbol{\Sigma}}$ $(k = 1, 2, 3, 4)$. Berger and Priam (2010, 2015) showed that (2) gives an approximately unbiased estimator for the variance of change.

Let $\check{p}_{t;i} = \delta\{y_{t;i} \leq 0.6\hat{Y}_{t;0.5}\}\pi^{-1}_{t;i}$ and $w_{t;i} = \pi^{-1}_{t;i}$. The multivariate model is given as follows,

$$\begin{pmatrix} \check{p}_{1;i} \\ w_{1;i} \\ \check{p}_{2;i} \\ w_{2;i} \end{pmatrix} = \begin{pmatrix} \alpha_{1;1}z_{1;i} + \alpha_{1;2}z_{2;i} + \alpha_{1;3}z_{1;i} \times z_{2;i} \\ \beta_{1;1}z_{1;i} + \beta_{1;2}z_{2;i} + \beta_{1;3}z_{1;i} \times z_{2;i} \\ \alpha_{2;1}z_{1;i} + \alpha_{2;2}z_{2;i} + \alpha_{2;3}z_{1;i} \times z_{2;i} \\ \beta_{2;1}z_{1;i} + \beta_{2;2}z_{2;i} + \beta_{2;3}z_{1;i} \times z_{2;i} \end{pmatrix} + \boldsymbol{\epsilon}_i. \tag{3}$$

The vector of the residuals $\boldsymbol{\epsilon_i}$ follow a multivariate distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$. Rotation of the sampling design is incorporated into the model through the model covariates: $z_{t;i} = \delta\{i \in s_t\}$ and $z_{1;i} \times z_{2;i} = \delta\{i \in s_1, i \in s_2\}$. It should be noted that the correlations $\widehat{\text{corr}}(\hat{\tau}_k, \hat{\tau}_\ell)$, with $(k, \ell = 1, 2, 3, 4)$, are obtained from the estimated residual covariance matrix $\hat{\boldsymbol{\Sigma}}$. The covariance terms on the nondiagonal part of the matrix $\widehat{\mathbf{var}}(\hat{\tau})$ are based on those estimated correlations $\widehat{\text{corr}}(\hat{\tau}_k, \hat{\tau}_\ell)$ and the estimated cross-sectional variance terms $\widehat{\text{var}}(\hat{\tau}_k)$. Note that this approach also accounts for a multistage sampling, using an "ultimate cluster approach" (e.g., Osier et al. 2013; Di Meglio et al. 2013).

Berger and Priam (2010, 2015) showed that the multivariate approach gives estimates which are approximately equal to the Hansen and Hurwitz (1943) variance estimator (e.g., Holmes and Skinner 2000).

The approach proposed can be easily extended to a stratified sampling. In this case, we assume that the sample sizes within each stratum are fixed (nonrandom) quantities. The model covariates $z_{t;i}$ are replaced by the stratum wave indicators $z_{th;i} = \delta\{i \in s_{th}\}$, where s_{th} is the sample for the stratum h at wave t. As the rotation is done within each stratum, we consider the interactions $z_{th;i} \times z_{(t+1)h;i}$.

4. Allowing for the Variability of the Poverty Threshold

Note that in (2), the variability of the poverty threshold is not taken into account because we treat $\hat{\theta}_1$ and $\hat{\theta}_2$ as ratios. Treating the poverty threshold as fixed might lead to an overestimation of the variances (e.g., Preston 1995; Berger and Skinner 2003; Verma and Betti 2011). Verma and Betti (2011) compared the ratio variance estimator (i.e., when the poverty threshold is treated as fixed) with linearisation and jackknife repeated replication. They found that the ratio variance estimator overestimated the standard errors for all the poverty measures and several complex statistics. However, these findings are related to the cross-sectional estimators and do not necessarily hold for the variance of change.

Taking into account the whole variability means that the sampling variation of the poverty threshold is also considered. However, the poverty rate is more complex than a ratio and cannot be expressed as a function of totals. We propose using the linearisation approach proposed by Deville (1999). The implementation of this approach for the poverty rate and

the inequality measures can be found in the literature (e.g., Berger and Skinner 2003; Verma and Betti 2005; Osier 2009; Muennich and Zins 2011; Verma and Betti 2011).

The linearised variable $L_{t;i}$ for individual i at wave t for the poverty rate is given by (see Osier 2009)

$$L_{t;i} = \frac{1}{\hat{N}_t}\left(\delta\{y_{t;i} \le 0.6\hat{Y}_{t;0.5}\} - \hat{\theta}_t\right) - \frac{0.6\hat{f}_t(0.6\hat{Y}_{t;0.5})}{\hat{N}_t \hat{f}_t(\hat{Y}_{t;0.5})}\left(\delta\{y_{t;i} \le \hat{Y}_{t;0.5}\} - 0.5\right), \quad (4)$$

where $\hat{f}_t(.)$ is an estimator of the density function, which is defined in (5). The second term in (4) is an additional term which reflects the sample variation originating from the randomness of the estimated median income.

The density functions can be estimated on the basis of the Gaussian kernel function as follows (e.g., Preston 1995):

$$\hat{f}_t(x) = \frac{1}{\hat{N}_t \hat{h}_t}\sum_{i \in s_t}\frac{1}{\pi_{t;i}} K\left(\frac{x - y_{t;i}}{\hat{h}_t}\right), \quad (5)$$

where $K(\eta) = (\sqrt{2\pi})^{-1}\exp(-\eta^2/2)$ is the Gaussian kernel, $\hat{N}_t = \sum_{i \in s_t}\pi_{t;i}^{-1}$ is the Horvitz and Thompson (1952) estimator of the population size at wave $t\,(t = 1, 2)$, and $\hat{h}_t$ is the bandwidth parameter, which can be defined in several ways (Silverman 1986, 45–48). For a normally distributed population and smooth densities, the following bandwidth parameter was recommended by Silverman (1986, 46):

$$\hat{h}_t = 1.06\hat{\sigma}_{t;\hat{Y}}\hat{N}_t^{-1/5}, \quad (6)$$

where

$$\hat{\sigma}_{t;\hat{Y}} = \sqrt{\frac{1}{\hat{N}_t}\left\{\sum_{i \in s_t}\frac{1}{\pi_{t;i}} y_{t;i}^2 - \frac{1}{\hat{N}_t}\left(\sum_{j \in s_t}\frac{1}{\pi_{t;j}} y_{t;j}\right)^2\right\}}$$

is the estimated standard deviation of the income distribution. However, for skewed and long-tailed distributions, Silverman (1986, 47) proposed using the interquartile range instead of the standard deviation of the distribution, that is,

$$\hat{h}_t = 0.79\hat{Y}_{t;iqr}\hat{N}_t^{-1/5}, \quad (7)$$

where $\hat{Y}_{t;iqr} = \hat{Y}_{t;0.75} - \hat{Y}_{t;0.25}$ is the weighted interquartile range of the income distribution. Another bandwidth, which is very suitable for many densities, even for the modest bimodal ones, was suggested by Silverman (1986, 48) as follows:

$$\hat{h}_t = 0.9\hat{A}_t\hat{N}_t^{-1/5}, \quad (8)$$

where $\hat{A}_t = \min(\hat{\sigma}_{t;\hat{Y}}, \hat{Y}_{t;iqr}/1.34)$. It should be noted that the bandwidth in (8) is smaller than the other bandwidths in (6) and (7). Thus we are likely to obtain less smooth densities with the bandwidth (8).

It is worth mentioning that choosing a bandwidth parameter is a crucial step in applications (e.g., Verma and Betti 2005; Graf 2013; Graf and Tillé 2014). For example, Verma and Betti (2005) showed that probability density functions are sensitive to the

chosen bandwidth parameter. A large value for the bandwidth parameter results in a smoother density. Graf (2013, 26–28) pointed out the potential danger of using standard deviation when estimating densities that might arise from extreme values in the data observed (for example, with income data). In such cases, Graf (2013) proposed using the logarithm to reduce the adverse impact of extreme values. He also remarked the fixed-bandwidth parameter might be problematic when observations are heaped up around some values. To avoid this problem, a more robust technique to estimate density involving nearest neighbours with minimal bandwidth was suggested by Graf (2013).

5. Estimation of Change Within Domains

In practice, we are often interested in change within domains of interest. For example, we may be interested in change in poverty within different age groups. According to the definition given by Eurostat (2003), the poverty threshold is calculated based on the overall estimated median income rather than the estimated median income within the domains. Hence, when we are interested in a domain, the threshold will be the same for all domains.

Consider $d_{t;i}$ to be a domain indicator for individual i at wave t defined by

$$d_{t;i} = \begin{cases} 1 & \text{if } i \in D \text{ at wave } t, \\ 0 & \text{if } i \notin D \text{ at wave } t, \end{cases}$$

where D refers to the domain of interest. The poverty rate over a domain is defined by

$$\hat{\theta}_{Dt} = \frac{\sum_{i \in s_t} d_{t;i} \delta\{y_{t;i} \leq \hat{Y}_{t;0.5}\} \pi_{t;i}^{-1}}{\sum_{i \in s_t} d_{t;i} \pi_{t;i}^{-1}}.$$

To estimate the variance of change within domains under the ratio approach (see (2)), we substitute $\breve{p}_{t;i}$ by $\breve{p}_{Dt;i} = d_{t;i}\breve{p}_{t;i}$, and $\omega_{t;i}$ by $\omega_{Dt;i} = d_{t;i}\omega_{t;i}$ in the model in (3). Note that the values of the response variables will be equal to zero for the units not included in the domain of interest.

For the linearisation approach, the linearised variables $L_{Dt;i}$ for individual i in domain D at wave t derived in Appendix B (see B.5) are given by

$$L_{Dt;i} = \frac{d_{t;i}}{\hat{N}_{Dt}} \left(\delta\{y_{t;i} \leq 0.6\hat{Y}_{t;0.5}\} - \hat{\theta}_{Dt} \right) - \frac{0.6\hat{f}_{Dt}(0.6\hat{Y}_{t;0.5})}{\hat{N}_t \hat{f}_t(\hat{Y}_{t;0.5})} \left(\delta\{y_{t;i} \leq \hat{Y}_{t;0.5}\} - 0.5 \right),$$

where

$$\hat{N}_{Dt} = \sum_{i \in s_t} \frac{d_{t;i}}{\pi_{t;i}},$$

$$\hat{f}_{Dt}(x) = \frac{1}{\hat{N}_{Dt}\hat{h}_{Dt}} \sum_{i \in s_t} \frac{d_{t;i}}{\pi_{t;i}} K_D\left(\frac{x - y_{t;i}}{\hat{h}_{Dt}} \right).$$

Here, $\hat{h}_{Dt}$ can be (6), (7), or (8) with $\hat{N}_{Dt}$, $\hat{Y}_{Dt;iqr} = \hat{Y}_{Dt;0.75} - \hat{Y}_{Dt;0.25}$,

$$\hat{\sigma}_{Dt;\hat{Y}} = \sqrt{\frac{1}{\hat{N}_{Dt}}\left\{\sum_{i\in s_t}\frac{d_{t;i}}{\pi_{t;i}}y_{t;i}^2 - \frac{1}{\hat{N}_{Dt}}\left(\sum_{j\in s_t}\frac{d_{t;j}}{\pi_{t;j}}y_{t;j}\right)^2\right\}},$$

and $\hat{A}_{Dt} = \min(\hat{\sigma}_{Dt;\hat{Y}}, \hat{Y}_{Dt;iqr}/1.34)$. Let $\hat{\Delta}_D = \hat{\theta}_{D2} - \hat{\theta}_{D1}$ be the change in poverty rate in domain D between Wave 1 and Wave 2. Thus the variance of domain change is estimated by

$$\widehat{\text{var}}(\hat{\Delta}_D) = \widehat{\text{var}}\left(\hat{\theta}_{D1}^L\right) + \widehat{\text{var}}\left(\hat{\theta}_{D2}^L\right) - 2\widehat{\text{corr}}\left(\hat{\theta}_{D1}^L, \hat{\theta}_{D2}^L\right)\sqrt{\widehat{\text{var}}\left(\hat{\theta}_{D1}^L\right)\widehat{\text{var}}\left(\hat{\theta}_{D2}^L\right)}, \quad (9)$$

with

$$\hat{\theta}_{Dt}^L = \sum_{i\in s_t}\frac{L_{Dt;i}}{\pi_{t;i}}. \quad (10)$$

We use the approach proposed by Berger and Priam (2010, 2015) by treating $\hat{\theta}_{D1}^L$ and $\hat{\theta}_{D2}^L$ in (10) as the estimators of totals. The correlation term $\widehat{\text{corr}}(\hat{\theta}_{D1}^L, \hat{\theta}_{D2}^L)$ in (9) is computed from the estimated residual covariance matrix $\hat{\boldsymbol{\Sigma}}$ of the following model,

$$\begin{pmatrix}\breve{L}_{D1;i}\\ \breve{L}_{D2;i}\end{pmatrix} = \begin{pmatrix}\alpha_{1;1}z_{1;i} + \alpha_{1;2}z_{2;i} + \alpha_{1;3}z_{1;i}\times z_{2;i}\\ \alpha_{2;1}z_{1;i} + \alpha_{2;2}z_{2;i} + \alpha_{2;3}z_{1;i}\times z_{2;i}\end{pmatrix} + \boldsymbol{\epsilon}_i,$$

with $\breve{L}_{Dt;i} = L_{Dt;i}\pi_{t;i}^{-1}$.

It should be noted that the domain information is incorporated into the model through the response variables, in contrast to the stratification (see Section 3). Note that the approach proposed can be used for strata domains and unplanned domains.

6. Simulation Study

In this section, the variance estimators from the ratio and the linearisation approaches are compared in terms of the relative bias (RB) and the root mean square error (RRMSE), respectively defined by (11) and (12). Additionally, we investigate whether the ratio approach gives more conservative estimates.

The income variables at Wave 1 and Wave 2 were generated according to different probability distributions (see Appendix A). For each wave, a gamma distribution (shape = 2.5, rate = 1), a lognormal distribution (mean = 1.119, standard deviation = 0.602) and a Weibull distribution (shape = 0.8, scale = 1) were used to generate populations with a size of N = 20,940. As stated by Salem and Mount (1974) and McDonald (1984), these distributions are good approximations of income distributions. The correlation coefficient between the variables of the first and the second wave is given by $\rho = 0.94$, which is the correlation observed from the common sample of the Turkish EU-SILC survey data. Note that this correlation and the correlation in (1) are different; in other words, the correlation $\rho = 0.94$ is the correlation between the variables of interest, whereas the correlation in (1) is the correlation between the point estimators.

The population is assumed fixed and the same sample size was used for both waves. We have 1,047 primary sampling units in the Turkish EU-SILC survey data. For this reason, we used $n_1 = n_2 = 1,047$ units for each wave. The fraction of the common sample is 75 percent. Hence, the number of units in the common sample is $n_c = 785$. Unequal and equal probabilities were used to select the samples. The Chao (1982) sampling design was used as unequal probability sampling (πps) design. The first-wave samples were selected without replacement with the inclusion probabilities proportional to a size variable x_i, which was generated by the model $x_i = \alpha + \rho y_{1;i} + e_i$, with $e_i \sim N(0, (1-\rho^2)\sigma^2_{y1})$, $\alpha = 5$ and $\rho = 0.7$. For the second wave, a simple random sample of n_c units were selected from the sample s_1; and $n_2 - n_c$ units were selected with the probabilities proportional to size $q_i = \pi_{1;i}/(1-\pi_{1;i})$ from the population $U \backslash s_1$. It can be shown that $\pi_{2;i} \approx \pi_{1;i}$ (Christine and Rocher 2012). For equal probability sampling designs, $\pi_{2;i} = \pi_{1;i} = n_1/N$.

We did six simulation studies for three populations and two sampling designs. For each simulation, 10,000 samples were selected. For each sample, the RB and the RRMSE were computed for the cross-sectional variance estimators, the variance estimator of change and the estimator of the correlation. The RB and the RRMSE are defined by

$$RB(\hat{\sigma}) = \frac{E(\hat{\sigma}) - \sigma}{\sigma} 100\%, \tag{11}$$

$$RRMSE(\hat{\sigma}) = \frac{\sqrt{(B-1)^{-1}\sum_{b=1}^{B}(\hat{\sigma}_b - \sigma)^2}}{\sigma} 100\%, \tag{12}$$

where $E(\hat{\sigma}) = B^{-1}\sum_{b=1}^{B}\hat{\sigma}_b$, with $B = 10{,}000$, is the empirical expectation; σ is either the empirical variance or the empirical correlation in (1); $\hat{\sigma}$ is the estimator of the quantity σ; $\hat{\sigma}_b$ is the estimate of the quantity σ for the b-th sample. For the linearisation, we considered three bandwidth parameters (see (6), (7) and (8)). The linearisations based on (6), (7) and (8) are respectively labelled Lin_Sd, Lin_Iqr, and Lin_A in Table 1 and Table 2.

For a gamma distribution, the poverty rates are 24.2 percent and 23.6 percent for the first and the second wave respectively. Hence, we have -0.59 percentage point change between two waves. For a lognormal distribution, the poverty rates are 19.4 percent and 19.9 percent. Thus there is a 0.54 percentage point change in this case. For a Weibull distribution, we have the highest poverty rates, which are 36.6 percent and 37.3 percent respectively. Hence, the change is 0.66 percentage points.

Table 1 shows the RB (%) of the variance and the correlation estimators for several distributions and sampling designs. Overall, the linearisation approach has lower RB compared to the ratio approach. Thus we achieve more accurate estimates with linearisation. Differences between the two approaches in terms of the RB are much more pronounced for the Weibull distribution, which is the most skewed distribution. For all situations except with the lognormal distribution, the ratio approach overestimates all the variances and the correlations. Therefore, the ratio approach may not always provide more conservative estimates. However, note that whenever we have a positive bias, we obtain relatively larger variance estimates with the ratio approach. When we compare the three linearisation methods based on different bandwidth, we obtained the largest RB with the smallest bandwidth (see (8)).

Table 1. Empirical RB (%) of the variance and correlation estimators for the poverty rates for three distributions and two sampling designs

Relative bias (%)								
	Gamma							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	41.3	2.4	2.6	3.1	50.9	7.2	7.4	7.8
Var Wave2	42.8	5.1	5.3	5.8	41.1	2.9	3.0	3.5
Var Change	8.1	1.0	1.2	1.8	13.0	2.1	2.4	2.9
Correlation	23.2	2.6	2.6	2.5	22.0	2.7	2.6	2.5
	Lognormal							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	15.6	0.9	2.2	2.9	22.7	−0.5	0.5	1.0
Var Wave2	24.1	6.4	7.6	8.2	28.9	4.2	5.1	5.6
Var Change	−14.1	1.3	2.6	3.4	−8.7	0.5	1.7	2.4
Correlation	38.1	3.1	2.9	2.8	35.5	1.6	1.3	1.1
	Weibull							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	140.1	4.3	6.5	6.7	132.9	2.9	4.8	5.1
Var Wave2	146.0	1.9	4.0	4.2	137.6	1.0	2.9	3.1
Var Change	26.6	0.9	4.2	4.4	28.3	2.0	5.2	5.5
Correlation	152.0	6.6	3.3	3.3	132.1	−0.4	−3.8	−3.8

As far as the RRMSE is concerned (see Table 2), we achieve more precise estimates with the linearisation approach. We observe the smallest RRMSE with bandwidth (6) and the largest RRMSE with bandwidth (8). The ratio approach provides less accurate point estimates. However, the differences between the two approaches can be negligible for the variance of change, except with the Weibull distribution.

7. An Application to the Turkish EU-SILC Survey

The 2007 and 2008 cross-sectional Turkish EU-SILC survey data was used. The Turkish EU-SILC survey has a stratified two-stage cluster probability sampling design. For the first stage, address blocks are selected within each stratum with a probability proportional to size (πps) without replacement sampling design. Each block is composed of approximately 100 addresses. Households within the selected address blocks are selected using a systematic sampling design. All individuals within the selected households participate in the survey. The cross-sectional survey weights in the "personal register" file (RB050) were used as inverses of the inclusion probabilities. The effect of calibration was not taken into account because we did not have any information about the auxiliary variables. The effect of imputation was ignored for the same reason.

Table 2. Empirical RRMSE (%) of the variance and correlation estimators for the poverty rates for three distributions and two sampling designs

Relative root mean square error (%)								
	Gamma							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	41.5	4.8	5.1	5.9	51.2	8.2	8.4	9.0
Var Wave2	36.8	6.8	7.1	7.9	41.4	4.9	5.1	5.9
Var Change	10.9	7.9	8.1	8.6	15.4	8.7	8.8	9.4
Correlation	20.0	6.0	6.5	6.5	22.7	7.3	7.3	7.3
	Lognormal							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	16.4	4.9	6.2	7.2	30.6	7.8	8.2	8.7
Var Wave2	24.6	8.1	9.8	10.8	35.2	8.9	9.8	10.5
Var Change	15.1	7.0	8.0	8.8	18.5	10.6	11.2	11.7
Correlation	38.5	7.1	7.0	7.0	37.4	11.1	11.0	11.0
	Weibull							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	140.1	6.5	8.5	9.0	133.0	6.5	8.0	8.5
Var Wave2	146.0	5.5	7.1	7.7	137.7	6.2	7.4	7.9
Var Change	27.1	5.7	7.8	8.4	29.1	7.0	9.3	9.9
Correlation	152.2	16.7	16.3	16.6	132.4	16.2	17.2	17.5

Table 3. Estimates when the poverty threshold is treated as fixed (see (2))

Domain	Pov'07(%)	Var'07	Pov'08	Var'08(%)	Change (in % point)	Var Change	Corr	*p*-value
Turkey	23.4	0.616	24.1	0.644	0.7	0.447	0.65	0.297
Male	23.0	0.650	23.7	0.665	0.7	0.494	0.62	0.328
Female	23.8	0.639	24.6	0.678	0.7	0.465	0.65	0.299
Owner	24.9	0.739	23.8	0.872	−1.1	0.593	0.63	0.140
Tenant	18.5	1.395	25.3	1.511	6.7	1.522	0.48	0.000
0–14	33.5	1.164	34.5	1.258	1.1	0.882	0.64	0.263
15–24	24.2	1.162	25.3	1.181	1.1	1.118	0.52	0.296
25–49	19.8	0.527	20.7	0.548	0.9	0.405	0.62	0.178
50–64	14.4	0.568	15.0	0.719	0.6	0.569	0.56	0.404
65 +	17.7	1.077	16.2	0.929	−1.5	0.988	0.51	0.120

Source: 2007 and 2008 cross-sectional data of the EU-SILC survey for Turkey conducted by TurkStat.

Table 4. Estimates when the sampling variation of the poverty threshold is taken into account (see Sections 4 and 5). The bandwidth parameter is based on the standard deviation of the income distribution (see (6)).

Domain	Pov'07(%)	Var'07	Pov'08	Var'08(%)	Change (in % point)	Var Change	Corr	*p*-value
Turkey	23.4	0.292	24.1	0.290	0.7	0.372	0.36	0.252
Male	23.0	0.314	23.7	0.306	0.7	0.416	0.33	0.287
Female	23.8	0.327	24.6	0.327	0.7	0.390	0.40	0.257
Owner	24.9	0.417	23.8	0.495	−1.1	0.527	0.42	0.117
Tenant	18.5	1.121	25.3	1.238	6.7	1.435	0.39	0.000
0–14	33.5	0.796	34.5	0.793	1.1	0.787	0.50	0.236
15–24	24.2	0.790	25.3	0.919	1.1	1.050	0.39	0.281
25–49	19.8	0.255	20.7	0.252	0.9	0.362	0.29	0.154
50–64	14.4	0.403	15.0	0.491	0.6	0.476	0.47	0.361
65 +	17.7	0.929	16.2	0.807	−1.5	0.978	0.44	0.118

Source: 2007 and 2008 cross-sectional data of the EU-SILC survey for Turkey conducted by TurkStat.

Table 5. Estimates when the sampling variation of the poverty threshold is taken into account (see Sections 4 and 5). The bandwidth parameter is based on the interquartile range of the income distribution (see (7)).

Domain	Pov'07(%)	Var'07	Pov'08	Var'08(%)	Change (in % point)	Var Change	Corr	*p*-value
Turkey	23.4	0.292	24.1	0.290	0.7	0.372	0.36	0.252
Male	23.0	0.316	23.7	0.306	0.7	0.416	0.33	0.287
Female	23.8	0.325	24.6	0.328	0.7	0.391	0.40	0.257
Owner	24.9	0.418	23.8	0.497	−1.1	0.530	0.42	0.118
Tenant	18.5	1.117	25.3	1.226	6.7	1.428	0.39	0.000
0–14	33.5	0.802	34.5	0.814	1.1	0.805	0.50	0.241
15–24	24.2	0.787	25.3	0.907	1.1	1.038	0.39	0.278
25–49	19.8	0.256	20.7	0.251	0.9	0.361	0.29	0.154
50–64	14.4	0.403	15.0	0.491	0.6	0.476	0.47	0.361
65 +	17.7	0.946	16.2	0.791	−1.5	0.976	0.44	0.118

Source: 2007 and 2008 cross-sectional data of the EU-SILC survey for Turkey conducted by TurkStat.

Table 6. Estimates when the sampling variation of the poverty threshold is taken into account (see Sections 4 and 5). The bandwidth parameter is based on parameter A (see (8)).

Domain	Pov'07(%)	Var'07	Pov'08	Var'08(%)	Change (in % point)	Var Change	Corr	*p*-value
Turkey	23.4	0.291	24.1	0.291	0.7	0.372	0.36	0.253
Male	23.0	0.316	23.7	0.306	0.7	0.416	0.33	0.287
Female	23.8	0.324	24.6	0.329	0.7	0.392	0.40	0.258
Owner	24.9	0.419	23.8	0.498	−1.1	0.531	0.42	0.119
Tenant	18.5	1.114	25.3	1.223	6.7	1.425	0.39	0.000
0–14	33.5	0.802	34.5	0.823	1.1	0.812	0.50	0.243
15–24	24.2	0.787	25.3	0.903	1.1	1.034	0.39	0.277
25–49	19.8	0.255	20.7	0.251	0.9	0.361	0.29	0.154
50–64	14.4	0.403	15.0	0.491	0.6	0.476	0.47	0.361
65 +	17.7	0.949	16.2	0.788	−1.5	0.977	0.44	0.118

Source: 2007 and 2008 cross-sectional data of the EU-SILC survey for Turkey conducted by TurkStat.

In Table 3, we give the estimates for several domains when the poverty threshold is treated as fixed (see (2)). We observe a significant change for the domain "tenant" at the 5 percent level.

In Table 4, we give the estimates obtained with the linearisation approach based on the bandwidth in (6) described in Section 4. Here, we again observe a highly significant change for the domain "tenant". We do not observe major differences in the p-values between Table 3 and Table 4. We observe a slight decrease in the p-values when the sampling variation of the poverty threshold is taken into account. This is due to the fact that the variances of changes are larger in Table 3.

The correlations in Table 4 are smaller than in Table 3 overall. Hence, the estimated correlations are smaller when the variability of the poverty threshold is taken into account.

The comparison of Table 3 and Table 4 also revealed that all variances were estimated more conservatively when the threshold is treated as fixed. Preston (1995), Berger and Skinner (2003), and Verma and Betti (2011) demonstrated that the cross-sectional variances are more conservative when the poverty threshold is treated as fixed. This finding was explained by Preston (1995) by the fact that the two sources of variability offset each other. This is more pronounced when the high fractions of the median are used.

For the variance of change, we cannot anticipate an increase in the variance when the poverty threshold is treated as fixed for the following reason. Let us assume that the cross-sectional variances are equal: $\widehat{\text{var}}(\hat{\theta}_1) = \widehat{\text{var}}(\hat{\theta}_2)$. Thus the variance estimator of change is given by $\widehat{\text{var}}(\hat{\Delta}_1) = 2\widehat{\text{var}}(\hat{\theta}_1)(1 - \widehat{\text{corr}}\,(\hat{\theta}_1, \hat{\theta}_2))$. Hence, the variance of change is affected in the same direction by the variance term, and in the opposite direction by the correlation term. Accordingly, when both the variance and the correlation terms increase or decrease concurrently, the direction of the effect on the variance of change cannot be predicted. Therefore, we may not necessarily have more conservative estimates of the variance of change when the poverty threshold is treated as fixed. With the Turkish EU-SILC survey data, we found that the variances of changes were more conservative, although the differences between the two approaches were not as pronounced as the differences between the cross-sectional variances (see Table 3 and Table 4).

In Table 4, the bandwidth parameter is given by (6). We also investigate the situations when the bandwidth parameter is given by (7) and (8). The results are given in Table 5 and Table 6. By comparing Table 5 and Table 6 with Table 3, we also observed smaller cross-sectional variances, variance of change and correlation when the bandwidth parameter is (7) and (8). When we compare Table 4, Table 5, and Table 6, the estimates do not differ significantly between the three linearisation approaches based on different bandwidth parameters, although we observe slight differences between them in terms of the RB and the RRMSE in the simulation study (see Section 6).

8. Conclusion

We applied a simple approach to estimate the variances of changes for the poverty rates over several domains by using the 2007–2008 Turkish EU-SILC survey data. Our approach involves a multivariate linear regression model proposed by Berger and Priam (2010, 2015), which can be easily applied. Survey characteristics such as rotation, stratification, and cluster sampling are all taken into account. The approach proposed is flexible and can be implemented for most of the EU-SILC surveys as long as sampling

fractions are negligible. This assumption implies that the second-order inclusion probabilities are not needed.

We have two ways of estimating the variances, depending on whether we treat the poverty threshold as fixed or not. When treated as fixed, we obtained more conservative variance estimates of change with the Turkish EU-SILC survey data. However, our simulation study shows that treating the threshold as fixed does not necessarily provide more conservative variance estimates of change. For the lognormal distribution, for example, variances of changes were underestimated with the ratio method. On the other hand, differences between the variance estimators of changes can be negligible in terms of the RB and the RRMSE, even though we observed significant differences between the cross-sectional variances and the correlations. For the latter, the linearisation approach gave more unbiased and more precise variance estimates. Thus based upon our results and due to the fact that linearisation involves complex numerical computations, the simple ratio approach may appear preferable to estimate the variance of change for the poverty rates. However, we should be careful with highly skewed distributions similar to a Weibull one. In this case, the linearisation approach is significantly better.

The approach proposed can also be used to estimate the variances of the other poverty and income inequality measures such as the relative median at-risk-of-poverty gap (RMPG), the quantile share ratio (QSR) and the Gini coefficient, which are included in the "Laeken" indicators (Eurostat 2003), by using linearisation (e.g., Berger 2008). The RMPG and the Gini coefficient can not be treated as a simple ratio, whereas the QSR can be. The linearised variables of many complex parameters are given by Verma and Betti (2005, 2011).

In this article, we implemented the fixed-bandwidth kernel method for its simplicity (Silverman 1986, 95). Note that the bandwidth in (8) is a suitable choice for a wide range of densities, as pointed out by Silverman (1986). If the distribution is heavily skewed, then an adaptive kernel method can be applied (Silverman 1986, chap. 5). This method uses a variable bandwidth, that is, for each observed data point, a different bandwidth is computed. It would be interesting to check whether an adaptive bandwidth improved the variance estimation in the presence of outliers.

Appendix A. Generation of the Income Variables for the Simulation Study

For the gamma random variables, we used the algorithm proposed by Schmeiser and Lal (1982, 358). First, three independent random variables were generated by a gamma distribution as follows:

$$Y_1 \sim Gamma\left(\alpha_1 - \rho\sqrt{\alpha_1}\sqrt{\alpha_2}, 1\right),$$
$$Y_2 \sim Gamma\left(\alpha_2 - \rho\sqrt{\alpha_1}\sqrt{\alpha_2}, 1\right),$$
$$Y_3 \sim Gamma\left(\rho\sqrt{\alpha_1}\sqrt{\alpha_2}, 1\right),$$

with $\alpha_1 = 2.5$, $\alpha_2 = 2.6$, and $\rho = 0.94$. Then, the income variables were obtained by the following expressions: $y_{1;i} = Y_1 + Y_3$ and $y_{2;i} = Y_2 + Y_3$, so that $y_{1;i} \sim Gamma(2.5, 1)$, $y_{2;i} \sim Gamma(2.6, 1)$, and $\rho(y_{1;i}, y_{2;i}) \approx 0.94$.

The Cholesky decomposition was used to generate the correlated lognormal variables. Hence, the log income variables with the correlation of $\rho = 0.95$, a mean of $\mu = 1.119$

and a standard deviation of $\sigma = 0.602$ were generated by

$$\log(y_{1;i}) = \mu + \sigma X_1,$$

$$\log(y_{2;i}) = \mu + \rho\sigma X_1 + \sqrt{1-\rho^2}\sigma X_2,$$

where X_1 and X_2 are independent standard normal variables. The correlation coefficient between the income variables was approximately 0.94.

For correlated Weibull variables, we followed the algorithm proposed by Feiveson (2002, 117). Firstly, two correlated standard normal variables Y_1 and Y_2 with a correlation of $\rho = 0.95$ were generated by using the Cholesky decomposition: $Y_1 = X_1$ and $Y_2 = \rho X_1 + \sqrt{1-\rho^2}X_2$, where X_1 and X_2 are independent standard normal variables. Secondly, correlated uniform variables were obtained by the standard normal cumulative distribution function transformation; such that $U_1 = \Phi(Y_1)$ and $U_2 = \Phi(Y_2)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Finally, uniform random variables were transformed by the inverse of the Weibull cumulative distribution function to achieve the correlated income variables as follows: $y_{1;i} = \mathrm{F}_U^{-1}(U_1) = (-\ln(1-U_1))^{5/4}$ and $y_{2;i} = \mathrm{F}_U^{-1}(U_2) = (-\ln(1-U_2))^{5/4}$, so that $y_{1;i}, y_{2;i} \sim$ *Weibull* $(0.8, 1)$ and $\rho(y_{1;i}, y_{2;i}) \approx 0.94$.

Appendix B. Derivation of the Influence Function of the Poverty Rate Over a Domain

Let M be a measure that assigns a unit mass to each unit i in the population U. For example, the population size N can be written as $N = \int d\mathrm{M} = \sum_{i\in U} 1$ and the total of a variable y can be expressed as $N = \int y d\mathrm{M} = \sum_{i\in U} y_i$ (Deville 1999). Let $\mathrm{F}(\mathrm{M}, x)$ be the income distribution function at x over the population U, that is,

$$\mathrm{F}(\mathrm{M},\, x) = \frac{1}{N}\sum_{i\in U}\delta\{y_i \le x\}.$$

Then, the income distribution function at the median of the income distribution is given by $\mathrm{F}(\mathrm{M}, Med(\mathrm{M})) = 0.5$. Thus the influence function of the functional $\mathrm{F}(\mathrm{M}, Med(\mathrm{M}))$ at i is equal to 0, that is, $\mathrm{IF}_i(\mathrm{M}, Med(\mathrm{M})) = 0$. By using "Rule 7" in Deville (1999, 198), the influence function of F at i (see also Osier 2009, 181–183) can be derived as follows:

$$\mathrm{IF}_i(\mathrm{M}, Med(\mathrm{M})) = \mathrm{IF}_i(\mathrm{M}, Med(\mathrm{M})|_{Med(\mathrm{M})\text{ fixed}}) + \frac{\partial \mathrm{F}(\mathrm{M}, x)}{\partial x}|_{x=Med(\mathrm{M})}\mathrm{IMed}_i(\mathrm{M}) = 0. \quad \text{(B.1)}$$

The influence function of F, when the median is fixed, is given by

$$\mathrm{IF}_i(\mathrm{M}, Med(\mathrm{M})|_{Med(\mathrm{M})\text{ fixed}}) = \frac{1}{N}[\delta\{y_i \le Med\} - 0.5].$$

Thus the influence function of the functional $Med(\mathrm{M})$ is obtained as

$$\mathrm{IMed}_i(\mathrm{M}) = -\frac{1}{N}\frac{1}{f(Med)}[\delta\{y_i \le Med\} - 0.5], \quad \text{(B.2)}$$

where

$$f(Med) = \frac{\partial \mathrm{F}(\mathrm{M}, x)}{\partial x}|_{x=Med(\mathrm{M})}$$

is the probability density function at the median of the income distribution.

Now define the income distribution function at x over a domain D as follows:

$$\mathrm{F}_D(\mathrm{M}, x) = \frac{1}{N_D}\sum_{i\in U} d_i \delta\{y_i \le x\}.$$

Hence, the income distribution function over a domain D at the poverty threshold T is defined by

$$\mathrm{F}_D(\mathrm{M},\ T(\mathrm{M})) = \frac{1}{N_D}\sum_{i\in U} d_i \delta\{y_i \le T(\mathrm{M})\},$$

where $T(\mathrm{M}) = 0.6Med(\mathrm{M})$ and d_i is the domain indicator, that is, 1 when $i \in D$, and 0 otherwise. $\mathrm{F_D}(\mathrm{M},\ T(\mathrm{M}))$ is equivalent to the poverty rate over a domain D (i.e., R_D). Thus we can obtain the influence function of the poverty rate analogously to (B.1), that is,

$$\mathrm{IF}_{D;i}(\mathrm{M}, T(\mathrm{M})) = \mathrm{IF}_{D;i}(\mathrm{M}, T(\mathrm{M}))|_{T(\mathrm{M})\text{ fixed}}) + \frac{\partial \mathrm{F}_D(\mathrm{M}, x)}{\partial x}|_{x=T(\mathrm{M})}\mathrm{IT}_i(\mathrm{M}) = \mathrm{IR}_{D;i}.$$

The influence function of F_D, when the threshold is fixed, is given by

$$\mathrm{IF}_{D;i}(\mathrm{M}, T(\mathrm{M}))|_{T(\mathrm{M})\text{ fixed}} = \frac{d_i}{N_D}[\delta\{y_i \le T\} - R_D].$$

Hence, the influence function of the poverty rate is obtained as follows:

$$\mathrm{IR}_{D;i} = \frac{d_i}{N_D}[\delta\{y_i \le T\} - R_D] + f_D(T)\mathrm{IT}_i(\mathrm{M}), \tag{B.3}$$

where

$$f_D(T) = \frac{\partial \mathrm{F}_D(\mathrm{M}, x)}{\partial x}|_{x=T(\mathrm{M})}$$

is the probability density function at the poverty threshold. The influence function of the functional T(M) at i is given by

$$\mathrm{IT}_i(\mathrm{M}) = 0.6\mathrm{IMed}_i(\mathrm{M}). \tag{B.4}$$

If we substitute $\mathrm{IMed}_i(\mathrm{M})$ in (B.2) into (B.4), we obtain the following:

$$\mathrm{IT}_i(\mathrm{M}) = -\frac{0.6}{N}\frac{1}{f(Med)}[\delta\{y_i \le Med\} - 0.5].$$

Therefore, the influence function of the poverty rate at i over a domain D given in (B.3) can be rewritten as follows:

$$\mathrm{IR}_{D;i} = \frac{d_i}{N_D}[\delta\{y_i \le T\} - R_D] - \frac{0.6}{N}\frac{f_D(T)}{f(Med)}[\delta\{y_i \le Med\} - 0.5]. \tag{B.5}$$

Note that we assume the derivatives of F and F_D exist and are strictly non-negative for all x.

9. References

Atkinson, A.B. and E. Marlier. 2010. "*Income and Living Conditions in Europe*." Publications Office of the European Union, Luxembourg. Available at: http://ec.europa.eu/eurostat/documents/3217494/5722557/KS-31-10-555-EN.PDF/e8c0a679-be01-461c-a08b-7eb08a272767 (accessed April 30, 2015).

Berger, Y.G. 2004. "Variance Estimation for Measures of Change in Probability Sampling." *Canadian Journal of Statistics* 32: 451–467. DOI: http://dx.doi.org/10.2307/3316027.

Berger, Y.G. 2008. "A Note on the Asymptotic Equivalence of Jackknife and Linearization Variance Estimation for the Gini Coefficient." *Journal of Official Statistics* 24: 541–555.

Berger, Y.G., T. Goedemé, and G. Osier. 2013. *Handbook on Standard Error Estimation and Other Related Sampling Issues in EU-SILC Second Network for the Analysis of EU-SILC, EuroStat.* Available at: http://www.cros-portal.eu/content/handbook-standard-error-estimation-and-other-related-sampling-issues-ver-29072013 (accessed February 6, 2013).

Berger, Y.G. and R. Priam. 2010. "Estimation of Correlations between Cross-Sectional Estimates from Repeated Surveys – an Application to the Variance of Change." In Proceedings of the 2010 Symposium of Statistics Canada, [26–29 October, 2010]. [10 pp.].

Berger, Y.G. and R. Priam. 2015. *A Simple Variance Estimator of Change for Rotating Repeated Surveys: an Application to the EU-SILC Household Surveys. Journal of the Royal Statistical Society, Series A (Statistics in Society)*. Available at: DOI: http://dx.doi.org/10.1111/rssa.12116. 22 pp. (accessed June 9, 2015).

Berger, Y.G. and C.J. Skinner. 2003. "Variance Estimation of a Low-Income Proportion." *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 52: 457–468. DOI: http://dx.doi.org/10.1111/1467-9876.00417.

Betti, G. and F. Gagliardi. 2007. "Jackknife Variance Estimation of Differences and Averages of Poverty Measures." Working Paper 68, Siena: Dipartimento di Metodi Quantitativi, Università degli Studi.

Chao, M.T. 1982. "A General Purpose Unequal Probability Sampling Plan." *Biometrika* 69: 653–656. DOI: http://dx.doi.org/10.1093/biomet/69.3.653.

Christine, M. and T. Rocher. 2012. "Construction d'échantillons astreints á des conditions de recouvrement par rapport un échantillon antérieur et á des conditions d'équilibrage par rapport á des variables courantes." Proceedings of the 10th Journée de Méthodologie Statistique de l'INSEE, January 24–26, 2012. [41 pp.]. Paris.

Demnati, A. and J.N.K. Rao. 2004. "Linearization Variance Estimators for Survey Data." *Survey Methodology* 30: 17–26.

Deville, J.C. 1999. "Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques." *Survey Methodology* 25: 193–203.

Di Meglio, E., G. Osier, T. Goedemé, Y. G. Berger, and E. Di Falco. 2013. "Standard Error Estimation in EU-SILC – First Results of the Net-SILC2 Project." In Proceedings of the Conference on New Techniques and Technologies for Statistics, [March 5–7,

2013]. [10 pp.]. Brussels. Available at: http://www.crosportal.eu/sites/default/files//NTTS2013%20Proceedings_0.pdf (last accessed April 30, 2015).

Eurostat. 2003. "'Laeken' Indicators-Detailed Calculation Methodology." Available at: http://www.cso.ie/en/media/csoie/eusilc/documents/Laeken%20Indicators%20-%20calculation%20algorithm.pdf (accessed February 4, 2014).

Eurostat. 2012. "European Union Statistics on Income and Living Conditions (EU-SILC)." Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eusilc (accessed January 7, 2013).

Feiveson, A.H. 2002. "Power by simulation." *The STATA Journal* 2: 107–124.

Gambino, J.G. and P.L.N. Silva. 2009. "Sampling and Estimation in Household Surveys." In *Handbook of Statistics, 29A: Design, Method and Applications*, edited by D. Pfeffermann and C.R. Rao, 407–439. Amsterdam: Elsevier.

Goedemé, T. 2010. "The Standard Error of Estimates Based on EU-SILC. An Exploration through the Europe 2020 Poverty Indicators." Working paper 10/09, [Herman Deleeck Centre for Social Policy, University of Antwerp, Belgium]. Available at: http://www.centrumvoorsociaalbeleid.be/index.php?q=node/2204/en (accessed April 30, 2015).

Graf, E. 2013. "Variance Estimation by Linearization for Indicators of Poverty and Social Exclusion in a Person and Household Survey Context." Paper presented at New Techniques and Technologies for Statistics, Brussels. Available at: http://www.cros-portal.eu/content/14a01ericgraf (last accessed February 5, 2014).

Graf, E. and Y. Tillé. 2014. "Variance Estimation Using Linearization for Poverty and Social Exclusion Indicators." *Survey Methodology* 40: 61–79.

Hansen, M.H. and W.N. Hurwitz. 1943. "On the Theory of Sampling from Finite Populations." *The Annals of Mathematical Statistics* 14: 333–362.

Holmes, D.J. and C.J. Skinner. 2000. "Variance Estimation for Labour Force Survey Estimates of Level and Change." The Office for National Statistics, London, United Kingdom. *Government Statistical Service Methodology Series* 21, 40 pp.

Horvitz, D.G. and D.J. Thompson. 1952. "A Generalization of Sampling Without Replacement From a Finite Universe." *Journal of the American Statistical Association* 47: 663–685. DOI: http://dx.doi.org/10.1080/01621459.1952.10483446.

Kalton, G. 2009. "Design for Surveys over Time." In *Handbook of Statistics, 29A: Design, Method and Applications*, edited by D. Pfeffermann and C.R Rao, 89–108. Amsterdam: Elsevier.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley and Sons.

Laniel, N. 1987. "Variances for a Rotating Sample from a Changing Population." In Proceedings of the Survey Research Methods Section, American Statistical Association, [August 17–20, 1987]. 496–500. Alexandria, VA: American Statistical Association.

McDonald, J.B. 1984. "Some Generalized Functions for the Size Distribution of Income." *Econometrica* 52: 647–664.

Muennich, R. and S. Zins. 2011. "Variance Estimation for Indicators of Poverty and Social Exclusion." Work package of the European project on Advanced Methodology for European Laeken Indicators (AMELI). Available at: http://www.uni-trier.de/index.php?id=24676 (accessed January 4, 2013).

Nordberg, L. 2000. "On Variance Estimation for Measures of Change when Samples Are Coordinated by the Use of Permanent Random Numbers." *Journal of Official Statistics* 16: 363–378.

Osier, G. 2009. "Variance Estimation for Complex Indicators of Poverty and Inequality Using Linearization Techniques." *Survey Research Methods* 3: 167–195.

Osier, G., Y.G. Berger, and T. Goedemé. 2013. "Standard Error Estimation for the EU-SILC Indicators of Poverty and Social Exclusion." Eurostat Methodologies and Working Papers series. Publications Office of the European Union, Luxembourg. Available at: http://ec.europa.eu/eurostat/documents/3888793/5855973/KS-RA-13-024-EN.PDF (accessed April 30, 2015).

Preston, I. 1995. "Sampling Distributions of Relative Poverty Statistics." *Applied Statistics* 44: 91–99.

Qualité, L. and Y. Tillé. 2008. "Variance Estimation of Changes in Repeated Surveys and its Application to the Swiss Survey of Value Added." *Survey Methodology* 34: 173–181.

Salem, A.B.Z. and T.D. Mount. 1974. "A Convenient Descriptive Model of Income Distribution: The Gamma Density." *Econometrica* 42: 1115–1127.

Schmeiser, B.W. and R. Lal. 1982. "Bivariate Gamma Random Vectors." *Operations Research* 30: 355–374. DOI: http://dx.doi.org/10.1287/opre.30.2.355.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Tam, S.M. 1984. "On Covariances from Overlapping Samples." *American Statistician* 38: 288–289. DOI: http://dx.doi.org/10.1080/00031305.1984.10483227.

Verma, V. and G. Betti. 2005. "Sampling Errors and Design Effects for Poverty Measures and Other Complex Statistics." Working Paper 53, Siena: Dipartimento di Metodi Quantitativi, Università degli Studi.

Verma, V. and G. Betti. 2011. "Taylor Linearisation Sampling Errors and Design Effects for Poverty Measures and Other Complex Statistics." *Journal of Applied Statistics* 38: 1549–1576. DOI: http://dx.doi.org/10.1080/02664763.2010.515674.

Wood, J. 2008. "On the Covariance Between Related Horvitz-Thompson Estimators." *Journal of Official Statistics* 24: 53–78.

Received July 2013
Revised February 2014
Accepted April 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 177–203, http://dx.doi.org/10.1515/JOS-2015-0013

ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys

Diego Zardetto[1]

ReGenesees is a new software system for design-based and model-assisted analysis of complex sample surveys, based on R. As compared to traditional estimation platforms, it ensures easier and safer usage and achieves a dramatic reduction in user workload for both the calibration and the variance estimation tasks. Indeed, *ReGenesees* allows the specification of calibration models in a symbolic way, using R model formulae. Driven by this symbolic metadata, the system automatically and transparently generates the right values and formats for the auxiliary variables at the sample level, and assists the user in defining and calculating the corresponding population totals. Moreover, *ReGenesees* can handle arbitrary complex estimators, provided they can be expressed as differentiable functions of Horvitz-Thompson or calibration estimators of totals. Complex estimators can be defined in a completely free fashion: the user only needs to provide the system with the symbolic expression of the estimator as a mathematical function. *ReGenesees* is in fact able to automatically linearize such complex estimators, so that the estimation of their variance comes at no cost at all to the user. Remarkably, all the innovative features sketched above leverage a particular strong point of the R programming language, namely its ability to process symbolic information.

Key words: Complex estimators; variance estimation; automated linearization; symbolic computation.

1. What is ReGenesees?

ReGenesees (R Evolved Generalized Software for Sampling Estimates and Errors in Surveys) is a full-fledged R software for design-based and model-assisted analysis of complex sample surveys. This system is the outcome of a long-term research and development project, aimed at defining a new standard for calibration, estimation and sampling error assessment to be adopted in all large-scale sample surveys routinely carried out by Istat (the Italian National Institute of Statistics).

The first public release of *ReGenesees* for general availability dates back to December 2011. The system is distributed as open source software under the European Union Public License (EUPL). It can be freely downloaded from JOINUP (the collaborative platform for interoperability and open source software of the European Commission) and from the Istat website.

Until the advent of *ReGenesees*, the estimation phase for sample surveys was handled at Istat by a SAS application named GENESEES (Falorsi and Falorsi 1997). The name of the

[1]Istat – Italian National Institute of Statistics, Via Cesare Balbo, 16 Rome, Rome 00184, Italy. Email: zardetto@istat.it

new R-based system has been deliberately chosen to emphasize Istat's seamless offer of software tools dedicated to that phase, while at the same time highlighting its *evolution* and *enhancement* through R. It is worth stressing, in any case, that the *ReGenesees* system is not the outcome of the simple migration to R of its SAS predecessor, but rather the fruit of a new, challenging and completely independent project.

The principal aim of this article is to introduce *ReGenesees* to the official statistics community. We will not try to provide a complete description of the statistical methods offered by *ReGenesees*, nor to describe its software implementation details: condensing this in an article would be beyond our ability. Instead, we will focus on few qualifying aspects which we perceive as *ReGenesees* "power features" and which, in our opinion, distinguish the system from other existing estimation platforms developed by National Statistical Institutes (NSIs). A broad overview of these qualifying aspects will be given in Section 3, after a brief discussion of the motivations of the project in Section 2.

Since many innovative features of *ReGenesees* can be traced back to a particular strong point of the R programming language, namely its ability to process *symbolic information*, the latter will form the leitmotif of the whole presentation.

R (R Core Team 2014) adheres to the functional programming paradigm and its semantics reveal some notable affinities with LISP (the ancestor of all functional languages). One relevant similarity is precisely the ability to manipulate symbolic expressions, a feature the R community usually refers to as "computing on the language". Perhaps the most popular materialization of this ability is the `formula` class with its ubiquitous usage in R statistical modelling functions (Chambers and Hastie 1992). Section 4 will be devoted to illustrate how *ReGenesees* exploits these potentialities in the calibration context.

A strictly related, though maybe less well known, fact is that R provides functionalities (e.g., symbolic differentiation and polynomial algebra) which are usually thought as hallmarks of Computer Algebra Systems (i.e., specialized software platforms where computation is performed on symbols representing mathematical objects rather than their numeric values). In Section 5 we will show how R facilities for computing symbolic partial derivatives of functions of several variables have been used in *ReGenesees* to fully automate the variance estimation of complex estimators.

Readers interested in assessing to what extent *ReGenesees* could cover the typical needs arising in NSIs during the estimation phase will find a list of the statistical methods made available by the system in Section 6. Section 7 will provide a quick overview of *ReGenesees* software design. There, we will acknowledge *ReGenesees*' indebtedness to the R *survey* package (Lumley 2004, 2010), but will also discuss those specific features which, in our opinion, make *ReGenesees* more fit than *survey* for the large-scale elaborations of the official statistics industry. The progress made so far in migrating Istat production processes to the new system will be reported in Section 8. Finally, Section 9 will address ongoing work and possible future extensions of the *ReGenesees* project.

2. Motivation of the ReGenesees Project

The tasks of calibrating survey weights, computing survey estimates, and assessing their precision, constitute a fundamental building block of the production process of official statistics. These are very complex tasks, whose correct execution requires a good

knowledge of the underlying statistical theory, full awareness of the adopted sampling plans, and often also some insight into the phenomena under investigation. Such skills, which rightfully contribute to define the ideal cultural background of a "good official statistician", cannot at present be entirely superseded by a software (nor probably will in the future), no matter how sophisticated and powerful it may be. However, most NSIs agree that the availability of highly evolved software systems (along with the definition of standard protocols for their optimal usage) is essential to ensure the *accuracy*, the *safety* and the *full reproducibility* of statistical production processes.

In the past, this strategic vision drove Istat and many other NSIs to invest in developing in-house software dedicated to the estimation phase: one may think, for example, of GES of Statistics Canada (Estevao et al. 1995), CLAN of Statistics Sweden (Andersson and Nordberg 1994), CALMAR and POULPE of French NSI INSEE (Sautory 1993; Caron 1998), BASCULA of Statistics Netherlands (Nieuwenbroek et al. 2000), g-CALIB of Statistics Belgium (Vanderhoeft 2001) and GENESEES of Istat (Falorsi and Falorsi 1997).

Today, the same strategic vision (even reinforced by the awareness of the ongoing rapid technological change, with its challenges and opportunities) pushes the same NSIs to renew, enrich or even redesign their software systems from scratch.

Some NSIs are continuously extending their traditional estimation platforms, trying to accommodate additional statistical capabilities as new needs arise in production and the methodology matures. Evidently, this approach is aimed at preserving the overall "look and feel" of the original system as much as possible (in terms of requirements, application logic and user experience), so that no abrupt and costly transitions can affect the production process. Relevant examples are Statistics Canada's project StatMx (Statistical Macro Extensions), which extends GES (Mohl 2007), as well as Statistics Sweden's new estimation tool ETOS (Estimation of Totals and Order Statistics), which enhances CLAN (Andersson 2009).

Istat decided to follow a different path, namely to redesign its estimation platform from scratch and to implement it in a different programming language. While the attempt to soften Istat dependence on proprietary technologies was undeniably a major driving factor, it was not the only relevant one. We were also interested in: (i) pushing our system towards *automation* (to reduce user workload and errors), (ii) improving its *modularity* (so that it could become easier to maintain and evolve), and (iii) providing a *wider choice* of statistical methods (beyond those formerly available). *ReGenesees* is the fruit of the efforts made by Istat in this direction.

3. ReGenesees: a Paradigm Shift

In the design phase of the *ReGenesees* project, it emerged fairly soon that the expectations described in Section 2 could only be met through a radical paradigm shift. As a consequence, *ReGenesees* has turned out to be rather different from its SAS predecessor GENESEES (and, incidentally, from most of other existing estimation tools) from the standpoint of both application logic and user experience. Indeed, besides allowing computing estimates and sampling errors for a much wider range of estimators, *ReGenesees* ensures easier and safer usage and a dramatic reduction in user workload. In a nutshell (see Sections 4, 4.1 and 5, 5.1 for more on points (1) and (2) below):

(1) User interaction with the new system takes place at a *very high level of abstraction*. *ReGenesees* users in fact no longer need to preprocess the survey data relying on ad hoc programs; instead, they only have to feed the software with (i) the data as they are, plus (ii) *symbolic metadata* that describe the adopted sampling design and calibration model (by "calibration model", we mean the assisting linear model underlying a specific calibration problem). At that point, it is up to the system itself to transform, in an automatic and transparent way, the survey data into the complex data structures required to solve the calibration problem and to compute estimates and errors.

(2) Besides totals and absolute frequency distributions (estimators that were already covered by GENESEES), *ReGenesees* is able to compute estimates and sampling errors with respect to means, ratios, multiple regression coefficients, quantiles, and, more generally, with respect to any *complex estimator*, provided it can be expressed as a differentiable function of Horvitz-Thompson or calibration estimators. It is worth stressing that such complex estimators can be defined in a completely free fashion: the user only needs to provide the system with the *symbolic expression* of the estimator as a mathematical function. *ReGenesees* is in fact able to automatically linearize such complex estimators, so that the estimation of their variance comes at no cost at all to the user.

Existing estimation software (the Istat traditional SAS system being no exception) generally gave little support to users in preparing auxiliary variables and population totals for calibration, or in deriving the Taylor expansion of nonlinear estimators and in computing the corresponding linearized variable for variance estimation purposes. As a consequence, ad hoc (often very complex) programs for data preparation, transformation and validity checking were developed and maintained outside the scope of the estimation system: a time-consuming and error-prone practice. *ReGenesees* frees its users from such needs, with an evident gain in terms of workload reduction, better usability and increased robustness against possible errors. Furthermore, letting *ReGenesees* carry out tasks that traditional platforms delegated to a (skilled) human makes the statistical production workflow fully reproducible, as the system persistently logs all the elaboration steps it executes.

Interestingly, both the innovative *ReGenesees* features sketched above leverage R's ability to process *symbolic information*. As a matter of fact, developing the same functionalities in SAS would have been prohibitive: a striking example of what we meant, in Section 2, when referring to "opportunities of technological change".

A technological shift, on the other hand, always involves challenges and some price to be paid. The most threatening challenge faced by the *ReGenesees* project has been to demonstrate that an R-based system would actually be able to manage efficiently the huge amounts of data involved in processing Istat large-scale surveys (see Section 7 for an illustrative example).

A lot of effort has been invested during the whole development cycle of the new system to meet this challenge. Thanks to the empirical evidence and to the reproducible results accumulated during an extensive and thorough testing campaign, we are certain that the challenge has been definitely overcome. Indeed, since its beta release, *ReGenesees* has been

successfully tested on both the Labour Force Survey (LFS) and the Small and Medium Enterprises Survey (SME): where the tasks of calibration and computation of estimates and errors are concerned, these two surveys constitute (each one in its own domain) the most severe test bed available at Istat. Furthermore, today about 20 Istat large-scale surveys have successfully integrated *ReGenesees* into their production workflow.

ReGenesees also underwent an independent validation, which was carried out by colleagues from the UK statistical institute (ONS). A first comparative study, performed on their Life Opportunities Survey, measured *ReGenesees* effectiveness and efficiency using Statistics Canada's GES as a benchmark. The outcome was that *ReGenesees* replicated the results achieved by GES exactly, while ensuring a significant increase in efficiency (in their testing environment, execution times turned out to be halved on average). This result, in turn, triggered a second ONS initiative. *ReGenesees* was used to calibrate three important surveys for the Scottish Government, whose weighting procedures had till then been contracted to three separate external companies: (i) Scottish Household Survey, (ii) Scottish Health Survey, and (iii) Scottish Crime and Justice Survey. Again the results were very satisfactory, and the ONS Methodology Advisory Service suggested *ReGenesees* as a possible "calibration engine" to be adopted in the novel *centralized weighting* framework designed for the Scottish Government (Davidson 2013). Eventually, *ReGenesees* was indeed used in production for the last round of all the aforementioned surveys (Scottish Government (2013a), (2013b), and (2014)).

4. Leveraging Symbolic Information: the Calibration Side

Real-world calibration tasks in the field of official statistics can simultaneously involve several hundreds of auxiliary variables (just to give an impression: each quarterly round of the Italian LFS entails calibrating to known population totals for over 4,000 auxiliary variables). Moreover, the construction of such auxiliary variables is in general highly non-trivial, as they need to be carefully derived from the original survey variables according to the (possibly very complex) adopted calibration models. With respect to such operations, traditional calibration facilities (as those listed in Section 2, mostly based on SAS) gave limited practical support to users, instead devoting dozens of user manuals' pages to describing the standard data structures they expected as input. As a consequence, users had to develop customized programs (typically SAS scripts) in order to generate the right input data to feed the calibration system, with "right" here meaning: (i) appropriate to the survey data and to the calibration task at hand, and (ii) compliant with all the documented rules imposed by the system.

Conversely, users interact with the *ReGenesees* system at very high level of abstraction, as they only need to specify the calibration model in a symbolic way, via R-model formulae. Model formulae are R objects of class `formula` (Chambers and Hastie 1992). Thanks to the flexible syntax of this class and to the powerful semantics of its methods, R-model formulae can be used to compactly specify a wide range of statistical models (far beyond the ANOVA context from which the inspiring notation of Wilkinson and Rogers (1973) originally came). In particular, R-model formulae have the expressive power to represent arbitrarily complex calibration models, including sophisticated modelling aspects such as, for example, interactions between numeric and categorical auxiliary variables, multi-way interactions,

factor crossing, nesting and conditioning, collinearity prevention and term aliasing, handling customized contrasts, referencing auxiliary variables defined on the fly, and so on.

Driven by a calibration-model formula, *ReGenesees* is able to transparently generate the right values and formats for the auxiliary variables at the sample level. In addition, the system assists in defining and calculating the population totals corresponding to the generated auxiliary variables. Indeed, by leveraging again the calibration model formula, *ReGenesees* provides the user with a *template* dataset appropriate to store the requested totals. Whenever the actual population totals are available to the user as such, that is, in the form of already computed aggregated figures, the user has only to fill in the template. This case typically occurs in Italy for household surveys (whose sampling design is two-stage: municipalities + households) because demographic balance figures are updated and released on a monthly basis, whereas a centralized population register is not yet available.

An even bigger benefit is achieved when the sampling frame of the survey is available as a single database table and the actual population totals can be calculated from this source. In such cases, *ReGenesees* is able to automatically compute the totals of the auxiliary variables from the sampling frame, and to safely arrange and format these values so that they can be directly used for calibration. This scenario applies to all the structural business surveys carried out in Italy, whose samples are drawn (typically with a one-stage design) from ASIA, the Istat comprehensive archive of about 4.5 million Italian active enterprises.

In the next section, the abstract considerations sketched above will be illustrated in practice by two calibration examples.

4.1. Two Calibration Examples: Global and Partitioned Calibration

Here we provide two simple examples illustrating how a *ReGenesees* user can tackle a calibration task. Both examples will be discussed in the light of the considerations set out in Section 4.

Both examples will address the same calibration problem, which will be solved *globally* in the first case, and in a *partitioned* way in the second case. For the sake of clarity, each example will be decomposed into atomic elaboration steps. For each elaboration step, first, we will provide a quick natural-language description of what is going on; then, we will show the corresponding R-code statements. Such code fragments will be reported as if they were typed by a user interacting with *ReGenesees* through the ordinary R command-line interface; the same elaborations could also be obtained by exploiting the graphical user interface of the system (see Section 7).

Both examples will involve two artificial datasets mimicking structural business statistics data, which ship with the *ReGenesees* system. The first dataset, `sbs`, represents a sample of enterprises, while the second, `sbs.frame`, is the sampling frame from which the sample has been selected.

We will calibrate `sbs` weights with calibration constraints imposed simultaneously on the total number of employees (`emp.num`) and enterprises (`ent`) inside domains obtained by:

i) crossing two-digit classification of economic activity or industry (`nace2`) and `region` (a convenience, threefold territorial division);
ii) crossing employment size classes (`emp.cl`), economic activity macro-sector (`nace.macro`) and `region`.

Let us now proceed step by step with the first example.

S1. Bind survey data (`sbs`) to sampling design metadata;

```
> sbsdes <- e.svydesign(data=sbs, ids=~id, strata=~strata, weights=~weight, fpc=~fpc)
> sbsdes
Stratified Independent Unit Sampling Design
- [664] strata
- [6909] units

Call:
e.svydesign(data = sbs, ids = ~id, strata = ~strata, weights = ~weight,
    fpc = ~fpc)
```

S2. Specify the calibration model symbolically (`calmodel`) and build a template dataframe to store the corresponding known population totals;

```
> pop <- pop.template(data=sbsdes,
+                     calmodel=~((emp.num+ent):(nace2+emp.cl:nace.macro)):region-1)
> dim(pop)
[1]   1 462
```

S3. Automatically compute the requested totals from the sampling frame (`sbs.frame`) and safely fill the template;

```
> pop <- fill.template(universe=sbs.frame, template=pop)
```

S4. Pass the known totals to the system and perform the calibration task;

```
> sbscal <- e.calibrate(design=sbsdes, df.population=pop, calfun="linear",
+                       bounds=c(0.01,3))
> sbscal
Calibrated, Stratified Independent Unit Sampling Design
- [664] strata
- [6909] units

Call:
e.calibrate(design = sbsdes, df.population = pop, calfun = "linear",
    bounds = c(0.01, 3))
```

Code fragment **S1** simply tells the *ReGenesees* system that the `sbs` sample was selected with a stratified one-stage unit sampling design without replacement. Since the present focus is on calibration, we cannot go into any detail on function `e.svydesign`. Rather, we only point out that it can handle a wide range of complex sampling designs (see Section 6 for a list).

The `calmodel` formula in code fragment **S2** specifies the assisting linear model underlying our complex calibration problem. Without going into syntax details: '$\sim$' declares a model formula, '**:**' means interaction, '**+**' means sum of effects (not arithmetic addition), '**-1**' means no intercept term needed (otherwise, it would be tacitly implied in R). More specifically, `calmodel` identifies the calibration constraints, as well as the related auxiliary variables. Given the nature of the `sbs` dataset, such a calibration model – which involves only six *original* survey variables, two numeric and four categorical – translates into 462 different numeric auxiliary variables.

Code fragment **S2** generates a template dataframe (`pop`) to store properly the known totals of these 462 variables (in fact `pop` has one row and 462 columns, as shown). It is a *template* dataframe in the sense that all the known totals it must be able to store are still

missing, but it has the right certified structure (in terms of dimension, column names, variable types, . . .) to be processed successfully by the calibration facility of the *ReGenesees* system once filled. Note that the `pop.template` function works in a *declarative* way: it avoids any need for the user to understand, comply with, or even be aware of, the structure of the template that is being built.

Function `fill.template`, in code fragment **S3**, transparently computes the requested population totals from the sampling frame, and arranges and formats such values according to the template structure. Note that these elaborations are again driven by the calibration model formula `calmodel`, which is attached as an *invisible* attribute to the template dataframe `pop` returned by the previous code fragment **S2**.

Function `e.calibrate` in code fragment **S4** first automatically generates the model matrix storing the values of the 462 numeric auxiliary variables for the whole sample, then computes the desired calibrated weights.

In conclusion, fragments **S1 – S4** show that we were able to perform a complex calibration task by passing to *ReGenesees* only the data as they were plus symbolic metadata, without any need to work out the 462 numeric auxiliary variables and their population totals.

Let us come back to the `calmodel` formula in code fragment **S2**. This formula identifies a *factorizable* calibration model, as variable `region` acts as a *common factor* interacting with *all* the other variables appearing in the formula. Factor variables with this property split the sample (and the target population) into nonoverlapping subsets known as *"calibration domains"* (*"model groups"* in the terminology of Estevao et al. 1995). The interest in factorizable calibration models lies in the fact that the *global* calibration problem they describe can actually be broken down into smaller *local* subproblems, one per calibration domain, which can be solved separately. This opportunity can, in many cases, result in a dramatic reduction in computational complexity. Obviously the computational efficiency gain increases with the size of the survey and (most importantly) with the number of auxiliary variables involved.

Now, code fragments **S2 – S4** above show how the *ReGenesees* calibration facility `e.calibrate` can be used to solve a factorizable calibration problem with a *global* approach. In the next example, we will instead show how to solve the same problem with a *partitioned* approach. Again no data preparation effort is required of the user.

We are now ready to go on with the second example.

S5. Specify symbolically the reduced calibration model (`calmodel`) and the calibration domains (`partition`), and build the appropriate known totals template;

```
> pop2 <- pop.template(data=sbsdes,
+                      calmodel=~((emp.num+ent):(nace2+emp.cl:nace.macro))-1,
+                      partition=~region)

> dim(pop2)

[1]   3 155
```

S6. Automatically compute the requested totals from the sampling frame and safely fill the template;

```
> pop2 <- fill.template(universe=sbs.frame, template=pop2)
```

S7. Pass the known totals to the system and perform the partitioned calibration task;

```
> sbscal2 <- e.calibrate(design=sbsdes, df.population=pop2, calfun="linear",
+                        bounds=c(0.01,3))

> sbscal2

Calibrated, Stratified Independent Unit Sampling Design
- [664] strata
- [6909] units

Call:
e.calibrate(design = sbsdes, df.population = pop2, calfun = "linear",
    bounds = c(0.01, 3))
```

S8. Check that the calibrated weights obtained through the partitioned algorithm are indeed identical to those obtained previously through the global approach;

```
> all.equal(weights(sbscal), weights(sbscal2))

[1] TRUE
```

Code fragment **S5** shows how we can tell *ReGenesees* to solve our factorizable calibration problem in a *partitioned* way. We only need to: (i) pass to the `calmodel` argument the *reduced* model obtained by factoring the common factor variable `region` out from the original model; (ii) pass the same variable to the `partition` argument in order to identify the calibration domains. Note that this implies a different structure of the known totals template `pop2` as compared to `pop` in **S2**. Specifically, we get as many rows as calibration domains (three in our example, since we have three regions) and a new column to identify such domains. Hence, the number of cells to be filled with actual population totals remains the same, that is, $462 = 3 \times (155 - 1)$, as it must be. Again the user does not need to take care of such format details, as they are automatically handled by function `fill.template` in code fragment **S6**.

Function `e.calibrate` in code fragment **S7** sequentially runs the three calibration subproblems corresponding to the calibration domains identified by `partition`. Lastly, code fragment **S8** shows that the calibrated weights achieved by the partitioned algorithm are indeed equal to those obtained previously through the global approach: we reported this result for illustration only, as it is guaranteed by construction.

We conclude by pointing out two technical aspects of the partitioned calibration task seen in code fragments **S5 – S7**. First, the computational efficiency gain of the partitioned approach is evident even for our toy example: execution time indeed turns out to be reduced by a factor of 10 with respect to the global calibration alternative. Second, solving the calibration problem in a partitioned way has some nontrivial consequences in the variance estimation phase, as discussed in Estevao et al. (1995). Nevertheless, as we will show in Subsection 5.1, the *ReGenesees* system will automatically take care of all the involved technical issues (mainly arising from the interplay between estimation domains and calibration domains), again without requiring any specific effort of the users.

5. Leveraging Symbolic Information: the Linearization Variance Side

The Taylor linearization method is a well-established, approximate tool (Woodruff 1971; Wolter 2007) for estimating the variance of complex estimators, namely estimators that can be expressed as nonlinear (but "smooth", say of class C^2 at least) functions of Horvitz-Thompson (HT) estimators of totals:

$$\hat{\theta} = f(\hat{Y}_1, ..., \hat{Y}_m) \tag{1}$$

where $\hat{Y}_j = \sum_{k \in s} d_k y_{jk}$ and the design weights d_k are reciprocals of first-order inclusion probabilities, $d_k = \pi_k^{-1}$.

The key assumption of the method, generally justified by large-sample arguments (see e.g., Krewski and Rao 1981), is that, as far as variance estimation is concerned, a complex estimator can be approximated by its first-order Taylor series expansion:

$$\hat{\theta} \approx \theta + \sum_{j=1}^{m} \left. \frac{\partial f}{\partial \hat{Y}_j} \right|_{\mathbf{Y}} (\hat{Y}_j - Y_j) \doteq \hat{\theta}_{lin} \tag{2}$$

The linear approximation of the original complex estimator (1) is then expressed (up to constant, though unknown, terms) as the HT total of a single artificial variable $\hat{z}$:

$$\hat{\theta}_{lin} \approx \sum_{k \in s} d_k \hat{z}_k + const \tag{3}$$

Variable $\hat{z}$ in Equation (3) is the so-called *linearized variable* (Woodruff 1971) of the complex estimator (1):

$$\hat{z}_k = \sum_{j=1}^{m} \left. \frac{\partial f}{\partial \hat{Y}_j} \right|_{\hat{\mathbf{Y}}} y_{jk} \tag{4}$$

The approximate identity symbol $\approx$ in Equation (3) and the hat over z in Equations (3)-(4) rest on having evaluated the gradient of function f at estimated totals $\hat{\mathbf{Y}}$ rather than at the corresponding true (but unknown) values $\mathbf{Y}$. From Equation (3) it follows that an estimator of the (approximate) variance of a complex estimator can be built for all sampling designs for which a good estimator of the variance of an HT total is known:

$$\hat{V}(\hat{\theta}) \approx \hat{V}(\sum_{k \in s} d_k \hat{z}_k) \tag{5}$$

Equation (5) summarizes the "golden rule" of the method: estimating the variance of a complex estimator boils down to the much simpler problem of estimating the variance of the HT total of its linearized variable $\hat{z}$, under the sampling design at hand.

The extension to "smooth" functions of calibration estimators (see Särndal 2007 and references therein) of totals is straightforward. Let us indicate an estimator of this kind as follows:

$$\hat{\theta} = f\left(\hat{Y}_1^{CAL}, \ldots, \hat{Y}_m^{CAL}\right) \tag{6}$$

where $\hat{Y}_j^{CAL} = \sum_{k \in s} w_k y_{jk}$ and the calibrated weights w_k are obtained by minimizing an appropriate distance function $G(\mathbf{w}, \mathbf{d})$ from the design weights d_k, subject to calibration constraints involving a set of auxiliary variables $\mathbf{x}$ and the corresponding known population totals: $\sum_{k \in s} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$.

The golden rule (5) still applies to estimators such as (6), the only relevant change being a different expression for the linearized variable (Deville 1999):

$$\hat{z}_k = \sum_{j=1}^{m} \left. \frac{\partial f}{\partial \hat{Y}_j^{CAL}} \right|_{\hat{\mathbf{Y}}^{CAL}} g_k \hat{e}_{jk} \tag{7}$$

namely the value of the original variable y_{jk} has been replaced by the product of the g-weight, $g_k = w_k/d_k$, with the estimated *residual* of that variable under the adopted calibration model:

$$\hat{e}_{jk} = y_{jk} - \boldsymbol{x}'_k \cdot \hat{\boldsymbol{\beta}}_j \tag{8}$$

The g-weighted residuals in (7) result from linearizing the GREG estimator (Särndal et al. 1989), and from Deville and Särndal's (1992) well-known finding that, under mild conditions on the involved distance functions, all calibration estimators are actually asymptotically equivalent to the GREG estimator.

Besides asymptotic theory, there is solid empirical evidence that the Taylor linearization method can yield reliable variance estimates, as long as: (i) the functional form of the estimator is well behaved, (ii) the sample is large and the sampling design is not awkward, and (iii) the variables involved in the estimator are not highly skewed at the population level. These conditions are frequently met in official statistics production, but there are notable exceptions, such as the skewness characterizing many business statistics variables. An investigation of the empirical properties of the variance estimators would therefore be an interesting topic for future study.

While the mathematical framework outlined by Equations (1)-(8) is clear, its software implementation involves some subtle and tricky technical points. For instance, domain estimation of standard errors tends to become cumbersome, especially for complicated functions of calibration estimators. Moreover, as anticipated in Subsection 4.1, when a *partitioned* calibration is performed, the interplay between estimation domains and calibration domains has to be carefully taken into account for variance estimation (see the "general case" discussed in sec. 5 of Estevao et al. 1995). Note that this issue has a relevant impact on software implementation, since the *partitioned* calibration approach is computationally far more efficient than the *global* one, if not even the only feasible alternative for some large-scale surveys (see Section 7 for an illustrative example).

From a software development standpoint, the linearization approach to variance estimation has a fundamental drawback: the Taylor series expansion of a nonlinear estimator *does* depend on its functional form *f*. Therefore, using traditional computing environments (e.g., SAS) that are *unable* to perform *symbolic differentiation*, different programs have to be developed separately for each nonlinear statistic *f*. As a direct consequence, traditional systems like those listed in Section 2 suffer from two main limitations.

First, they support only a rather limited set of nonlinear estimators, typically ratios of totals (EUROSTAT 2002, 2013). As a somewhat extreme example, Istat traditional system GENESEES cannot automatically handle any nonlinear estimator. At the other extreme, Statistics Sweden's system CLAN, despite being unquestionably more general and versatile, still can only handle *rational* functions of totals.

The second limitation is that traditional platforms generally cannot allow their users to define *their own* complex estimators, namely statistics which are not built in. Whenever users of such systems need non-built-in estimators, they have to develop ad hoc programs to compute the appropriate linearized variables on their own.

Even CLAN users must actually write SAS (%FUNCTION) macros to let the system understand the functional form of the rational function they are interested in. Moreover, those macros become more and more complicated as the complexity of the estimator grows (see the examples reported in the Appendix), with the risk of impairing the usability of the system (Davies and Smith 1999; Ollila et al. 2004). In fact, CLAN users have to successively decompose their original function into simpler subfunctions until no further simplification is possible, using the four elementary algebraic operations (i.e., *addition*, *subtraction*, *multiplication* and *division*) provided by the system as preprogrammed macros (i.e., %ADD, %SUB, %MULT and %DIV). The purpose of this coding burden is to enable CLAN to compute the linearized variable of a complex rational function of totals in a stepwise fashion, that is, by successively applying appropriate Woodruff transformations corresponding to elementary *binary* functions $+$, $-$, $\times$, and /, which are the only ones the system can *directly* cope with.

ReGenesees overcomes both the limitations mentioned above, again leveraging R's ability to process symbolic information. We achieved this goal through the following steps. First, we devised a simple syntax for specifying arbitrary complex estimators through their functional form, and enabled it by exploiting R methods for manipulating `expression` objects (see the next section for further details). Then, we used advanced R facilities for calculating *symbolic derivatives* to develop a sort of "universal" linearization program. Once equipped with it, we were in the position to add to the system new nonlinear estimators almost for free (see Section 6 for a list). Lastly, we engineered our universal linearization program, making it friendly and fully visible to users. The resulting function, named `svystatL`, handles arbitrary user-defined complex estimators, as we show in the next section with two practical examples.

5.1. *Two Examples of Complex Estimators: Geometric Mean and Standard Deviation of a Variable*

As we have already sketched above, we equipped *ReGenesees* with a simple syntax for specifying arbitrary complex estimators through their functional form, via R `expression` objects. According to this syntax:

i) the estimator of the *total* of a variable is simply represented by the *name* of the variable itself;
ii) the convenience name *ones* identifies an *artificial* variable whose value is 1 for each elementary unit.

Evidently, the system variable `ones` can be used directly to estimate the size (in terms of elementary units) of the population, as well as to define the estimator of the mean of a given survey variable.

By combining rules (i) and (ii) above, and by making use of whatever algebraic operators and mathematical functions R understands, *ReGenesees* users can actually define any estimator they need.

Here are some elementary examples:

- $\hat{Y}$ *maps to* `expression(y)`
- $\hat{N}$ *maps to* `expression(ones)`
- $\hat{R} = \frac{\hat{Y}}{\hat{X}}$ *maps to* `expression(y/x)`
- $\hat{\mu}_y = \frac{\hat{Y}}{\hat{N}}$ *maps to* `expression(y/ones)`
- $\hat{B} = \frac{\hat{T}_{in} - \hat{T}_{out}}{\hat{T}_{in} + \hat{T}_{out}}$ *maps to* `expression((in − out)/(in + out))`

with $\hat{T}_{\text{var}}$ in the last item indicating the estimator of the total of variable *var*.

After this necessary preamble, let us now switch to our complex estimator examples. For each example, we will proceed step by step, illustrating each atomic elaboration step with the same style we adopted in Subsection 4.1.

Note that, since the estimators we are going to tackle cannot be expressed as rational functions of estimators of totals, they could *not* be handled directly (i.e., automatically) by any of the traditional estimation platforms listed in Section 2, not even by CLAN which is the most general and flexible among them.

Our first example will address the *geometric mean* of a (non-negative) survey variable *y*. To this end, recall that the geometric mean of *y* can be expressed as the exponential of the average of the logarithm of *y*, so that our complex estimator reads:

$$\hat{G}_y = e^{(\hat{T}_{log(y)}/\hat{N})} \tag{9}$$

We will work with the same sbs-like data we used in Subsection 4.1, and will select number of employees (`emp.num`) as study variable *y*.

S9. Add a new computed variable, i.e. the log in equation (9), to the original survey design object (`sbsdes`);

```
> sbsdes <- des.addvars(sbsdes, log.emp.num=log(emp.num))
```

S10. Estimate the geometric mean of number of employees (`emp.num`), its standard error and confidence interval;

```
> G <- svystatL(sbsdes, expression(exp(log.emp.num/ones)), conf.int=TRUE)
```

S11. Print on screen the obtained results;

```
> G
                        Complex       SE   CI.l(95%)   CI.u(95%)
exp(log.emp.num/ones)   20.5156   0.0608     20.3965     20.6347
```

The purpose of code fragment **S9** ought to be clear: if we included `log(emp.num)` *directly* inside the `expression` in **S10**, this would have been interpreted – according to the syntax introduced before – as the *logarithm of the estimator of the total* of `emp.num`, rather than as the *estimator of the total of the logarithm* of `emp.num`, which is what Equation (9) actually dictates.

As a whole, code fragments **S9 – S11** above show that *ReGenesees* was able to estimate the variance of a user-defined complex estimator in a completely automated manner, overcoming any need for developing ad hoc programs.

Our second example focuses on the estimator of the *Standard Deviation* of a survey variable y:

$$\hat{S}_y = \sqrt{\frac{\hat{N}}{\hat{N}-1}\left[\hat{\mu}_{y^2} - (\hat{\mu}_y)^2\right]} \tag{10}$$

We will keep working with the same survey data and we will select value added (`va`) as study variable. This time we will compute estimates and sampling errors on a *calibrated* design object, more specifically on the result of our *partitioned* calibration example of Subsection 4.1, `sbscal2`. Moreover, we will not estimate `va`'s standard deviation for the whole population, but rather for domains identified by employment size classes, `emp.cl`: note that each of these estimation domains intersects all the calibration domains of `sbscal2` (which were identified by variable `region`). The purpose of such choices is to show that no additional effort is actually required of a *ReGenesees* user for handling the additional technical complexities of this second example.

S12. Add a new computed variable, i.e. the square in equation (10), to the calibrated survey design object (`sbscal2`);

```
> sbscal2 <- des.addvars(sbscal2, va2=va^2)
```

S13. Estimate the standard deviation of value added (`va`), its standard error and confidence interval for employment size classes (`emp.cl`);

```
> S <- svystatL(sbscal2, expression(sqrt((ones/(ones-1))*((va2/ones) - (va/ones)^2))),
+               by=~emp.cl, conf.int=TRUE)
```

S14. Print on screen the obtained results;

```
> S
  emp.cl     Complex    SE.Complex   CI.l(95%).Complex   CI.u(95%).Complex
   [6,9]     4970.64        429.58             4128.69             5812.60
  (9,19]     4252.76        231.81             3798.42             4707.10
 (19,49]     7535.00        780.52             6005.21             9064.78
 (49,99]     8883.48        588.78             7729.49            10037.48
(99,Inf]    20022.11          0.00            20022.11            20022.11
```

Code fragment **S12** has exactly the same purpose as **S9**. Code fragments **S12 – S14** above demonstrate that *ReGenesees* was actually able to cope with the very complex analysis we set up for our second example in a completely automated way, again overcoming any need for developing ad hoc programs. More specifically, function `svystatL` in fragment **S13** handled all the following tasks rigorously (though transparently to the user): (i) compute symbolically the gradient of the function specifying the complex estimator, f, with respect to the estimators of totals f depends on (i.e., $\hat{Y}$, $\hat{T}_{y^2}$ and $\hat{N}$); (ii) compute the (calibrated) estimates of such totals for all the requested estimation domains; (iii) for each estimation domain, evaluate the gradient of f at the corresponding estimated totals; (iv) for each elementary unit belonging to a given estimation domain, compute the g-weighted residuals of the variables whose totals appear in f, taking into account the different estimated regression coefficients pertaining to the calibration domains which happen to be intersected by the given estimation domain;

(v) for each elementary unit in the sample, compute the linearized variable obtained by putting together all the aforementioned ingredients according to Formula (7); (vi) pass the linearized variable and the direct weights to the variance estimation algorithm appropriate for an HT total under the sampling design at hand; (vii) return the obtained results for the requested estimation domains.

6. ReGenesees Statistical Methods in a Nutshell

From a statistical point of view, the *ReGenesees* system is quite rich and flexible, as it is able to handle a wide range of sampling designs, calibration models and estimators. As anticipated in the article outline, we cannot provide a thorough description of the methods offered by *ReGenesees* here. Instead, we will report them concisely in a list (see Table 1 below) and limit ourselves to clarifying those expressions which might not be self-evident to readers who are not official statistics practitioners. Further details, along with a wealth

Table 1. A summary of ReGenesees statistical methods

- **Complex sampling designs**
  - Multistage, stratified, clustered, sampling designs
  - Sampling with equal or unequal probabilities, with or without replacement
  - "Mixed" sampling designs (i.e., with both self-representing and non-self-representing strata)
- **Calibration**
  - Global and partitioned (for factorizable calibration models)
  - Unit-level and cluster-level weights adjustment
  - Homoscedastic and heteroscedastic models
  - Linear, raking and logit distance functions
  - Bounded and unbounded weights adjustment
  - Multi-step calibrations
- **Basic estimators**
  - Horvitz-Thompson
  - Calibration estimators
- **Variance estimation**
  - Multistage formulation
  - Ultimate cluster approximation
  - Collapsed strata technique for handling lonely PSUs
  - Taylor linearization of nonlinear "smooth" estimators
  - Generalized variance functions method
- **Estimates and sampling errors (standard error, variance, coefficient of variation, confidence interval, design effect) for:**
  - Totals
  - Means
  - Absolute and relative frequency distributions (marginal, conditional and joint)
  - Ratios between totals
  - Multiple regression coefficients
  - Quantiles
- **Estimates and sampling errors for complex estimators**
  - Handles arbitrary differentiable functions of Horvitz-Thompson or calibration estimators
  - Complex estimators can be freely defined by the user
  - Automated Taylor linearization
  - Design covariance and correlation between complex estimators
- **Estimates and sampling errors for subpopulations (domains)**
  - All the analyses above can be carried out for arbitrary domains

of practical examples, can be found in the reference manual of the system (Zardetto 2014). Lastly, we will point out some resources for assessing how *ReGenesees* statistical methods compare to those offered by other existing systems cited in Section 2.

In multistage sampling, it is common jargon to call "self-representing" those primary sampling units (PSUs) which are selected with probability one. Thus, for instance, all PSUs belonging to a "take-all" stratum are self-representing. Sometimes efficiency considerations push survey designs to the extreme of selecting just a single PSU per stratum: this happens, for instance, in the Italian LFS. In these cases, population strata containing just a single self-representing PSU are referred to as self-representing strata (SR strata); all the other strata, containing many PSUs among which just a single one is randomly selected with probability less than one, are called non-self-representing (NSR strata).

The resulting sampling design is sometimes called "mixed", because the actual stages of selection differ for SR and NSR strata. For instance, the Italian LFS is actually a two-stage cluster sampling design inside NSR strata (namely a *pps* selection of municipalities followed by a *srs* of households) and a one-stage cluster sampling inside SR strata (namely a *srs* of households inside those municipalities which are always included in the sample). Note also that for the Italian LFS *all* the PSUs selected inside NSR strata are "lonely PSUs" *by design*.

Both the mixed nature of a sampling design and the occurrence of lonely PSUs (i.e., PSUs which are alone inside a NSR stratum at the sample level) are issues that need to be carefully taken into account in variance estimation, and this is ensured by *ReGenesees*. In particular, the system overcomes problems arising from lonely PSUs by adopting the collapsed strata technique, as proposed by Rust and Kalton (1987).

When clusters selected at subsequent sampling stages ($k \geq 2$) have *equal* inclusion probabilities (e.g., for a stratified two-stage design with *srs* of both PSUs and SSUs), *ReGenesees* correctly estimates the full multistage variance, without neglecting subleading contributions arising from stages after the first ($2, \ldots, k$). This exploits a recursive algorithm similar to the one proposed in Bellhouse (1985), inherited and extended from package *survey* (see Section 7 for details).

Conversely, for *unequal* probability sampling *without* replacement (e.g., the pps selection of PSUs in NSR strata of the Italian LFS), in order to get exact variance estimates second-order inclusion probabilities should be known, which is generally unfeasible. In such cases, *ReGenesees* resorts to so-called Ultimate Cluster approximation (Kalton 1979), which rests on pretending that PSUs were sampled *with* replacement, even if this is not actually the case. If PSUs were sampled with replacement, the *only* contribution to the variance would come from estimated PSU totals, in that one could simply ignore any available information about subsequent sampling stages – which explains the phrase *"ultimate clusters"*. This approximation is known to result in conservative variance estimates, with an upward bias which is negligible as long as the actual sampling fractions of PSUs are very small.

As a rule, *ReGenesees* computes variance estimates of nonlinear estimators with the Taylor approach summarized in Section 5. Quantiles – being *implicit*, *non-smooth* functions of totals – are a mandatory exception. Here *ReGenesees* switches to the method proposed in Woodruff (1952), whose main ingredients are the estimation and the local inversion of the cumulative distribution function of the interest variable, again using facilities from and extending the *survey* package.

The Generalized Variance Functions (GVF) method (see, e.g., sec. 7 of Wolter 2007) is the first (and so far only) *ReGenesees* incursion in the model-based realm. The GVF method (whose justification is largely empirical, with few exceptions) amounts to modelling the variance of an estimator as a function of its expected value, and using the fitted model to *predict* variance estimates, rather than computing them directly.

It is worth stressing that only a rather limited subset of the statistical methods covered by *ReGenesees* was already available inside its SAS predecessor GENESEES. For instance, the only estimators provided were totals and absolute frequencies, and variance estimation in multistage designs could be tackled only under the ultimate cluster approximation.

Readers interested in assessing how *ReGenesees* statistical methods compare to those offered by existing systems developed by NSIs are referred to the following resources. EUROSTAT (2002) provides a synoptic table summarizing the suitability of software tools for sampling designs and related issues on variance estimation (p. 34): this covers BASCULA, CLAN, GENESEES, GES, and POULPE. A similar table, this time also taking into account *ReGenesees* (but loosing GES), can be found in Appendix 7 of EUROSTAT (2013). Davies and Smith (1999) review and compare CLAN and GES in depth. Lastly, Ollila et al. (2004) offer the most comprehensive and thorough comparative study we are aware of, even though restricted to BASCULA, CLAN, and POULPE (only a brief overview is given of g-CALIB and GENESEES).

7. ReGenesees Software Design: a Quick Overview

System Architecture. The *ReGenesees* system has a clear-cut two-layer architecture. The application layer of the system is embedded into an R package named `ReGenesees` (Zardetto 2014). A second R package, called `ReGenesees.GUI` (Zardetto and Cianchetta 2014), implements the presentation layer of the system (namely a Tcl/Tk GUI). Both packages can be run in Windows as well as in Mac, Linux and most of the Unix-like operating systems. While the `ReGenesees.GUI` package requires the `ReGenesees` package, the latter can be used also without the GUI on top. This means that the statistical functions of the system will always be accessible to users interacting with R through the traditional command-line interface (as for code fragments in Subsections 4.1 and 5.1). Conversely, less experienced R users will benefit from the user-friendly mouse-click graphical interface.

Related R Projects. It is worth mentioning that, especially in terms of software design principles, the `ReGenesees` package owes a lot to the beautiful, rich and still growing `survey` package written by Thomas Lumley (Lumley 2004, 2010). Retrospectively, the original seeds of the *ReGenesees* project can be traced back to late 2006, when we were trying to optimize and extend `survey` in order to enable its critical functions to successfully process Istat's large-scale surveys. Fairly soon, this attempt required us to rethink the technical implementation of the package globally, that is, consider its internal structure at a deeper level. Over time, this line of work coagulated into an R package in its own right, with many advanced and useful new features that were not covered by `survey`. A tentative list of *ReGenesees* user-visible improvements over the `survey` package is presented in Table 2.

Table 2. Some ReGenesees improvements over the survey package

Feature	How it works
Calibration and variance estimation functions can efficiently process large-scale surveys even in environments with low computational resources (e.g., ordinary PCs)	• Exploits calibration models factorization through a dedicated divide and conquer algorithm • Accelerates execution and saves memory by means of ad hoc optimizations of many internal functions (e.g., variance, design effects and domain estimation for nonlinear estimators)
Provides estimates and sampling errors for arbitrary user-defined complex estimators, i.e., any nonlinear differentiable function of Horvitz-Thompson or calibration estimators of totals	• Enables users to define their own complex estimators symbolically (i.e., as mathematical functions) by means of R expressions • Exploits R symbolic differentiation facilities to linearize complex estimators automatically, so that their variance is estimated on the fly
Assists users in computing and organizing population totals for calibration tasks	• Driven by the calibration model formula, automatically generates a template dataframe to be filled with actual population totals • If the sampling frame of the survey is available, the template is filled automatically • Able to cope with sampling frames of several million rows and thousands of auxiliary variables by means of a dedicated adaptive chunking algorithm
Interaction with all summary statistics functions (i.e., estimators of totals, means, frequencies, ratios, quantiles, multiple regression coefficients, and complex estimators) has been standardized, so that they are easier to assemble in an industrialized process	• All estimators share (nearly) the same interface, even for domain estimation • All estimators produce return values with the same structure, even for subpopulation estimation • Estimates and sampling errors can be written to database tables or exported to external files in a common data model
New statistical capabilities and utilities	• Hints on feasible bounds for range-restricted calibration • Quick estimates of auxiliary variables totals • Compression of nested factors to reduce model-matrix sparseness in calibration tasks • Detailed diagnostics on the calibration process and on its results • Merge of new survey data into existing design objects • Collapsed strata technique for getting rid of lonely PSUs in variance estimation • Detailed diagnostics on the collapsing process and on the generated superstrata • Covariance and correlation between complex estimators • A generalized variance functions (GVF) infrastructure, i.e., facilities for defining, fitting, testing and plotting GVF models, and to exploit them to predict variance estimates

Table 2. Continued

Feature	How it works
Provides a comprehensive and user-friendly point-and-click GUI	• Pure R implementation, relies on tcltk and tcltk2 packages

Providing a comprehensive head-to-head comparison of *ReGenesees* to `survey` here would be beyond the present article's scope. However, we are able to present to the interested reader at least one example highlighting how the development trajectory of our system happened to diverge from `survey`. This example relates to the *partitioned* calibration functionality of *ReGenesees*, which was introduced in Subsection 4.1 and whose impact on variance estimation has been stressed in Sections 5 and 5.1.

In late 2006 we started studying, analyzing and testing `survey` in order to verify whether that package was able to satisfy, at least partially, the typical needs of Istat sample surveys. By using data from the Italian LFS as empirical test case, we soon realized that `survey` could not have been adopted at Istat "as it was" (Scannapieco et al. 2007). Indeed, every attempt to exploit its calibration function `calibrate` on LFS data invariably led either to memory allocation failures or to unaffordable execution times, whatever testing environment (i.e., hardware and operating system configuration) we set up. The point was that, despite being anything but naive, `survey` code was not optimized for processing such huge amounts of data.

To be slightly more specific: for each quarterly sample, typical LFS datasets have about 200,000 rows, and survey weights must be calibrated using over 4,000 numeric auxiliary variables. If the calibration task was to be tackled *globally* (and this is indeed the only available option in `survey`) those numbers would require: (i) computing a sample model matrix of over 8×10^8 numeric entries (i.e., about 6 GB of memory space), (ii) computing the cross-product of that model matrix, yielding a matrix of over 1.6×10^7 elements (i.e., about 120 MB), (iii) computing the (generalized) inverse of (an appropriate scaling of) that cross-product matrix, and (iv) repeating steps (ii)-(iii) for all the needed iterations of a Newton-Raphson algorithm until convergence. Note that, in spite of appearances, point (iii) – rather than (i) – turns out to be the hardest one, due to the high (typically cubic) computational and space complexity of generalized inverse algorithms.

As anticipated, the viable alternative we figured out was a "divide and conquer" calibration program exploiting a common feature of most Istat large-scale surveys, namely the *factorizable* nature of their calibration tasks. Coming back to the Italian LFS example: since known population benchmarks are defined at NUTS 2 level (i.e., for the 21 Italian administrative regions), *ReGenesees* calibration facility `e.calibrate` can split the unaffordable *global* calibration problem into 21 smaller subproblems. With respect to points (i)–(iii) defined above, each one of these local calibration subproblems involves matrices whose size is, on average, about 400 (i.e., 21^2) times smaller than those arising in the global approach.

Unfortunately, it would have been impossible to overcome `survey` limitations by locally modifying and extending just the calibration function, because the side effects of our partitioned algorithm would have propagated through `survey`'s variance estimation

backbone. The reason is that, as discussed in Section 5 of Estevao et al. (1995), solving the calibration problem in a partitioned way has subtle, nontrivial consequences in the variance estimation phase. Stated differently: in order to ensure that computed variance estimates of calibration estimators (as well as of functions of calibration estimators) are *identical*, *irrespective* of whether the underlying calibration problem has been solved globally or in a partitioned way, the software program addressing variance estimation must behave *differently* in the two cases. This explains why we were forced to override `survey`'s variance estimation facility and to implement dedicated solutions appropriate to our partitioned calibration approach.

Similarly, technical issues arising from the interplay between estimation domains and calibration domains when estimating the variance of partitioned calibration estimators in subpopulations (recall point (iv) at the end of Subsection 5.1.) prevented us from retaining `survey`'s workhorse function for domain estimation `svyby`. Again we had to override that function and to develop suitable alternatives.

In summary, none of the user-visible *ReGenesees* features reported in Table 2 as improvements over the `survey` package has been obtained as a simple add on. Instead, each one is the result of extensive and thorough programming effort at a deeper level. This is not surprising, given the sophisticated and highly specialized nature of both software tools.

Software Interoperability. As testified by recent standardization initiatives such as GSIM (UNECE 2013a) and CSPA (UNECE 2013b), NSIs are striving to modernize their production workflows in such a way that software components could be shared between different organizations and assembled "LEGO-wise" into industrialized processes. Devising and implementing a common information model and a standard production architecture are mandatory steps still to be completed to reach this challenging goal. Anyway, we think *ReGenesees* complies with many of the requirements implied by this modernization vision. Just to mention some of them: (i) it is free, (ii) it is open source, (iii) it is cross-platform, (iv) it supports input and output of data in "open" formats, (v) it is a collection of modules (namely R functions) with clearly defined interfaces, (vi) its main functionalities have been designed to reduce (or avoid, whenever possible) human intervention, (vii) its main functionalities are clearly mapped to GSBPM (UNECE 2013c), (viii) it is available for download and sharing by all interested users, (ix) its technical and end-user documentation is available in English.

8. Migrating Istat Procedures Towards ReGenesees

As already stated above, the first public release of the *ReGenesees* system is quite recent (December 2011). The software began to spread in Istat from late 2010 onwards during its beta-testing cycle. A recent (informal) internal survey revealed that – to date – it is being used in production by 20 Istat large-scale surveys. These include: (i) ten surveys on enterprises carried out in compliance with Eurostat regulations, including seven structural business surveys (e.g., Community Innovation Survey, Information and Communication Technology, Labour Cost Survey) and three short-term statistics surveys (e.g., Services Turnover Indices); (ii) three agri-environmental surveys (e.g., Survey on Permanent Crops); (iii) five surveys in the social demographic domain (e.g., Time Use, Health Conditions and Use of Medical

Services, Safety of Women); (iv) and two census-related surveys (Post Enumeration Survey of the 6th Agricultural Census and 9th Industry and Services Census).

In the opinion of survey statisticians involved, the migration of the standard calibration and estimation procedures from GENESEES to *ReGenesees* resulted in a significant reduction in both user workload and execution time. The observed 3:1 prevalence of business surveys on household surveys is arguably related to what we discussed at the end of Section 4: when samples are drawn from a centralized frame (like the Italian archive of enterprises ASIA), *ReGenesees* automatically computes the totals of the auxiliary variables from the sampling frame, and safely arranges and formats those values so that they can be directly used for calibration.

Though as yet partial, *ReGenesees* penetration can be deemed quite satisfactory, especially considering that SAS has been a *de facto* standard for statistical elaborations in Istat since the early '80s. Many Istat statisticians have strong SAS skills, some of them have been familiar with Istat's legacy estimation platform GENESEES for decades, and there are always some costs involved in changing a consolidated production workflow: it would have been unrealistic to expect an immediate transition to *ReGenesees*.

9. Ongoing Work and Future Extensions

Since its beta release, *ReGenesees* has been steadily gaining ground in Istat: to date, as already sketched above, it has been successfully integrated into the production workflow of about 20 large-scale surveys. Moreover, other Istat surveys are migrating to the new system at present. Surveys with a bigger "mass" (i.e., involving more actors operating in a more complex context) arguably have a greater "inertia" (i.e., are more resilient to changes). In this respect, our next challenge will be to undertake the migration towards *ReGenesees* of the production workflow of the Italian LFS and SME, a task whose *technical* feasibility has already been proven. Internal training courses dedicated to the new R-based system, whose first edition was launched in 2013, will allow an even faster and wider diffusion of *ReGenesees* in Istat production processes.

In the meantime, the *ReGenesees* project is in full swing and still growing. *ReGenesees* version 1.6, whose public release took place in April 2014, added facilities implementing the *Generalized Variance Functions* (GVF) method (Wolter 2007). The option of *predicting* variance estimates based on a fitted model explaining the variance of an estimator in terms of its expected value, rather than directly *computing* such variance estimates, can benefit surveys with very demanding dissemination schedules and publication plans.

We are currently assessing the feasibility of integrating the EVER package (Zardetto 2012) with *ReGenesees*, thus bringing the extended DAGJK technique (Kott 2001) for variance estimation into the latter system. This would make it possible to handle estimators which cannot be expressed as closed-form mathematical functions of sample observations, for example, due to hot-deck imputation variability (Miller and Kott 2011).

Another open line of research points towards the software implementation of the *generalized linearization* technique, which hinges upon the notion of *functional derivatives* of estimators (influence functions in the seminal paper of Deville (1999)). This would be useful whenever the ordinary Taylor method cannot be applied, for example, for

estimators which are based on order statistics or expressed as non-smooth, implicit functions of totals, like the Gini index, the at-risk-of-poverty rate, and other Laeken indicators (Osier 2009). Besides quantiles (which are available in *ReGenesees* too), Statistics Sweden's tool ETOS already provides sampling errors for the Gini index and quantile shares. Anyway, our aim would be to enable *ReGenesees* to cope with arbitrary user-defined functions of estimators of quantiles and cumulative distribution functions. Of course, we are aware of the difficulty of this task.

We are confident that some of the aforementioned methodological enrichments will be included in the next major release of the *ReGenesees* system, very likely together with further developments on the software engineering side.

Appendix

With respect to the ability to *automatically* estimate the variance of nonlinear estimators with the Taylor approach (thus excluding replication methods), Statistics Sweden's tool CLAN (nowadays extended by ETOS) is, to the best of our knowledge, the most general and flexible of the traditional estimation platforms cited in Section 2. Indeed, it is able to cope with arbitrary *rational* functions of estimators of totals. However, as anticipated in Section 5, this comes at the price of asking CLAN users to program SAS macros which become more and more complicated as the complexity of the desired estimator grows. The examples below illustrate this point, draw a comparison with *ReGenesees*, and eventually distil some insights from it.

Example 1

Suppose we want to estimate the mean of income (variable `income`) and its variance for each cell of a two-way table crossing age group (variable `agegrp`) and sector of activity (variable `sector`).

Ex1: CLAN Solution

The %FUNCTION SAS macro we would have to write in CLAN is as follows:

```
%macro function(a, b);
   %tot(tab, income, (sector = &a) and (agegrp = &b))
   %tot(nab, 1, (sector = &a) and (agegrp = &b))
   %div(rab, tab, nab)
   %estim(rab)
%mend;
```

Afterwards, in order to practically obtain the desired results in CLAN, we would have to embed the macro above into an enclosing SAS program ending with a call to macro %CLAN, and run that program.

Ex1: ReGenesees Solution

As *ReGenesees* users we would obtain the desired results by directly invoking function `svystatL` as follows:

```
> svystatL (ex1, expression(income/ones), by=~sector:agegrp)
```

Ex1: Comment

The most relevant point in this example is not the different amount of involved lines of code in itself, but rather the reason for such a difference: the logic with which users interact with CLAN and *ReGenesees*. In fact, while *ReGenesees* users must only ask the system *what* they need (i.e., they have to *invoke* a program), CLAN users must also provide the system with an explanation of *how* it has to compute *what* they need (i.e., they have to *write* a program).

As we are going to see in the next example, the effects induced by these diverse interaction paradigms become more evident as soon as a more complex estimator is addressed.

Example 2

Suppose we want to compute estimates and sampling errors for a product of two ratios between totals, say $\hat{Q} = (\hat{Y}_1/\hat{Y}_2) \times (\hat{Y}_3/\hat{Y}_4)$, again for each cell of the two-way table used in Example 1.

Ex2: CLAN Solution

The %FUNCTION SAS macro we would have to write in CLAN is as follows:

```
%macro function(a, b);
   %tot(y1ab, y1, (sector = &a) and (agegrp = &b))
   %tot(y2ab, y2, (sector = &a) and (agegrp = &b))
   %div(r1ab, y1ab, y2ab)
   %tot(y3ab, y3, (sector = &a) and (agegrp = &b))
   %tot(y4ab, y4, (sector = &a) and (agegrp = &b))
   %div(r2ab, y3ab, y4ab)
   %tot(q1ab, r1ab, (sector = &a) and (agegrp = &b))
   %tot(q2ab, r2ab, (sector = &a) and (agegrp = &b))
   %mult(Q, q1ab, q2ab)
   %estim(Q)
%mend;
```

Afterwards, in order to practically obtain the desired results in CLAN, we would have to embed the macro above into an enclosing SAS program ending with a call to macro %CLAN, and run that program.

Ex2: ReGenesees Solution

As *ReGenesees* users we would obtain the desired results by directly invoking function `svystatL` as follows:

```
> svystatL(ex2, expression((y1/y2)*(y3/y4)), by=~sector:agegrp)
```

Ex2: Comment

The most relevant point here is that the increased complexity of the estimator (as compared to Example 1) doesn't affect a *ReGenesees* user at all, whereas a CLAN user has to write a trickier and lengthier SAS macro. As explained in Section 5, this is because CLAN users always need to tell CLAN *how* to tackle the estimator they are interested in, and this is achieved by successively decomposing it into simpler subfunctions, until no further simplification is possible. This decomposition requires more and more steps as the estimator complexity grows.

10. References

Andersson, C. 2009. Using Auxiliary Information in the Calculation of Order Statistics and Estimated Totals in a Large Scale Production Environment. In Proceedings of the 57th Session of the International Statistical Institute (ISI). Durban, South Africa, 16–22 August 2009. Available at: http://isi.cbs.nl/iamamember/CD8-Durban2009/index.htm (accessed May 2015).

Andersson, C. and L. Nordberg. 1994. "A Method for Variance Estimation of Non-Linear Functions of Totals in Surveys – Theory and Software Implementation." *Journal of Official Statistics* 10: 395–405.

Bellhouse, D.R. 1985. "Computing Methods for Variance Estimation in Complex Surveys." *Journal of Official Statistics* 1: 323–329.

Caron, N. 1998. Le logiciel POULPE: aspects méthodologiques. In: INSEE: Actes des Journées de Méthodologie. Available at: http://jms.insee.fr/files/documents/1998/513_1-JMS1998_S3-1_CARON_P173-200.PDF (accessed May 2015).

Chambers, J.M. and T.J. Hastie. 1992. *Statistical Models in S*. London: Chapman & Hall.

Davidson, M. 2013. *Scottish Population Surveys Centralised Weighting Project. Weighting project report of the Scottish Government*. Available at: http://www.scotland.gov.uk/Topics/Statistics/About/Surveys/WeightingProjectReport (accessed August 2014).

Davies, P. and P. Smith. 1999. *Model Quality Report in Business Statistics. Volume II: Comparison of Variance Estimation Software and Methods, EUROSTAT*. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/MODEL%20QUALITY%20REPORT%20VOL%202.pdf (accessed August 2014).

Deville, J.C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87: 376–382.

Deville, J.C. 1999. "Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques." *Survey Methodology* 25: 193–203.

Estevao, V., M.A. Hidiroglou, and C.-E. Särndal. 1995. "Methodological Principles for a Generalized Estimation System at Statistics Canada." *Journal of Official Statistics* 11: 181–204.

EUROSTAT. 2002. Variance estimation methods in the European Union. Monographs of Official Statistics, Publications Office of the European Union, Luxembourg. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/MOS_20VARIANCE_ESTIMATION_202002.pdf (accessed August 2014).

EUROSTAT. 2013. Handbook on precision requirements and variance estimation for ESS households surveys. Methodologies & Working papers, Publications Office of the European Union, Luxembourg. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-13-029/EN/KS-RA-13-029-EN.PDF (accessed August 2014).

Falorsi, P.D. and S. Falorsi. 1997. "The Italian Generalised Package for Weighting Persons and Families: Some Experimental Results with Different Non-Response Models." *Statistics in Transition* 3: 357–381.

Kalton, G. 1979. "Ultimate Cluster Sampling." *Journal of the Royal Statistical Society* 142: 210–222.

Kott, P.S. 2001. "The Delete-A-Group Jackknife." *Journal of Official Statistics* 17: 521–526.

Krewski, D. and J.N.K. Rao. 1981. "Inference from Stratified Sample: Properties of Linearization, Jackknife, and Balanced Repeated Replication Methods." *The Annals of Statistics* 9: 1010–1019.

Lumley, T. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9: 1–19.

Lumley, T. 2010. *Complex Surveys: A Guide to Analysis Using R*. New York: John Wiley & Sons.

Miller, D. and P.S. Kott. 2011. "Using the DAG Jackknife to Measure the Variance of an Estimator in the Presence of Item Nonresponse." In Proceedings of the JSM (July 30–August 4, 2011) Alexandria, VA: American Statistical Association, 1121–1129. Available at: http://nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/conferences/JSM-2011/JSM-2011-Miller.pdf

Mohl, C. 2007. "The Continuing Evolution of Generalized Systems at Statistics Canada for Business Survey Processing." In Proceedings of the Third International Conference on Establishment Surveys (ICESIII) (June 18–21, 2007), American Statistical Association, 758–768. Available at: http://www.amstat.org/meetings/ices/2007/proceedings/ICES2007-000135.PDF

Nieuwenbroek, N., R. Renssen, and L. Hofman. 2000. "Towards a Generalized Weighting System." In Proceedings of the Second International Conference on Establishment Surveys (ICESII) (June 17–21, 2000), American Statistical Association, 667–676. Available at: http://www.amstat.org/meetings/ices/2000/proceedings/S09.pdf

Ollila, P., Y. Berger, H.J. Boonstra, A. Davison, A. Laaksonen, K. Magg, R. Munnich, D. Ohly, S. Sardy, K. Sostra, and J. van den Brakel. 2004. "Evaluation of Software for Variance Estimation in Complex Surveys, DACSEIS project, Deliverables 4.1 and 4.2." . Available at: https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Dacseis_Deliverables/DACSEIS-D4-1-4-2.pdf (accessed August 2014).

Osier, G. 2009. "Variance Estimation for Complex Indicators of Poverty and Inequality Using Linearization Techniques." *Survey Research Methods* 3: 167–195.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org (accessed August 2014).

Rust, K. and G. Kalton. 1987. "Strategies for Collapsing Strata for Variance Estimation." *Journal of Official Statistics* 3: 69–81.

Särndal, C.-E., B. Swensson, and J. Wretman. 1989. "The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total." *Biometrika* 76: 527–537. Doi: http://dx.doi.org/10.1093/biomet/76.3.527.

Särndal, C.-E. 2007. "The Calibration Approach in Survey Theory and Practice." *Survey Methodology* 33: 99–119.

Sautory, O. 1993. La macro CALMAR: Redressement d'un Echantillon par Calage sur Marges. Document de travail de la Direction des Statistiques Démographiques et Sociales, no. F9310. Available at: http://www.insee.fr/fr/methodes/outils/calmar/doccalmar.pdf (accessed May 2015).

Scannapieco, M., D. Zardetto, and G. Barcaroli. 2007. *La Calibrazione dei Dati con R: una Sperimentazione sull'Indagine Forze di Lavoro ed un Confronto con GENESEES/SAS*. Collana Contributi, 4, Istat, Italy. Available at: http://www3.istat.it/dati/pubbsci/contributi/Contributi/contr_2007/2007_4.pdf (accessed August 2014).

Scottish Government. 2013a. *Scottish Household Survey – Methodology and Fieldwork Outcomes 2012*. Available at: http://www.scotland.gov.uk/Resource/0044/00443332.pdf (accessed August 2014).

Scottish Government. 2013b. *Scottish Health Survey 2012 – Volume 2 Technical Report*. Available at: http://www.scotland.gov.uk/Resource/0043/00434643.pdf (accessed August 2014).

Scottish Government. 2014. *Scottish Crime and Justice Survey 2012/13 – Technical Report*. Available at: http://www.scotland.gov.uk/Resource/0044/00445791.pdf (accessed August 2014).

UNECE. 2013a. *Generic Statistical Information Model (GSIM), version 1.1*. Available at: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId = 59703371 (accessed August 2014).

UNECE. 2013b. *Common Statistical Production Architecture (CSPA), version 1.0*. Available at: http://www1.unece.org/stat/platform/display/CSPA/CSPA+v1.0 (accessed August 2014).

UNECE. 2013c. *Generic Statistical Business Process Model (GSBPM), version 5.0*. Available at: http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model (accessed August 2014).

Vanderhoeft, C. 2001. *Generalised Calibration at Statistics Belgium. SPSS Module g–CALIB–S and Current Practices*. Statistics Belgium Working Paper no. 3. Available at: http://statbel.fgov.be/nl/binaries/paper03%5B1%5D_tcm325-35412.pdf (accessed May 2015).

Wilkinson, G.N. and C.E. Rogers. 1973. Symbolic Description of Factorial Models for Analysis of Variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 22: 392–399.

Wolter, K.M. 2007. *Introduction to Variance Estimation*, Second Edition. New York: Springer.

Woodruff, R.S. 1952. Confidence Intervals for Medians and Other Position Measures. *Journal of the American Statistical Association* 47: 635646. Doi: http://dx.doi.org/10.1080/01621459.1952.10483443

Woodruff, R.S. 1971. "A Simple Method for Approximating the Variance of a Complicated Estimate." *Journal of the American Statistical Association* 66: 411–414. Doi: http://dx.doi.org/10.1080/01621459.1971.10482279.

Zardetto, D. 2012. *EVER: Estimation of Variance by Efficient Replication*. R package version 1.2, Istat, Italy. Available at: http://cran.r-project.org/web/packages/EVER/index.html (accessed August 2014).

Zardetto, D. 2014. *ReGenesees: R Evolved Generalized Software for Sampling Estimates and Errors in Surveys*. R package version 1.6, Istat, Italy. Available at: https://joinup.ec.europa.eu/software/regenesees/description (accessed August 2014).

Zardetto, D. and R. Cianchetta. 2014. *ReGenesees.GUI: a TclTk Interface for the ReGenesees Package*. R package version 1.6, Istat, Italy. Available at: https://joinup.ec.europa.eu/software/regenesees/description (accessed August 2014).

Received July 2013
Revised August 2014
Accepted August 2014

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 205–229, http://dx.doi.org/10.1515/JOS-2015-0014

Dwelling Price Ranking versus Socioeconomic Clustering: Possibility of Imputation

Larisa Fleishman[1], *Yury Gubman*[1], *and Aviad Tur-Sinai*[1]

In order to characterize the socioeconomic profile of various geographic units, it is common practice to use aggregated indices. However, the process of calculating such indices requires a wide variety of variables from various data sources available concurrently. Using a number of administrative databases for 2001 and 2003, this study examines the question of whether dwelling prices in a given locality can serve as a proxy for its socioeconomic level. Based on statistical and geographic criteria, we developed a Dwelling Price Ranking (DPR) methodology. Our findings show that the DPR can serve as a good approximation for the socioeconomic cluster (SEC) calculated by the Israel Central Bureau of Statistics for years when the required data was available. As opposed to the SEC, the suggested DPR indicator can easily be calculated, thus ensuring a continuum of socioeconomic index series. Both parametric and nonparametric statistical analyses have been carried out in order to examine the additional social, demographic, location, crime and security effects that are exogenous to SEC. Complementary analysis on recently published SEC series for 2006 and 2008 show that our conclusions remain valid. The proposed methodology and the obtained findings may be applicable for different statistical purposes in other countries which possess dwelling transactions data.

Key words: Housing market; urban locality; index construction.

1. Introduction

The socioeconomic profile of a residential area can be identified and characterized in two different ways: by using specific demographic and socioeconomic factors, or by estimating aggregated indices based on a range of these factors. In Israel, a socioeconomic index (SEI) for local authorities was developed at the Central Bureau of Statistics (CBS) in the mid-1990s. The main and most important use of the data on SEI is its contribution to the design and implementation of different policies in various ministries and other governmental agencies relating to local authorities, including various resource allocation procedures. However, the main limitation in using SEI is that the SEI data is not available for all localities every year, since the process of SEI calculation requires a wide variety of variables from various data sources available concurrently for various localities.

In previous studies, there is well-documented evidence of the relationship between the various demographic and socioeconomic characteristics of a locality and its dwelling prices. In Israel, the main source of information about dwelling prices is the record of real estate transactions kept by the Israel Tax Authority (ITA). The Israel CBS has been receiving this database on a monthly basis since 1990. Using this data source, our study

[1] Israeli Central Bureau of Statistics, 66 Kanfei Nesharin St. Jerusalem 91342, Israel. Emails: larisaf@cbs.gov.il, yuryg@cbs.gov.il, and aviadts@cbs.gov.il

proposes a method for assessing the socioeconomic level of those localities and for those time points for which these data are missing.

The current study has two main goals: 1) to examine if dwelling prices can serve as a proxy for the socioeconomic level of various urban localities; 2) to examine the extent of correlation between social and demographic characteristics not included in the SEI calculations and the level of dwelling prices in a given locality.

In order to achieve the first goal, relevant urban localities in Israel were graded according to their dwelling price level, and this was compared to their socioeconomic level. A similar analysis was conducted for the years 2006 and 2008. To achieve the second goal, parametric models were estimated. In this way conclusions were drawn regarding the possibilities for imputation of missing data on SEI values.

This study augments the relevant research literature in several ways. First, previous studies have focused primarily on the effect of demographic and socioeconomic factors on the level of dwelling prices in specific cities or metropolitan areas. We extended this investigation to a national level. Furthermore, it should be noted that Israel is a highly urbanized country with more than ninety percent of Israelis living in urban areas. Second, our investigation focused on the correlation between socioeconomic indices and dwelling price indicators. Only limited research has been carried out on the degree of correspondence between these two indicators. Third, this study proposes, examines and discusses a reasonable alternative method for assessing the socioeconomic level of urban localities. This method does not require a range of multivariate procedures and various statistical data available concurrently, although it allows for a sound approximation to a socioeconomic level. Research thus far has not addressed this issue.

The rest of this article is organized as follows: Section 2 presents some theoretical background and a survey of relevant literature, and explains the construction of the socioeconomic index in Israel. Section 3 describes the main source of data, the construction of the key variable reflecting dwelling price level, and defines geographic units for which this variable was created. Section 4 presents and discusses the degree of correspondence between the dwelling price level and socioeconomic level, describes additional sources of information and defines the variables used in the analysis. Section 5 presents the statistical models that we used for the empirical analysis, gives the results and discusses them. Section 6 concludes the study.

2. Literature Review

2.1. Dwelling Price in Light of Socioeconomic Level Characteristics

The price of a residential property on the free market reflects the willingness of the purchaser to pay not only for the property itself, but also for a specific residential environment – in other words, for the quality of "social space" (Reed 2001).

Research findings from around the world testify to the range of factors which reflect the socioeconomic essence of a residential area, the most important of which are income, education, employment, and the demographic characteristics of its population. Earlier studies show a significant positive correlation between the three major factors characterizing the residential area's socioeconomic level – income, education and

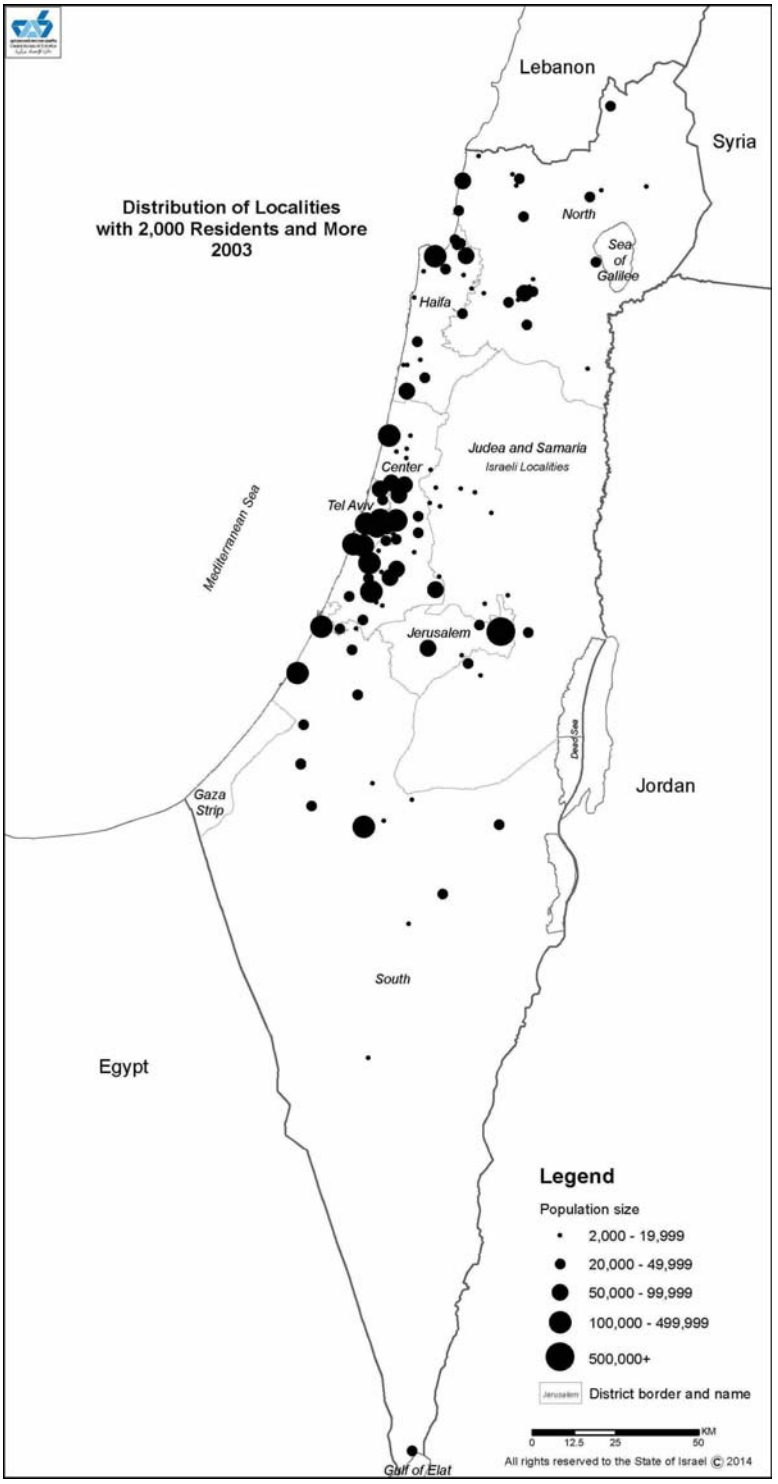
employment – and the price of dwellings (Heikkila 1992; Potepan 1996; Goodman and Thibodeau 1998; Greenberg 1999; Des Rosiers et al. 2002; Yates 2002). Income is considered the primary factor (Ozanne and Thibodeau 1983; Malpezzi et al. 1998). Those with relatively high incomes choose their residential area in an attempt to avoid neighbors with a low socioeconomic status. The popular viewpoint considers social problems, such as crime, drug use, and the neighborhood's economic decline resulting in neglected buildings, as all directly linked to neighborhoods characterized by a high proportion of unemployed and low levels of education and income (Harris 1999; Jackson et al. 2007). The study of Cummings et al. (2002) examined education as one of the dimensions of the socioeconomic level of residents of various neighborhoods in the city of Philadelphia and its influence on the price of residential dwellings in those neighborhoods. The study findings show a 21 percent increase in dwelling prices with every ten percent rise in the proportion of adults with post-high school education.

Aside from the aforementioned factors that characterize the socioeconomic space and impact dwelling price, the relevant literature has examined the relationship between the demographic characteristics of residents, such as age and marital status, and dwelling prices in that neighborhood (Myers 1990; Heikkila 1992). There are studies indicating that ethnic composition and personal security in a residential environment may also contribute to the socioeconomic space, and as a result affect the price of dwellings. In particular, previous studies provide sound evidence of a strong positive correlation between the level of personal security in a residential area and its dwelling prices (Thaler 1978; Dubin and Goodman 1982; Buck et al. 1991; Hazam and Felsenstein 2007).

Earlier studies have also addressed the effect of immigrant groups and the racial-ethnic context of residential areas on the local housing prices (Kiel and Zabel 1996). However, there is no common agreement on either the existence or the magnitude of the effect of immigration shocks on the housing market. The magnitude of the effect immigrants have on housing prices depends heavily on the reaction of natives to the presence of immigrants in the area. For example, some studies in the USA have found that blacks and Hispanics own homes of lesser value than the white population. This held true even when the researchers controlled for the characteristics of dwellings (Horton and Thomas 1998; Krivo 1995; Lewin-Epstein et al. 1997). Harris (1999) found that dwelling prices in the USA decline by an average of 16 percent when the Afro-American population exceeds ten percent in a neighborhood. Furthermore, a far more dramatic drop in prices occurs when the Afro-American population exceeds 60 percent of the neighborhood residents. According to this study, the explanation is not necessarily ethnic preference, but may be related to social problems stemming from the socioeconomic status of the Afro-American population, which is usually lower. However, a study conducted in Darwin (Australia) by Jackson et al. (2007) revealed high positive correlation between housing prices and ethnicity for people born overseas and for those who speak other languages.

Along with the abovementioned demographic and socioeconomic characteristics of a population, religiosity contributes substantially to the residential profile of a locality (Blanchard 2007). Among several religiosity patterns displayed by the Israeli Jewish population, both ultra-Orthodox and Orthodox streams play a significant role with regard to the socioeconomic, ethnic and spatial divide in Israeli society, thus notably effecting local dwelling market price level (Cahaner 2012).

Map 1. Distribution of Urban Localities in Israel



In addition, the importance of the geographical location of dwellings in the context of their price level has been emphasized in several studies (McCluskey et al. 2000; Bourassa et al. 2003). In Israel this issue is of special importance. According to the official administrative division, there are six main administrative districts in Israel. Map 1 shows the distribution of urban localities as well as administrative districts. Tel Aviv district is composed of the central city of Tel Aviv and other cities enclose it on three sides. This district is the geographical center of Israel, and is characterized by a very highly concentrated urban population. It is the financial, economic, social, and cultural center of the country.

Tel Aviv district is dominant and influential in the domains of employment and communication, as well as the domain of land and dwellings prices (Soffer and Bystrov 2006). Therefore, the geographical proximity of a locality to the center, or to the periphery, affects many aspects of life in a locality, including the socioeconomic level of the population and dwelling price level. Summarizing this short literature review, it can be concluded that a sizeable body of literature provides evidence of a strong association between various demographic, social and economic characteristics of a locality and the price of its dwellings . This evidence serves as a theoretical foundation for the premise behind this study: the price of dwellings in a specific locality can serve as an alternate measure of its socioeconomic level.

2.2. *The Socioeconomic Index and the Socioeconomic Cluster*

In order to characterize and document the socioeconomic profile of various localities, it is common practice in the official statistics of various countries to use aggregated indices (Burck and Kababia 1996, 1999; Australian Board of Statistics 2006). These indices are based on different theoretical assumptions and estimation methods, and may be classified in accordance with two main approaches. Using a "deterministic" approach, a socio-economic cluster is specified by applying predefined classification criteria based on an underlying conceptual model. For instance, Rose and Prevalin (2001) suggest the set of employment status and occupation variables for socioeconomic classification in the UK, following the social class classification methodology earlier suggested by Olausson and Vagero (1991) for Swedish register data.

The "stochastic" approach assumes the existence of a latent continuous variable (Y) for the socioeconomic level of a given locality. It is also assumed that Y may be assessed using multivariate analysis methodology on a set of observed variables. Finally, localities are clustered by the estimated values of Y. For example, Jackson et al. (2007) suggest estimating socioeconomic level using principal component methodology applied to a wide set of socioeconomic characteristics which includes income, age, family status, dwelling data, and so on. Generally, Principal Component Analysis (PCA) is a technique that is useful for the compression and classification of data. The main idea of PCA is to reduce the dimensionality of a data set which contains a large number of interrelated variables. This reduction is achieved by finding a new set of uncorrelated variables (the principal components) smaller than the original set of variables that nonetheless retains most of the variation present in the original data set (Jolliffe 2002).

The stochastic approach is widely used for socioeconomic index calculation in the official statistics of different countries. The Office for National Statistics in the UK devises the

socioeconomic index for areas within local authorities by means of principal component analysis based on population censuses. The Australian Bureau of Statistics produces five socioeconomic indices that measure various socioeconomic aspects of residential areas based on the population census. Surveys are used for updating the index in the periods between population censuses. A similar methodology is used in New Zealand.

In Israel, a socioeconomic index (SEI) was developed at the Central Bureau of Statistics (CBS) in the mid-1990s on the basis of the 1995 Population and Housing Census data (Burck and Kababia 1996). It is based on five groups of variables that include 14 variables. The variables used to construct the index reflect all of the aspects related to the socioeconomic makeup of the population of different localities, subject to the availability of the data (for more details on the selection of the variables, see CBS 2000). These five groups contain the following variables: (1) demographic characteristics (dependency ratio, median age, percentage of families with four or more children); (2) education and schooling (percentage of the students studying for a bachelor's or higher degree, percentage eligible for a matriculation certificate); (3) standard of living (level of motorization, percentage of new motor vehicles, average income per capita); (4) labor force statistics (percentage of job seekers, percentage of salaried workers and self-employed persons earning up to minimum wage, percentage of salaried workers earning more than twice the average salary); (5) support/pension (percentage receiving unemployment benefits, percentage receiving income supplements; percentage receiving old age pensions with income supplements). SEI is based on the stochastic approach and calculated using principal component analysis.

Principal components (factors) are essentially new variables, calculated as a linear combination (weighted average) of the original, standardized variables (i.e., each variable has a mean of 0 and variance of 1). The weights of the original standardized variables are determined mathematically so as to maximize the differences in the scores between the geographical units, subject to some normalization restrictions. The factors are determined sequentially, so that the first factor is the linear combination that accounts for the maximum amount of the variance of the variables. Hence, the first factor has the greatest ability to discern between the localities. The second factor accounts for a maximum variance not accounted for by the first factor, and so on. The optimal number of factors that should be used to explain the maximum amount of the variance of the variables is determined by statistical testing. It is noteworthy that since the variables are standardized, the total variance of the original variables is equal to the number of variables. These factors define an orthogonal set of axes in the multidimensional variable space where each factor is a linear combination of the original variables. This type of factor analysis can be defined as PCA (CBS 2000).

We can represent the socioeconomic index as follows:

$$SEI = a_1X_1 + a_2X_2 + ... + a_{14}X_{14} \quad (1)$$

with SEI indicating the socioeconomic index (continuous), $a_1, \ldots, a_{14}$ the coefficients calculated using principal component analysis, and $X_1, \ldots, X_{14}$ the variables constituting the index which were specified above.

The set $X_1, \ldots, X_{14}$ has been defined based on the methodological background identified in the relevant literature, while considering local conditions and data

availability. It should be noted that, as opposed to Jackson et al. (2007), the dwelling prices are not included in the SEI calculations.

The SEI estimates cover most local authorities for which all the variables listed above are available at the relevant time point. For the sake of consistency and comparability of the SEI series, the set of variables and the calculation methodology have not been changed over the years.

Using cluster analysis, the local authorities, for which the SEI is calculated, are then divided into ten socioeconomic clusters (SEC), with Cluster 1 including authorities with the lowest socioeconomic level, and Cluster 10 including authorities with the highest socioeconomic level.

After 1995, the SEI was updated for the years 1999, 2001, 2003, and 2006 when the required data was available. The SEI calculated in 2008 on the basis of the 2008 Population Census is the most recent. As dwelling price data are available annually, we suggest a univariate deterministic approach for approximation of the socioeconomic level of a given locality at a given time.

In this context, it is worth noting that the relevant literature is mostly dedicated to the correlation between the different demographic and socioeconomic characteristics of a locality and the dwelling prices in it, and the examination of the degree of correspondence between aggregated socioeconomic indices and dwelling prices remains beyond the scope of research. Thus, to reach a conclusion as to the ability of dwelling prices to serve as a proxy for the socioeconomic level, we first need to check the degree of correspondence between them. Afterwards, we examine certain additional social and demographic characteristics that are not currently included in the SEI calculations, but are known to affect dwelling prices.

3. Data and Definitions

The study is based on files of dwelling transactions in the housing market in 2001 and 2003. Transaction data are provided annually by the Israel Tax Authority. In total, the basic file from 2001 included 60,851 transactions, and the file from 2003 included 57,223 transactions.

In order to compare the SEC of a given locality with its aggregate dwelling price level, the same coding scheme had to be used for both indicators. Dwelling Price Ranking (DPR) was constructed based on the following steps. First, using the transactions data, dwelling price level (DPL) for each relevant locality k was calculated:

$$DPL_k = log(median(Y_k)), \tag{2}$$

where Y denotes price per square meter. Using price per square meter as an underlying variable for DPR, we neutralize the effect of apartment size, one of the main variables which explains differences in dwelling prices (Lozano-Gracia and Anselin 2012), and represents as far as possible the market value of a dwelling at the aggregate level for a given locality. The median is used for reasons of robustness, and the log-transformation stabilizes the variance and generally makes the data normally distributed.

Second, localities for which the DPR was created were selected using the following criteria: (1) total population of 2,000 or more in locality, which corresponds to the

definition of "urban" in Israel; (2) the number of transactions in a locality should be sufficient enough to represent the price level in the housing market (at least 15 in the current study). Localities that did not match the above criteria were excluded from the analysis. The final data set includes 104 localities in 2001 and 112 localities in 2003, covering about 90 percent of the Israeli population.

Third, the selected localities were divided into ten clusters, alongside the SEC. Localities with the lowest DPL were ranked as Level 1 (DPR $= 1$), while the localities with the highest dwelling prices were ranked as Level 10 (DPR $= 10$). Each of the resulting DPR clusters contains approximately the same number of localities.

4. Dwelling Price Ranking vs. Socioeconomic Cluster

In order to examine the degree of correspondence between the SEC and the DPR, a correlation analysis was carried out. The obtained Spearman correlation coefficients are equal to 0.69 and 0.67 for 2001 and 2003, respectively.

Table 1 presents the detailed results for 2003; the analysis for 2001 revealed similar results. In Table 1, the digit in each cell indicates the number of localities with DPR and the SEC as they appear in the rows and columns, respectively. The cases in which both rankings are identical appear in bold print; there is an exact correspondence between the SEC and the DPR for 21 localities (out of 112).

Based on these results, it can be concluded that a lack of correspondence between the SEC and the DPR is more typical for localities where the SEC is low or low-medium. The minimal gap between the SEC and the DPR (± 1 range) is observed for 28 percent of the localities. For localities where the gap between the SEC and the DPR is greater than 2, it was found that localities with a DPR higher than their SEC are mainly situated close to the Tel Aviv district, that is, close to the center of the country. Those are localities where the SEC is medium-high to high (6–9). For localities with a low-medium SEC or medium SEC (3–5), in most cases the DPR was lower than the SEC. Those localities are located in the more peripheral areas.

Table 1. The socioeconomic cluster vs. the dwelling price ranking 2003

	The socioeconomic cluster										
DPR	1	2	3	4	5	6	7	8	9	10	Total
1	–	1	2	6	1	1	–	–	–	–	11
2	–	–	–	3	7	–	1	–	1	–	12
3	–	1	–	4	4	2	–	–	–	–	11
4	–	–	2	**4**	2	1	–	1	1	–	11
5	1	1	–	–	**5**	4	–	1	–	–	12
6	–	1	–	1	1	**3**	4	1	–	–	11
7	–	–	–	–	1	2	**5**	1	–	–	9
8	–	–	–	–	1	2	3	**3**	2	1	12
9	–	1	–	1	–	1	5	4	–	–	12
10	–	–	–	–	–	–	–	9	1	**1**	11
Total	1	5	4	19	22	16	18	20	5	2	112

These findings indicate the spatial aspects contained in the correlation between the SEC and the DPR. In particular, it can be concluded that the dependence of dwelling prices on the distance from the Tel Aviv district is stronger than the spatial dependence for the SEC. That is, it is rare to find very expensive dwellings in the peripheral regions, while there are some peripheral localities with a comparatively high socio-economic profile.

It can also be seen that the degree of correspondence between the two indicators increases with the rise in the SEC scale of the localities. Maps 2 and 3 illustrate spatial distribution of the localities according to the SEC and the DPR for 2003.

We conclude that the suggested dwelling price ranking appears to be a sufficiently good approximation for the socioeconomic cluster. However, a gap is revealed between two indicators, and it is therefore reasonable to assume that there are other factors influencing dwelling prices. Using these factors, we attempted to correct the developed dwelling price ranking by reducing the gap between the DPR and the SEC.

In order to examine additional factors that are not included in the SEC calculation but are assumed to influence dwelling prices, the following administrative databases are used. First, the Population Registry from which information on population characteristics by locality was obtained (e.g., percentage of immigrants from the former USSR and Ethiopia in 2001 and 2003). Second, the "Level of Religiosity" administrative file that was developed at the CBS serves as a basis for such variables as the percentage of ultra-Orthodox and Orthodox population by locality. Third, a crime database was provided by the Israeli Police. Finally, terror incidents data for relevant years was created by using information from different sources available from the International Institute for Counter-Terrorism (ICT) at the Interdisciplinary Center (IDC) Herzliya, Ministry of Foreign Affairs and the Prime Minister's Office. Additionally, spatial information regarding the location of the localities relative to the center of the country was provided by the Geographic Information System (GIS).

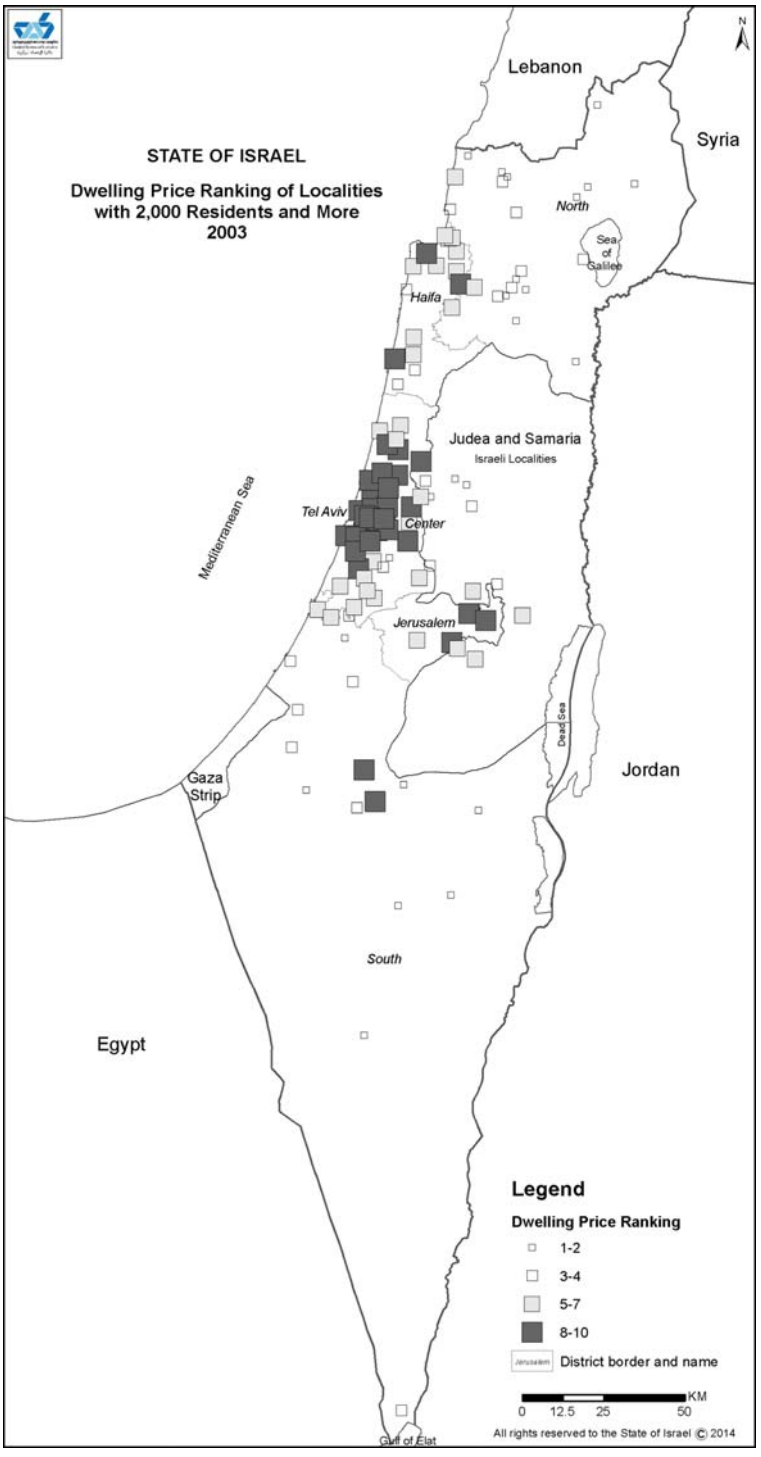
On the basis of these databases, a set of explanatory variables were selected based on the existing literature in this field partly reviewed in Subsection 2.1. Appendix 1 presents and defines the variables that we used in the study, their means, standard deviations and medians.

In order to examine the degree of correlation between the DPL and the selected variables, a correlation analysis was carried out. The Spearman correlation coefficients are presented in Table 2. Note that the Spearman coefficient is used since the distribution of most explanatory variables is skewed.

Of all the variables having a significant correlation with the DPL, there are six variables characterized by a positive correlation with the DPR: the total population in a locality, the rate of cases of property crimes, the number of terror incidents and the three variables indicating the geographic district of a locality – Jerusalem District, Center District and Tel Aviv District.

The degree of correlation between the SEC and the selected variables was also examined. The results of this test are presented in Appendix 2. The correlation analysis on both the DPL (Table 2) and the SEC variables (Appendix 2) revealed similar results.

Map 2. Dwelling Price Ranking of Localities



Map 3. Socioeconomic Cluster of Localities

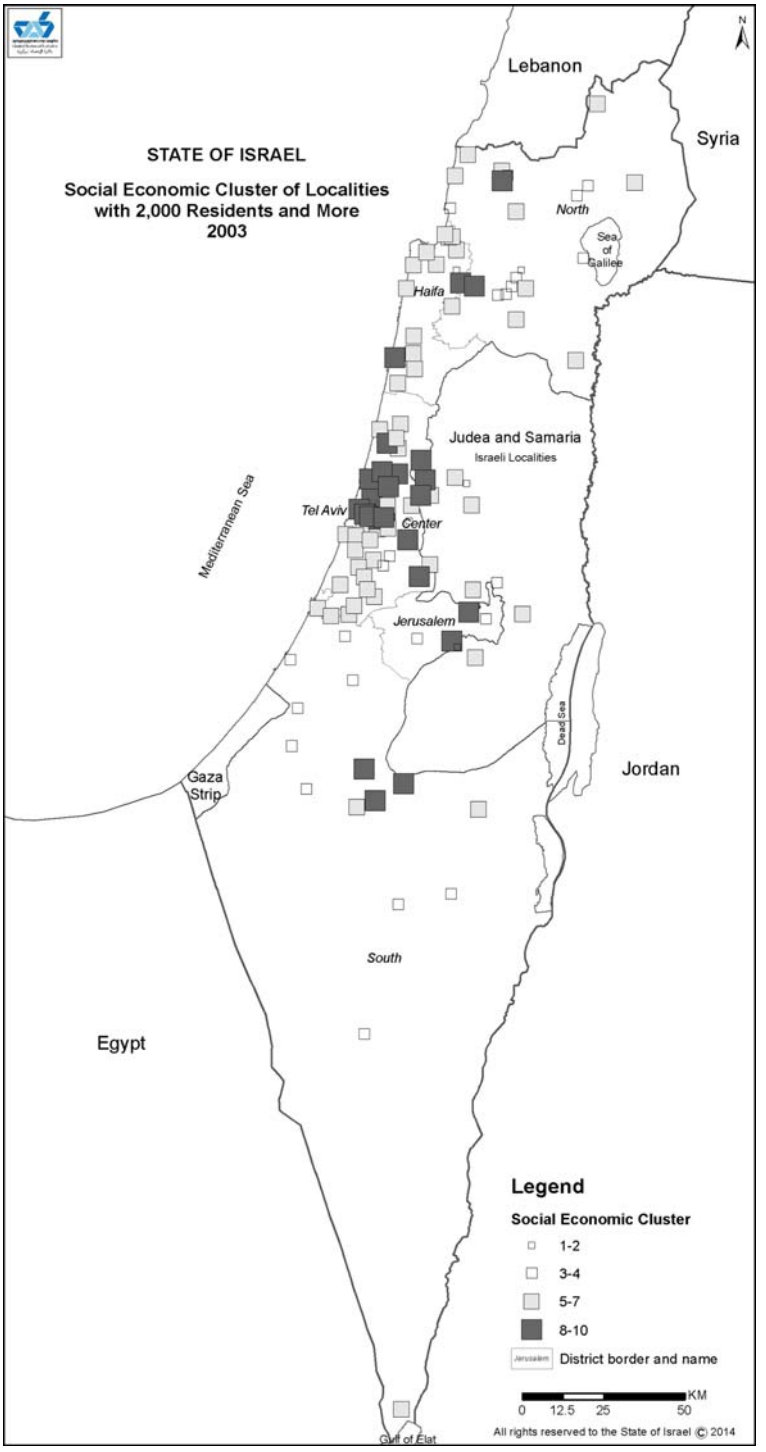


Table 2. Correlations between the DPL and the explanatory variables

Variables	2001		2003	
	Correlation coefficient	Level of significance	Correlation coefficient	Level of significance
Total population	0.169	0.086	0.264	0.005
Percentage of Arab population	−0.493	<0.001	−0.418	<0.001
Percentage of Orthodox population	−0.307	0.002	−0.340	<0.001
Percentage of ultra-Orthodox population	−0.307	0.002	−0.265	0.005
Percentage of immigrants from the former USSR since 1990	−0.423	<0.001	−0.379	<0.001
Percentage of Ethiopian immigrants	−0.205	0.040	−0.068	0.481
Rate of cases of bodily injury crimes	−0.453	<0.001	−0.459	<0.001
Rate of cases of property crimes	0.294	0.002	0.292	0.002
Number of terror incidents	0.204	0.038	0.132	0.166
Districts: Jerusalem	0.124	0.211	0.174	0.066
North	−0.446	<0.001	−0.473	<0.001
Haifa	−0.014	0.885	0.018	0.849
Center	0.346	<0.001	0.410	<0.001
Tel Aviv	0.449	<0.001	0.440	<0.001
South	−0.353	0.001	−0.315	<0.001
Distance from the Tel Aviv district	−0.695	<0.001	−0.721	<0.001

5. Parametric Models and Findings

Two parametric models were estimated: multinomial logistic regression for the SEC variable and the OLS model for the DPL. The models were estimated only for the Jewish sector for the following reasons. The housing market in the Arab sector operates under different conditions than that in the Jewish sector, and some of the explanatory variables are irrelevant to the Arab sector (such as the percentage of new immigrants from the former USSR and Ethiopia). Furthermore, the number of localities in the Arab sector with sufficient number of transactions was inadequate for performing statistical analyses for the Arab sector solely (four localities).

In order to avoid possible multicollinearity, those explanatory variables that were found to be highly correlated with other explanatory variables (Pearson correlation coefficient is more than 0.5) were excluded from the parametric models. These variables were included in the nonparametric analysis presented in Subsection 5.3.

5.1. A Multinomial Logistic Model for the SEC Variable

To estimate the marginal contribution of each of the above factors to the SEC, a regression analysis was carried out. Since the SEC is categorical, a multinomial logistic regression model was estimated, with the dependent variable being the probability of being in cluster i:

$$P(SEC = i) = log\,it(\alpha_i + \beta DPR + \gamma X) \tag{3}$$

Table 3. Multinomial models for the socioeconomic cluster

Variable	2001			2003		
	Estimate	*p*-value	Odds ratio	Estimate	*p*-value	Odds ratio
Intercept for ranking = 10	−6.16	0.000	–	−5.22	0.000	–
Intercept for ranking = 9	−4.05	0.004	–	−3.58	0.005	–
Intercept for ranking = 8	−1.38	0.310	–	−0.44	0.707	–
Intercept for ranking = 7	0.75	0.589	–	2.35	0.059	–
Intercept for ranking = 6	2.42	0.083	–	4.67	0.000	–
Intercept for ranking = 5	5.52	0.000	–	8.32	<0.001	–
Intercept for ranking = 4	9.31	<0.001	–	18.85	<0.001	–
Intercept for ranking = 3	11.22	<0.001	–	21.64	<0.001	–
Intercept for ranking = 2	12.79	<0.001	–	33.82	<0.001	–
DPR	0.61	<0.001	1.85	0.50	<0.001	1.65
Percentage of Orthodox population	−0.04	0.042	0.96	−0.03	0.051	0.97
Percentage of ultra-Orthodox population	−0.09	<0.001	0.91	−0.29	<0.001	0.75
Percentage of Ethiopian immigrants	−0.15	0.318	0.86	−0.59	0.000	0.55
Percentage of immigrants from the former USSR since 1990	−0.12	<0.001	0.89	−0.12	<0.001	0.89
Rate of cases of bodily injury crimes	−0.27	<0.001	0.76	−0.30	<0.001	0.74
Rate of cases of property crimes	−0.31	0.068	0.74	−0.0003	0.43	1.00
North district	1.46	0.015	4.33	0.25	0.66	1.29
South district	1.64	0.032	5.12	0.97	0.209	2.65
Number of terror events	−0.18	0.081	0.84	−0.02	0.839	0.98
Number of observations			98			107
Percent concordant			94.2			94.6

with $\alpha_i, i = 2, \ldots, 10$ being the intercepts of the model for cluster values 2,. . .,10 respectively, where cluster 1 was chosen to be the reference category. In (3), γ denotes a vector of the regression coefficients to be estimated and X the set of explanatory variables.

Table 3 presents the final estimated models for 2001 and 2003, with variables that are significant for at least one of the years (significance level 0.10).

It can be seen that there is a positive correlation between the DPR and the probability of appearing in a higher SEC, given all the other controlled variables.

However, it was found that this influence is partially offset as a result of the effect of minorities, for example the percentage of religious population (both ultra-Orthodox and Orthodox), the percentage of immigrants from Ethiopia, and the percentage of immigrants from the former USSR.

Given all other controlled variables, including DPR, it appears that the location in peripheral districts increases the odds for being in a higher SEC.

The effects of both bodily injuries and property crimes as well as the number of terrorism incidents were found to be significant and negative.

5.2. *OLS Model on DPL*

A regression model was estimated for the continuous DPL variable which served for constructing the DPR.

In order to validate the obtained estimators, we carried out appropriate statistical tests to identify possible multicollinearity, residual dependence and residual normality. Figures 1A and 2A (Appendix 3) demonstrate that the DPL is close to normally distributed, justifying use of the OLS model. Figures 3A and Table 1A (Appendix 3) display the residuals normal probability plots for the estimated OLS models, showing that the residuals' distribution is approximately normal. Statistical test results (such as the Durbin-Watson test and 1st Order Autocorrelation test for residual independence,

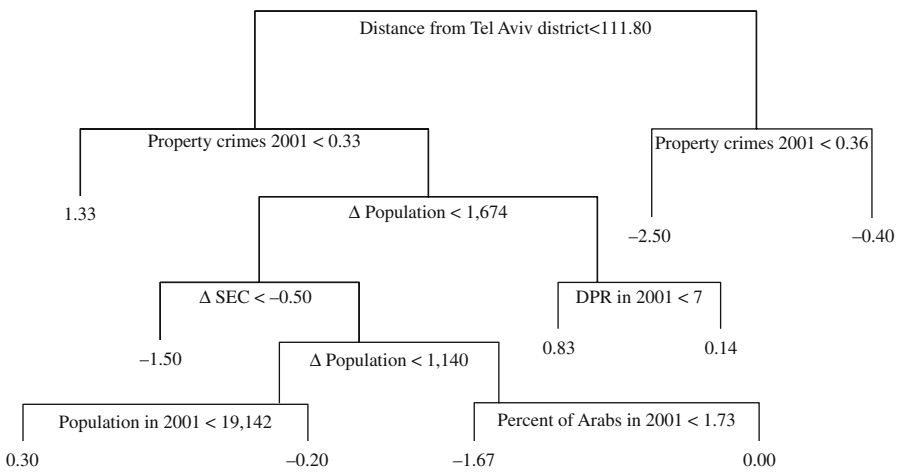


Fig. 1. Regression tree for change in the DPR between 2001 and 2003

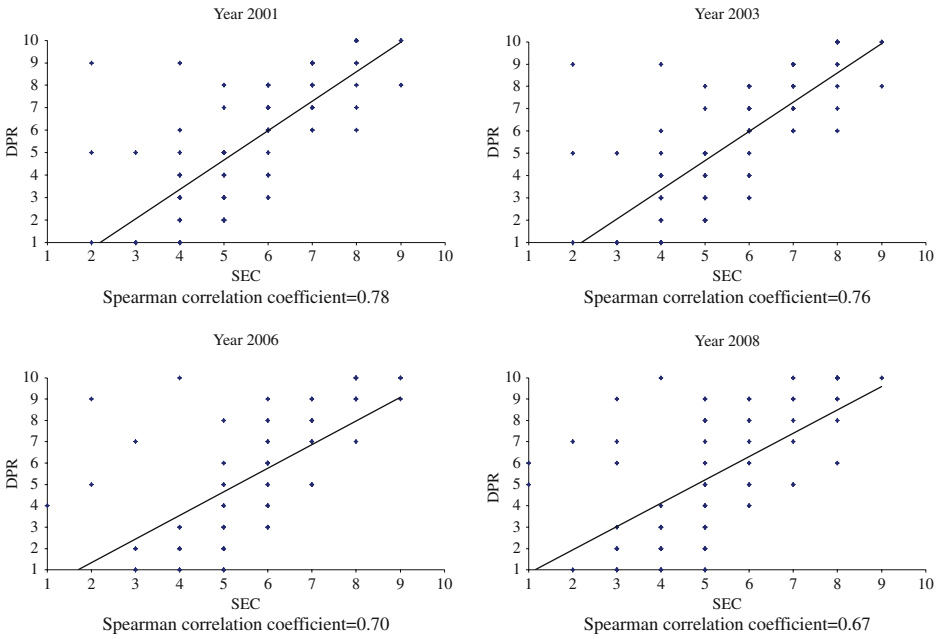


Fig. 2. Socioeconomic cluster vs. dwelling price ranking (Sources: Tax Authority, Central Bureau of Statistics)

and the Shapiro-Wilk and Kolmogorov-Smirnov tests for residual normality) show that at five percent significance level we cannot indicate residual dependence or significant deviation from normality. In addition, results from the correlation analysis do not reveal any evidence of multicollinearity.

The model can be represented as:

$$DPL_k = \alpha + \beta SEC_k + \gamma X_k + \varepsilon \tag{4}$$

In (4), β denotes the regression coefficient for SEC variable in locality k, X_k is a set of explanatory variables for this locality, γ is a vector of coefficients of X_k to be estimated, and ε denotes model residuals with zero expected value and constant variance.

Table 4 shows that most of the effects found to be significant in Model (3) on the SEC variable are also significant in (4) estimated on the DPL and follow the same directions, such as overall effect of crime and the effect of minorities (percentage of Orthodox, percentage of the immigrants from Ethiopia and from the former USSR). The positive and significant estimate of the property crimes variable in 2003 might be explained by "special" correspondence between property crime and dwelling prices, where more expensive residential areas "invite" property crimes. In addition, the negative estimate of the distance from the Tel Aviv district reflects a strong peripherality effect in Israel. However, this influence is nonlinear. A positive sign of the squared term means that the peripherality effect weakens as distance from the center of national economic activity

Table 4. OLS models on "logarithm of the median dwelling price in a locality"

Variable	2001		2003	
	Estimate	*p*-value	Estimate	*p*-value
Intercept	8.819	<0.001	8.647	<0.001
SEC	0.062	<0.001	0.067	<0.001
Percentage of Orthodox population	−0.005	0.003	−0.001	0.443
Percentage of immigrants from the former USSR since 1990	−0.006	0.001	−0.006	0.010
Percentage of Ethiopian immigrants	−0.048	0.000	0.036	0.017
Rate of cases of bodily injury crimes	0.0001	0.975	−0.010	0.096
Rate of cases of property crimes	−0.0006	0.432	0.002	0.052
Distance from Tel Aviv district	−0.005	<0.001	−0.005	<0.001
Distance from Tel Aviv district −squared function	0.00001	0.000	0.00001	0.002
Small locality*	−0.141	0.003	−0.208	<0.001
Total population in a locality (tens of thousands)	0.005	0.029	0.006	0.011
Number of observations		98		107
Adjusted R^2		0.77		0.73

*Localities with 10,000 residents or less (dummy variable)

(Tel Aviv) increases. This occurs due to the existence of additional employment centers in various peripheral towns and the decreasing effect of distance from Tel Aviv in regions that are very far from it.

Estimated Equation (4) for locality k is given by: $DPL_k = \hat{\alpha} + \hat{\beta}SEC_k + \hat{\gamma}X_k$. It follows, that the SEC variable can be expressed as: $SEC_k \approx (DPL_k - \hat{\alpha} - \hat{\gamma}X_k)/\hat{\beta}$. It should be noted that the above approximation for SEC may not be an integer due to continuous characteristics of $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$.

Therefore, this approximation for SEC in a locality k is given by:

$$\overline{SEC}_k = DPR_{corrected} = Round\left[\frac{(DPL_k - \hat{\alpha} - \hat{\gamma}X_k)}{\hat{\beta}}\right] \quad (5)$$

Using (4) and (5), we can examine whether the gap between the SEC and the DPR defined earlier can be at least partly bridged. Table 5 presents the distribution of absolute values of differences between the SEC and the DPR, before and after the correction.

Table 5. Distribution of absolute differences between the SEC and the DPR

	Mean		Min		Max		Percentiles					
							10th		50th		90th	
	2001	2003	2001	2003	2001	2003	2001	2003	2001	2003	2001	2003
Original	1.673	1.776	0	0	7	7	0	0	1	2	3	3
Corrected	0.235	0.234	0	0	1	1	0	0	0	0	1	1

The results presented in Table 5 allow us to conclude that the examined socioeconomic factors that are not currently included in the calculation of the SEC may contribute to a better approximation of the suggested indicator. Furthermore, the earlier obtained Spearman correlation coefficients are also improved to some extent, now being equal to 0.70 and 0.71 for 2001 and 2003, respectively.

5.3. Regression Tree Analysis

In order to draw relevant conclusions on factors influencing the dynamics of the DPR between 2001 and 2003, a nonparametric regression tree was built.

This method of analysis was chosen for the following reasons. First, nonparametric methodology allows for the inclusion of localities in the Arab sector, despite a very small number of available observations and significant differences between the housing markets in Jewish and Arab sectors as mentioned in Section 5. Second, the variables which were removed from the regression models due to multicollinearity can be included in the nonparametric analysis (such as percentage of the Arab population). Given these explanations, the nonparametric regression method is designed to complete and enrich the results obtained from the analysis presented in Subsections 4, 5.1 and 5.2. In this analysis, the dependent variable was defined as the difference between the DPR in 2003 and the DPR in 2001.

To the explanatory variables used in Model (4) we added the differences between the values of these variables in 2003 and 2001. The regression tree method divides observations into homogeneous groups of a dependent variable, given a set of explanatory variables. A detailed description of the regression tree methodology is presented in Breiman et al. (1984).

The algorithm is iterative and works as follows. Initially, from the explanatory variables and their values, the algorithm finds a variable and its values which divides all the observations into two distinct groups, so that the variance of the dependent variable within each group ("leaf") is minimal and the variance between these two groups is maximal (among all possible combinations). This value is fixed as the "split point". The same process is repeated until a specified stopping criterion is fulfilled. At each stage, analysis is performed on the full set of the input variables; therefore, the same explanatory variable can be used several times. This method reveals nonlinear relationships between the dependent and explanatory variables.

The R^2 index was used to test goodness of fit. Let SSW_k denote the estimated variance within the final group k. Since the groups are independent, the total variance within all the groups is calculated by: $SSW = \sum_k SSW_k$. Using the definition of the index R^2, it is given by:

$$R^2 = 1 - \frac{SSW}{SST} \tag{6}$$

where SST denotes the total variance of the dependent variable. The higher the value of the R^2, the better the classification achieved (in terms of the homogeneity of the final "leaves") relative to a previous iteration. In our case, the value of the R^2 index is equal to 0.65.

In Figure 1, the height of the lines between the split points shows the reduction in variance within the group as a result of the division described above, while the numerical value in the final group shows the average increase/decrease in the dependent variable for those leaves. The left branch of each bifurcation corresponds to the "yes" alternative, that is, the condition being fulfilled.

It appears that the dominant factors for change in the DPR are a locality's geographical location and its crime level. For example, in the localities that are situated rather close to the Tel Aviv district (less than 111.8 km) and where the property crime rate was less than 0.33 in 2001, the dwelling price ranking rose by an average of 1.33 (the "leaf" furthest to the left). A decrease in the SEC in a locality caused a consequent decrease in its DPR, given changes in its population and property crime rates in 2001. It also appears that given other controlled variables, an increase in total population is correlated with an increase in the DPR. Additionally, the low percentage of the Arab population in 2001 is correlated with a decrease in the DPR.

6. Conclusions

The current study examined the question of whether dwelling prices in a given locality can serve as an approximation to its socioeconomic level.

The study is based on a number of administrative databases available at a national level. It was found that during the research period (2001 and 2003) there was a strong association between the locality's socioeconomic cluster and the value of its dwellings, with the Spearman correlation coefficients almost identical for these two years.

An analysis of recently obtained SEC data for the years 2006 and 2008 shows that the results obtained for 2001 and 2003 remained consistent, as did the results for 2006 and 2008 (Figure 2). However, a gap has been found between the SEC and the DPR. Our results show that this gap may be explained by location, other social and demographic factors, crime and security characteristics that are exogenous to SEC. In particular, a significant correlation was found between dwelling prices in a specific locality and the percentage of those belonging to defined population groups. It was also found that the size of a locality has a positive correlation with the level of the dwelling prices there. It appears that the effect of the distance from the center of Israel's economic activity is negative, as expected.

It was found that these effects, which reflect other social and demographic characteristics that are not currently included in the SEC calculation, may bridge, at least partly, the gap between the SEC and the DPR.

Overall, we conclude that the ranking based on dwelling prices can serve as a rather good approximation to the socioeconomic level of most urban localities in Israel.

Obviously, this approximation may not always be accurate for some of the localities. Nevertheless, the proposed methodology can provide the required information on socioeconomic profile. This finding is extremely important since the process of SEI and SEC calculation requires a wide variety of variables from various data sources available concurrently for various localities, while the dwelling price ranking allows a rather simple approximation of SEC for different localities for every given year.

For the sake of the methodological consistency and comparability of the SEC series, at the current stage we do not suggest any changes in the set of variables used for the SEC calculations. Rather, we propose the DPR index as an approximation to SEC values for those localities and for years when SEC variables are not available ("intermittent points"). In such cases, using the DPR index, particularly after corrections are made according to Equation (5), can serve as an important working tool for the users of the SEC, such as the Ministry of Finance, Ministry of the Interior, planning authorities and others ensuring a continuum of the index series.

The proposed methodology and the obtained findings are likely to be valid and applicable for different statistical purposes in other countries which possess administrative data on dwelling transactions. For countries that compute socioeconomic indices, the proposed methodology may be used for assessing SEC values for time points and localities for which this index is missing. For countries that do not compute such indices, dwelling price ranking may be used to characterize the socioeconomic profile of a given locality. The suggested approximation may also be used for studying trends in SEC compared with DPR over the years.

Further development and application of an adjustment methodology for SEC imputation is a subject of future research.

Appendix 1. Descriptive Statistics

Name of variable	2001			2003		
	Mean	Std Dev	Median	Mean	Std Dev	Median
Log (DPL)	8.72	0.36	8.64	8.70	0.38	8.68
Total population in a locality	49,287	85.706	21.441	47,299	86,918	21,074
Percentage of Arab population (from the total population in a locality)	11.10	20.56	4.58	10.68	19.98	4.22
Percentage of Orthodox population (from the Jewish population in a locality)	13.56	11.26	11.41	14.90	15.16	11.79
Percentage of ultra-Orthodox population (from the Jewish population in a locality)	13.66	32.20	4.78	12.43	23.09	4.38
Percentage of immigrants from the former USSR since 1990 (from the Jewish population in a locality)	11.73	11.41	7.25	10.86	10.99	6.55
Percentage of Ethiopian immigrants (from the Jewish population in a locality)	0.87	1.47	0.18	0.81	1.43	0.14
Rate of cases of bodily injury crimes (per 1,000 residents)	0.9	0.64	0.78	0.82	0.56	0.69
Rate of cases of property crimes (per 1,000 residents)	40.36	27.40	35.59	38.14	25.81	37.76
Number of terror incidents in a locality	0.43	2.16	0	0.70	3.00	0
Districts: Jerusalem	0.03	0.17	0	0.04	0.18	0
North	0.22	0.42	0	0.20	0.40	0
Haifa	0.15	0.36	0	0.14	0.35	0
Center	0.27	0.45	0	0.28	0.45	0
Tel Aviv	0.11	0.31	0	0.10	0.30	0
South	0.14	0.35	0	0.14	0.35	0
(Dummy variables: 1 if defined district, 0 –otherwise)						
Distance from the Tel Aviv district (km)	55.31	50.76	46.24	54.60	51.12	44.31

Appendix 2. Correlations Between the SEC and the Explanatory Variables

Variables	2001		2003	
	Correlation coefficient	Level of significance	Correlation coefficient	Level of significance
Total population	−0.038	0.702	−0.037	0.702
Percentage of Arab population	−0.445	<0.001	−0.351	0.000
Percentage of Orthodox population	−0.307	0.002	−0.202	0.035
Percentage of ultra-Orthodox population	−0.363	0.000	−0.635	<0.001
Percentage of immigrants from the former USSR since 1990	−0.431	<0.001	−0.359	0.000
Percentage of Ethiopian immigrants	−0.223	0.026	−0.229	0.016
Rate of cases of bodily injury crimes	−0.487	<0.001	−0.434	<0.001
Rate of cases of property crimes	−0.081	0.413	0.055	0.562
Number of terror incidents in a locality	−0.062	0.536	−0.059	0.538
District: Jerusalem	−0.041	0.684	0.014	0.881
North	−0.263	0.007	−0.256	0.006
Haifa	0.071	0.474	0.044	0.647
Center	0.217	0.027	0.233	0.013
Tel Aviv	0.195	0.049	0.181	0.056
South	−0.138	0.163	−0.143	0.134
Distance from the Tel Aviv District	−0.301	0.002	−0.307	0.001

Appendix 3. Validation of OLS Assumptions

3.1. Distribution of the Dependent Variable

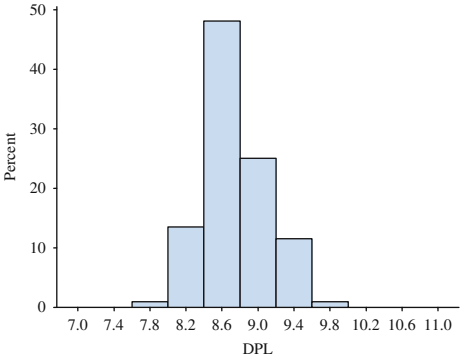


Fig. 1A. Distribution of the DPL 2001

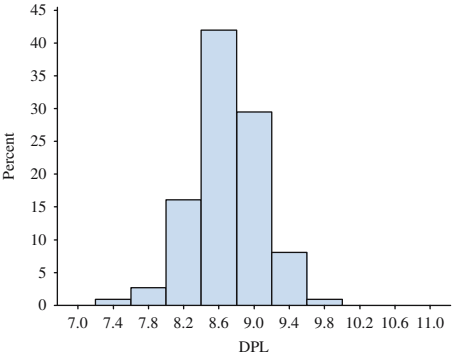


Fig. 2A. Distribution of the DPL 2003

3.2. Residual Normality

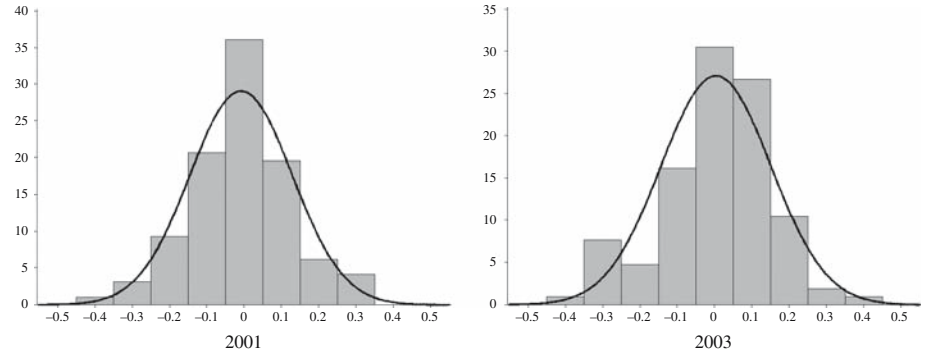


Fig. 3A. Histograms of the model residuals and normal density curves, 2001 and 2003

Table 1A. Test for normality of residuals

	Test (*p*-value)	
Year	Kolmogorov-Smirnov	Shapiro-Wilk
2001	0.66	0.07
2003	0.15	0.10

7. References

Australian Bureau of Statistics. 2006. "An Introduction to Socio-Economic Indexes for Areas (SEIFA)." Information Paper No. 2039.0. Available at: http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/2039.0Main%20Features22006?opendocument&tabname=Summary&prodno=2039.0&issue=2006&num=&view (accessed April 2015).

Blanchard, T.C. 2007. "Conservative Protestant Congregation and Racial Residential Segregation: Evaluating the Closed Community Thesis in Metropolitan and Nonmetropolitan Counties." *American Sociological Review* 72: 416–433. Doi: http://dx.doi.org/10.1177/000312240707200305.

Bourassa, S. C., M. Hoesli, and V.S. Peng. 2003. "Do Housing Submarkets Really Matter." *Journal of Housing Economics* 12: 12–28. Available at: http://www.sciencedirect.com/science/article/pii/S1051137703000032 (accessed April 2015).

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and Regression Trees*. Wadsworth: Belmont.

Buck, A.J., J. Deutsch, J. Hakim, U. Spiegel, and J. Weinblatt. 1991. "A Von Thunen Model of Crime, Casinos and Property Values in New Jersey." *Urban Studies* 28: 673–686. Doi: http://dx.doi.org/10.1080/00420989120080861.

Burck, L. and Y. Kababia. 1996. *Characterization and Ranking of Local Authorities according to the Socio-Economic Level of the Population in 1995*. Publication No. 1039, Central Bureau of Statistics, Jerusalem, (Hebrew).

Burck, L. and Y. Kababia. 1999. *Characterization and Ranking of Local Authorities according to the Socio-Economic Level of the Population in 1999, Based on the 1995 Census of Population and Housing*. Publication No. 1118, Central Bureau of Statistics, Jerusalem. (Hebrew).

Cahaner, L. 2012. "Expansion Processes of the Jewish ultra-Orthodox Population in Haifa." In *Themes in Israel Geography, Special Issue of Horizons in Geography*, edited by J.O. Maos and I. Charney, 70–87, University of Haifa.

Central Bureau of Statistics (CBS). 2000. Characterization and Classification of Geographical Units by the Socio-Economic Level of the Population. 1995 Census of Population and Housing Publications, No. 13, Jerusalem.

Cummings, J. L., D. DiPasquale, and M. E Kahn. 2002. "Measuring the Consequences of Promoting Inner City Homeownership." *Journal of Housing Economics* 11: 330–359. Available at: http://www.cityresearch.com/pubs/dipasquale.pdf (accessed April 2015).

Des Rosiers, F., M. Theriault, Y. Kestens, and P. Villeneuve. 2002. "Landscaping and House Values: An Empirical Investigation." *Journal of Real Estate Research* 23: 139–161.

Dubin, R.A. and A.C. Goodman. 1982. "Valuation of Education and Crime Neighborhood Characteristics Through Hedonic Housing Price." *Population and Environment* 5: 166–181. Doi: http://dx.doi.org/10.1007/BF01257055.

Goodman, A.C. and T.G. Thibodeau. 1998. "Housing Market Segmentation." *Journal of Housing Economics* 7: 121–143. Doi: http://dx.doi.org/10.1006/jhec.1998.0229.

Greenberg, M.R. 1999. "Improving Neighborhood Quality: A Hierarchy of Needs." *Housing Policy Debate* 20: 601–624. Doi: http://dx.doi.org/10.1080/10511482.1999.9521345.

Harris, D.R. 1999. "Property Values Drop When Blacks Move in, Because. . .: Racial and Socioeconomic Determinants of Neighborhood Desirability." *American Sociological Review* 64: 461–479.

Hazam, S. and D. Felsenstein. 2007. "Terror, Fear and Behaviour in the Jerusalem Housing Market." *Urban Studies* 44: 2529–2546. Doi: http://dx.doi.org/10.1080/00420980701558392.

Heikkila, E. 1992. "Describing Urban Structure: A Factor Analysis of Los Angeles." *Review of Urban and Regional Development Studies* 4: 84–101. Doi: http://dx.doi.org/10.1111/j.1467-940X.1992.tb00035.x.

Horton, H.D. and M.E. Thomas. 1998. "Race, Class, and Family Structure: Differences in Housing Values for Black and White Homeowners." *Sociological Inquiry* 68: 114–136. Doi: http://dx.doi.org/10.1111/j.1475-682X.1998.tb00456.x.

Jackson, E., V. Kupke, and P. Rossini. 2007. The Relationship between socio-economic indicators and residential property values in Darwin. Paper presented at the 13th Annual Pacific-Rim Real Estate Society Conference. Fremantle, Western Australia. Available at: http://scholar.google.com.au/citations?view_op=view_citation&hl=en&user=KbV4jccAAAAJ&citation_for_view=KbV4jccAAAAJ:3fE2CS-JIrl8C (accessed April 2015).

Jolliffe, I.T. 2002. *Principal Component Analysis*, 2nd ed. Springer Series in Statistics. New York: Springer.

Kiel, K.A. and J.E. Zabel. 1996. "House Price Differentials in U.S. Cities: Households and Neighborhood Racial Effects." *Journal of Housing Economics* 5: 143–165. Doi: http://dx.doi.org/10.1006/jhec.1996.0008.

Krivo, L.J. 1995. "Immigrant Characteristics and Hispanic-Anglo Housing Inequality." *Demography* 32: 599–615. Doi: http://dx.doi.org/10.2307/2061677.

Lewin-Epstein, N., Y. Elmelech, and M. Semyonov. 1997. "Ethnic Inequality in Home Ownership and the Value of Housing: The Case of Immigrants in Israel." *Social Forces* 75: 1439–1462. Doi: http://dx.doi.org/10.1093/sf/75.4.1439.

Lozano-Gracia, N. and L. Anselin. 2012. "Is the Price Right? Assessing Estimates of Cadastral Values for Bogota, Colombia." *Regional Science Policy & Practice* 4: 495–508. Doi: http://dx.doi.org/10.1111/j.1757-7802.2012.01062.x.

Malpezzi, S., G.H. Chun, and R.K. Green. 1998. "New Place-to Place Housing Price Indexes for U.S. Metropolitan Areas, and Their Determinants." *Real Estate Economics* 26: 235–274. Doi: http://dx.doi.org/10.1111/1540-6229.00745.

McCluskey, W. J., W. G. Deddis, I. G. Lamont, and R. A. Borst. 2000. "The Application of Surface Generated Interpolation Models for the Prediction of Residential Property values." *Journal of Property Investment & Finance* 18: 162–176. Available at: http://eprints.ulster.ac.uk/10588/ (accessed April 2015).

Myers, D. 1990. *Housing Demography: Linking Demographics Structure and Housing Markets*. Madison: University of Wisconsin Press.

Olausson, P.O. and D. Vagero. 1991. "Miscellanea, A Swedish Classification Into Social Classes Based on Census Information and Comparable to The British Classification–A Proposal." *Journal of Official Statistics* 7: 93–103.

Ozanne, L. and T. Thibodeau. 1983. "Explaining Metropolitan Housing Price Differences." *Journal of Urban Economics* 13: 51–66. Doi: http://dx.doi.org/10.1016/0094-1190(83)90045-1.

Potepan, M.J. 1996. "Explaining Intermetropolitan Variation in Housing Prices, Rents and Land Prices." *Real Estate Economics* 24: 219–245. Doi: http://dx.doi.org/10.1111/1540-6229.00688.

Reed, R. 2001. "The Significance of Social Influences and Established Housing Values." *Appraisal Journal, October* 1.

Rose, D. and D. Pevalin. 2001. *The National Statistics Socio-Economic Classification: Unifying Official and Sociological Approaches to the Conceptualization and Measurement of Social Classes*. Institute of Social and Economic Research Working Papers, November 2001–4. Available at: https://www.iser.essex.ac.uk/files/iser_working_papers/2001-04.pdf (accessed April 2015).

Soffer, A. and E. Bystrov. 2006. *Tel Aviv State: A Threat to Israel*. Haifa, Reuven Chaikin Chair in Geostrategy, University of Haifa. Available at: http://geo.haifa.ac.il/ ~ ch-strategy/publications/english/Tel_Aviv_State.pdf (accessed April 2015).

Thaler, R. 1978. "A Note on the Value of Crime Control: Evidence from the Property Market." *Journal of Urban Economics* 5: 137–145. Doi: http://dx.doi.org/10.1016/0094-1190(78)90042-6.

Yates, J. 2002. *A Spatial Analysis of Trends in Housing Markets and Changing Patterns of Household Structure and Income*. Positioning Paper 30 of the Australian Housing and Urban Research Institute. Sydney Research Centre. Available at: http://www.ahuri.edu.au/publications/download/ahuri_60064_fr (accessed April 2015).

Received July 2013
Revised August 2014
Accepted October 2015

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 205–229, http://dx.doi.org/10.1515/JOS-2015-0014

Dwelling Price Ranking versus Socioeconomic Clustering: Possibility of Imputation

Larisa Fleishman[1], *Yury Gubman*[1], *and Aviad Tur-Sinai*[1]

In order to characterize the socioeconomic profile of various geographic units, it is common practice to use aggregated indices. However, the process of calculating such indices requires a wide variety of variables from various data sources available concurrently. Using a number of administrative databases for 2001 and 2003, this study examines the question of whether dwelling prices in a given locality can serve as a proxy for its socioeconomic level. Based on statistical and geographic criteria, we developed a Dwelling Price Ranking (DPR) methodology. Our findings show that the DPR can serve as a good approximation for the socioeconomic cluster (SEC) calculated by the Israel Central Bureau of Statistics for years when the required data was available. As opposed to the SEC, the suggested DPR indicator can easily be calculated, thus ensuring a continuum of socioeconomic index series. Both parametric and nonparametric statistical analyses have been carried out in order to examine the additional social, demographic, location, crime and security effects that are exogenous to SEC. Complementary analysis on recently published SEC series for 2006 and 2008 show that our conclusions remain valid. The proposed methodology and the obtained findings may be applicable for different statistical purposes in other countries which possess dwelling transactions data.

Key words: Housing market; urban locality; index construction.

1. Introduction

The socioeconomic profile of a residential area can be identified and characterized in two different ways: by using specific demographic and socioeconomic factors, or by estimating aggregated indices based on a range of these factors. In Israel, a socioeconomic index (SEI) for local authorities was developed at the Central Bureau of Statistics (CBS) in the mid-1990s. The main and most important use of the data on SEI is its contribution to the design and implementation of different policies in various ministries and other governmental agencies relating to local authorities, including various resource allocation procedures. However, the main limitation in using SEI is that the SEI data is not available for all localities every year, since the process of SEI calculation requires a wide variety of variables from various data sources available concurrently for various localities.

In previous studies, there is well-documented evidence of the relationship between the various demographic and socioeconomic characteristics of a locality and its dwelling prices. In Israel, the main source of information about dwelling prices is the record of real estate transactions kept by the Israel Tax Authority (ITA). The Israel CBS has been receiving this database on a monthly basis since 1990. Using this data source, our study

[1] Israeli Central Bureau of Statistics, 66 Kanfei Nesharin St. Jerusalem 91342, Israel. Emails: larisaf@cbs.gov.il, yuryg@cbs.gov.il, and aviadts@cbs.gov.il

proposes a method for assessing the socioeconomic level of those localities and for those time points for which these data are missing.

The current study has two main goals: 1) to examine if dwelling prices can serve as a proxy for the socioeconomic level of various urban localities; 2) to examine the extent of correlation between social and demographic characteristics not included in the SEI calculations and the level of dwelling prices in a given locality.

In order to achieve the first goal, relevant urban localities in Israel were graded according to their dwelling price level, and this was compared to their socioeconomic level. A similar analysis was conducted for the years 2006 and 2008. To achieve the second goal, parametric models were estimated. In this way conclusions were drawn regarding the possibilities for imputation of missing data on SEI values.

This study augments the relevant research literature in several ways. First, previous studies have focused primarily on the effect of demographic and socioeconomic factors on the level of dwelling prices in specific cities or metropolitan areas. We extended this investigation to a national level. Furthermore, it should be noted that Israel is a highly urbanized country with more than ninety percent of Israelis living in urban areas. Second, our investigation focused on the correlation between socioeconomic indices and dwelling price indicators. Only limited research has been carried out on the degree of correspondence between these two indicators. Third, this study proposes, examines and discusses a reasonable alternative method for assessing the socioeconomic level of urban localities. This method does not require a range of multivariate procedures and various statistical data available concurrently, although it allows for a sound approximation to a socioeconomic level. Research thus far has not addressed this issue.

The rest of this article is organized as follows: Section 2 presents some theoretical background and a survey of relevant literature, and explains the construction of the socioeconomic index in Israel. Section 3 describes the main source of data, the construction of the key variable reflecting dwelling price level, and defines geographic units for which this variable was created. Section 4 presents and discusses the degree of correspondence between the dwelling price level and socioeconomic level, describes additional sources of information and defines the variables used in the analysis. Section 5 presents the statistical models that we used for the empirical analysis, gives the results and discusses them. Section 6 concludes the study.

2. Literature Review

2.1. Dwelling Price in Light of Socioeconomic Level Characteristics

The price of a residential property on the free market reflects the willingness of the purchaser to pay not only for the property itself, but also for a specific residential environment – in other words, for the quality of "social space" (Reed 2001).

Research findings from around the world testify to the range of factors which reflect the socioeconomic essence of a residential area, the most important of which are income, education, employment, and the demographic characteristics of its population. Earlier studies show a significant positive correlation between the three major factors characterizing the residential area's socioeconomic level – income, education and

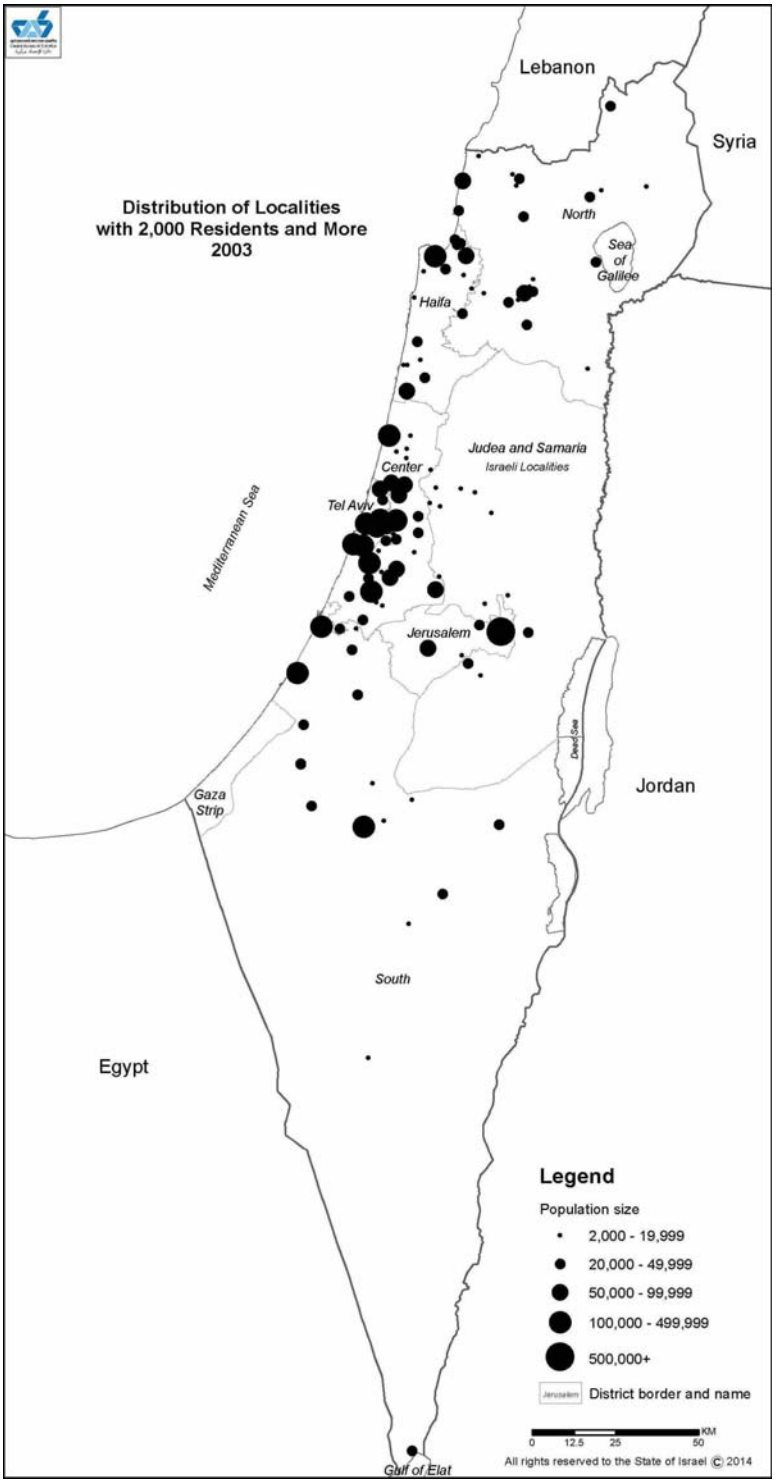
employment – and the price of dwellings (Heikkila 1992; Potepan 1996; Goodman and Thibodeau 1998; Greenberg 1999; Des Rosiers et al. 2002; Yates 2002). Income is considered the primary factor (Ozanne and Thibodeau 1983; Malpezzi et al. 1998). Those with relatively high incomes choose their residential area in an attempt to avoid neighbors with a low socioeconomic status. The popular viewpoint considers social problems, such as crime, drug use, and the neighborhood's economic decline resulting in neglected buildings, as all directly linked to neighborhoods characterized by a high proportion of unemployed and low levels of education and income (Harris 1999; Jackson et al. 2007). The study of Cummings et al. (2002) examined education as one of the dimensions of the socioeconomic level of residents of various neighborhoods in the city of Philadelphia and its influence on the price of residential dwellings in those neighborhoods. The study findings show a 21 percent increase in dwelling prices with every ten percent rise in the proportion of adults with post-high school education.

Aside from the aforementioned factors that characterize the socioeconomic space and impact dwelling price, the relevant literature has examined the relationship between the demographic characteristics of residents, such as age and marital status, and dwelling prices in that neighborhood (Myers 1990; Heikkila 1992). There are studies indicating that ethnic composition and personal security in a residential environment may also contribute to the socioeconomic space, and as a result affect the price of dwellings. In particular, previous studies provide sound evidence of a strong positive correlation between the level of personal security in a residential area and its dwelling prices (Thaler 1978; Dubin and Goodman 1982; Buck et al. 1991; Hazam and Felsenstein 2007).

Earlier studies have also addressed the effect of immigrant groups and the racial-ethnic context of residential areas on the local housing prices (Kiel and Zabel 1996). However, there is no common agreement on either the existence or the magnitude of the effect of immigration shocks on the housing market. The magnitude of the effect immigrants have on housing prices depends heavily on the reaction of natives to the presence of immigrants in the area. For example, some studies in the USA have found that blacks and Hispanics own homes of lesser value than the white population. This held true even when the researchers controlled for the characteristics of dwellings (Horton and Thomas 1998; Krivo 1995; Lewin-Epstein et al. 1997). Harris (1999) found that dwelling prices in the USA decline by an average of 16 percent when the Afro-American population exceeds ten percent in a neighborhood. Furthermore, a far more dramatic drop in prices occurs when the Afro-American population exceeds 60 percent of the neighborhood residents. According to this study, the explanation is not necessarily ethnic preference, but may be related to social problems stemming from the socioeconomic status of the Afro-American population, which is usually lower. However, a study conducted in Darwin (Australia) by Jackson et al. (2007) revealed high positive correlation between housing prices and ethnicity for people born overseas and for those who speak other languages.

Along with the abovementioned demographic and socioeconomic characteristics of a population, religiosity contributes substantially to the residential profile of a locality (Blanchard 2007). Among several religiosity patterns displayed by the Israeli Jewish population, both ultra-Orthodox and Orthodox streams play a significant role with regard to the socioeconomic, ethnic and spatial divide in Israeli society, thus notably effecting local dwelling market price level (Cahaner 2012).

Map 1. Distribution of Urban Localities in Israel



In addition, the importance of the geographical location of dwellings in the context of their price level has been emphasized in several studies (McCluskey et al. 2000; Bourassa et al. 2003). In Israel this issue is of special importance. According to the official administrative division, there are six main administrative districts in Israel. Map 1 shows the distribution of urban localities as well as administrative districts. Tel Aviv district is composed of the central city of Tel Aviv and other cities enclose it on three sides. This district is the geographical center of Israel, and is characterized by a very highly concentrated urban population. It is the financial, economic, social, and cultural center of the country.

Tel Aviv district is dominant and influential in the domains of employment and communication, as well as the domain of land and dwellings prices (Soffer and Bystrov 2006). Therefore, the geographical proximity of a locality to the center, or to the periphery, affects many aspects of life in a locality, including the socioeconomic level of the population and dwelling price level. Summarizing this short literature review, it can be concluded that a sizeable body of literature provides evidence of a strong association between various demographic, social and economic characteristics of a locality and the price of its dwellings . This evidence serves as a theoretical foundation for the premise behind this study: the price of dwellings in a specific locality can serve as an alternate measure of its socioeconomic level.

2.2. *The Socioeconomic Index and the Socioeconomic Cluster*

In order to characterize and document the socioeconomic profile of various localities, it is common practice in the official statistics of various countries to use aggregated indices (Burck and Kababia 1996, 1999; Australian Board of Statistics 2006). These indices are based on different theoretical assumptions and estimation methods, and may be classified in accordance with two main approaches. Using a "deterministic" approach, a socio-economic cluster is specified by applying predefined classification criteria based on an underlying conceptual model. For instance, Rose and Prevalin (2001) suggest the set of employment status and occupation variables for socioeconomic classification in the UK, following the social class classification methodology earlier suggested by Olausson and Vagero (1991) for Swedish register data.

The "stochastic" approach assumes the existence of a latent continuous variable (Y) for the socioeconomic level of a given locality. It is also assumed that Y may be assessed using multivariate analysis methodology on a set of observed variables. Finally, localities are clustered by the estimated values of Y. For example, Jackson et al. (2007) suggest estimating socioeconomic level using principal component methodology applied to a wide set of socioeconomic characteristics which includes income, age, family status, dwelling data, and so on. Generally, Principal Component Analysis (PCA) is a technique that is useful for the compression and classification of data. The main idea of PCA is to reduce the dimensionality of a data set which contains a large number of interrelated variables. This reduction is achieved by finding a new set of uncorrelated variables (the principal components) smaller than the original set of variables that nonetheless retains most of the variation present in the original data set (Jolliffe 2002).

The stochastic approach is widely used for socioeconomic index calculation in the official statistics of different countries. The Office for National Statistics in the UK devises the

socioeconomic index for areas within local authorities by means of principal component analysis based on population censuses. The Australian Bureau of Statistics produces five socioeconomic indices that measure various socioeconomic aspects of residential areas based on the population census. Surveys are used for updating the index in the periods between population censuses. A similar methodology is used in New Zealand.

In Israel, a socioeconomic index (SEI) was developed at the Central Bureau of Statistics (CBS) in the mid-1990s on the basis of the 1995 Population and Housing Census data (Burck and Kababia 1996). It is based on five groups of variables that include 14 variables. The variables used to construct the index reflect all of the aspects related to the socioeconomic makeup of the population of different localities, subject to the availability of the data (for more details on the selection of the variables, see CBS 2000). These five groups contain the following variables: (1) demographic characteristics (dependency ratio, median age, percentage of families with four or more children); (2) education and schooling (percentage of the students studying for a bachelor's or higher degree, percentage eligible for a matriculation certificate); (3) standard of living (level of motorization, percentage of new motor vehicles, average income per capita); (4) labor force statistics (percentage of job seekers, percentage of salaried workers and self-employed persons earning up to minimum wage, percentage of salaried workers earning more than twice the average salary); (5) support/pension (percentage receiving unemployment benefits, percentage receiving income supplements; percentage receiving old age pensions with income supplements). SEI is based on the stochastic approach and calculated using principal component analysis.

Principal components (factors) are essentially new variables, calculated as a linear combination (weighted average) of the original, standardized variables (i.e., each variable has a mean of 0 and variance of 1). The weights of the original standardized variables are determined mathematically so as to maximize the differences in the scores between the geographical units, subject to some normalization restrictions. The factors are determined sequentially, so that the first factor is the linear combination that accounts for the maximum amount of the variance of the variables. Hence, the first factor has the greatest ability to discern between the localities. The second factor accounts for a maximum variance not accounted for by the first factor, and so on. The optimal number of factors that should be used to explain the maximum amount of the variance of the variables is determined by statistical testing. It is noteworthy that since the variables are standardized, the total variance of the original variables is equal to the number of variables. These factors define an orthogonal set of axes in the multidimensional variable space where each factor is a linear combination of the original variables. This type of factor analysis can be defined as PCA (CBS 2000).

We can represent the socioeconomic index as follows:

$$SEI = a_1X_1 + a_2X_2 + ... + a_{14}X_{14} \quad (1)$$

with *SEI* indicating the socioeconomic index (continuous), $a_1, \ldots, a_{14}$ the coefficients calculated using principal component analysis, and $X_1, \ldots, X_{14}$ the variables constituting the index which were specified above.

The set $X_1, \ldots, X_{14}$ has been defined based on the methodological background identified in the relevant literature, while considering local conditions and data

availability. It should be noted that, as opposed to Jackson et al. (2007), the dwelling prices are not included in the SEI calculations.

The SEI estimates cover most local authorities for which all the variables listed above are available at the relevant time point. For the sake of consistency and comparability of the SEI series, the set of variables and the calculation methodology have not been changed over the years.

Using cluster analysis, the local authorities, for which the SEI is calculated, are then divided into ten socioeconomic clusters (SEC), with Cluster 1 including authorities with the lowest socioeconomic level, and Cluster 10 including authorities with the highest socioeconomic level.

After 1995, the SEI was updated for the years 1999, 2001, 2003, and 2006 when the required data was available. The SEI calculated in 2008 on the basis of the 2008 Population Census is the most recent. As dwelling price data are available annually, we suggest a univariate deterministic approach for approximation of the socioeconomic level of a given locality at a given time.

In this context, it is worth noting that the relevant literature is mostly dedicated to the correlation between the different demographic and socioeconomic characteristics of a locality and the dwelling prices in it, and the examination of the degree of correspondence between aggregated socioeconomic indices and dwelling prices remains beyond the scope of research. Thus, to reach a conclusion as to the ability of dwelling prices to serve as a proxy for the socioeconomic level, we first need to check the degree of correspondence between them. Afterwards, we examine certain additional social and demographic characteristics that are not currently included in the SEI calculations, but are known to affect dwelling prices.

3. Data and Definitions

The study is based on files of dwelling transactions in the housing market in 2001 and 2003. Transaction data are provided annually by the Israel Tax Authority. In total, the basic file from 2001 included 60,851 transactions, and the file from 2003 included 57,223 transactions.

In order to compare the SEC of a given locality with its aggregate dwelling price level, the same coding scheme had to be used for both indicators. Dwelling Price Ranking (DPR) was constructed based on the following steps. First, using the transactions data, dwelling price level (DPL) for each relevant locality k was calculated:

$$DPL_k = log(median(Y_k)), \tag{2}$$

where Y denotes price per square meter. Using price per square meter as an underlying variable for DPR, we neutralize the effect of apartment size, one of the main variables which explains differences in dwelling prices (Lozano-Gracia and Anselin 2012), and represents as far as possible the market value of a dwelling at the aggregate level for a given locality. The median is used for reasons of robustness, and the log-transformation stabilizes the variance and generally makes the data normally distributed.

Second, localities for which the DPR was created were selected using the following criteria: (1) total population of 2,000 or more in locality, which corresponds to the

definition of "urban" in Israel; (2) the number of transactions in a locality should be sufficient enough to represent the price level in the housing market (at least 15 in the current study). Localities that did not match the above criteria were excluded from the analysis. The final data set includes 104 localities in 2001 and 112 localities in 2003, covering about 90 percent of the Israeli population.

Third, the selected localities were divided into ten clusters, alongside the SEC. Localities with the lowest DPL were ranked as Level 1 (DPR $= 1$), while the localities with the highest dwelling prices were ranked as Level 10 (DPR $= 10$). Each of the resulting DPR clusters contains approximately the same number of localities.

4. Dwelling Price Ranking vs. Socioeconomic Cluster

In order to examine the degree of correspondence between the SEC and the DPR, a correlation analysis was carried out. The obtained Spearman correlation coefficients are equal to 0.69 and 0.67 for 2001 and 2003, respectively.

Table 1 presents the detailed results for 2003; the analysis for 2001 revealed similar results. In Table 1, the digit in each cell indicates the number of localities with DPR and the SEC as they appear in the rows and columns, respectively. The cases in which both rankings are identical appear in bold print; there is an exact correspondence between the SEC and the DPR for 21 localities (out of 112).

Based on these results, it can be concluded that a lack of correspondence between the SEC and the DPR is more typical for localities where the SEC is low or low-medium. The minimal gap between the SEC and the DPR (± 1 range) is observed for 28 percent of the localities. For localities where the gap between the SEC and the DPR is greater than 2, it was found that localities with a DPR higher than their SEC are mainly situated close to the Tel Aviv district, that is, close to the center of the country. Those are localities where the SEC is medium-high to high (6–9). For localities with a low-medium SEC or medium SEC (3–5), in most cases the DPR was lower than the SEC. Those localities are located in the more peripheral areas.

Table 1. The socioeconomic cluster vs. the dwelling price ranking 2003

	The socioeconomic cluster										
DPR	1	2	3	4	5	6	7	8	9	10	Total
1	–	1	2	6	1	1	–	–	–	–	11
2	–	–	–	3	7	–	1	–	1	–	12
3	–	1	–	4	4	2	–	–	–	–	11
4	–	–	2	**4**	2	1	–	1	1	–	11
5	1	1	–	–	**5**	4	–	1	–	–	12
6	–	1	–	1	1	**3**	4	1	–	–	11
7	–	–	–	–	1	2	**5**	1	–	–	9
8	–	–	–	–	1	2	3	**3**	2	1	12
9	–	1	–	1	–	1	5	4	–	–	12
10	–	–	–	–	–	–	–	9	1	**1**	11
Total	1	5	4	19	22	16	18	20	5	2	112

These findings indicate the spatial aspects contained in the correlation between the SEC and the DPR. In particular, it can be concluded that the dependence of dwelling prices on the distance from the Tel Aviv district is stronger than the spatial dependence for the SEC. That is, it is rare to find very expensive dwellings in the peripheral regions, while there are some peripheral localities with a comparatively high socio-economic profile.

It can also be seen that the degree of correspondence between the two indicators increases with the rise in the SEC scale of the localities. Maps 2 and 3 illustrate spatial distribution of the localities according to the SEC and the DPR for 2003.

We conclude that the suggested dwelling price ranking appears to be a sufficiently good approximation for the socioeconomic cluster. However, a gap is revealed between two indicators, and it is therefore reasonable to assume that there are other factors influencing dwelling prices. Using these factors, we attempted to correct the developed dwelling price ranking by reducing the gap between the DPR and the SEC.

In order to examine additional factors that are not included in the SEC calculation but are assumed to influence dwelling prices, the following administrative databases are used. First, the Population Registry from which information on population characteristics by locality was obtained (e.g., percentage of immigrants from the former USSR and Ethiopia in 2001 and 2003). Second, the "Level of Religiosity" administrative file that was developed at the CBS serves as a basis for such variables as the percentage of ultra-Orthodox and Orthodox population by locality. Third, a crime database was provided by the Israeli Police. Finally, terror incidents data for relevant years was created by using information from different sources available from the International Institute for Counter-Terrorism (ICT) at the Interdisciplinary Center (IDC) Herzliya, Ministry of Foreign Affairs and the Prime Minister's Office. Additionally, spatial information regarding the location of the localities relative to the center of the country was provided by the Geographic Information System (GIS).

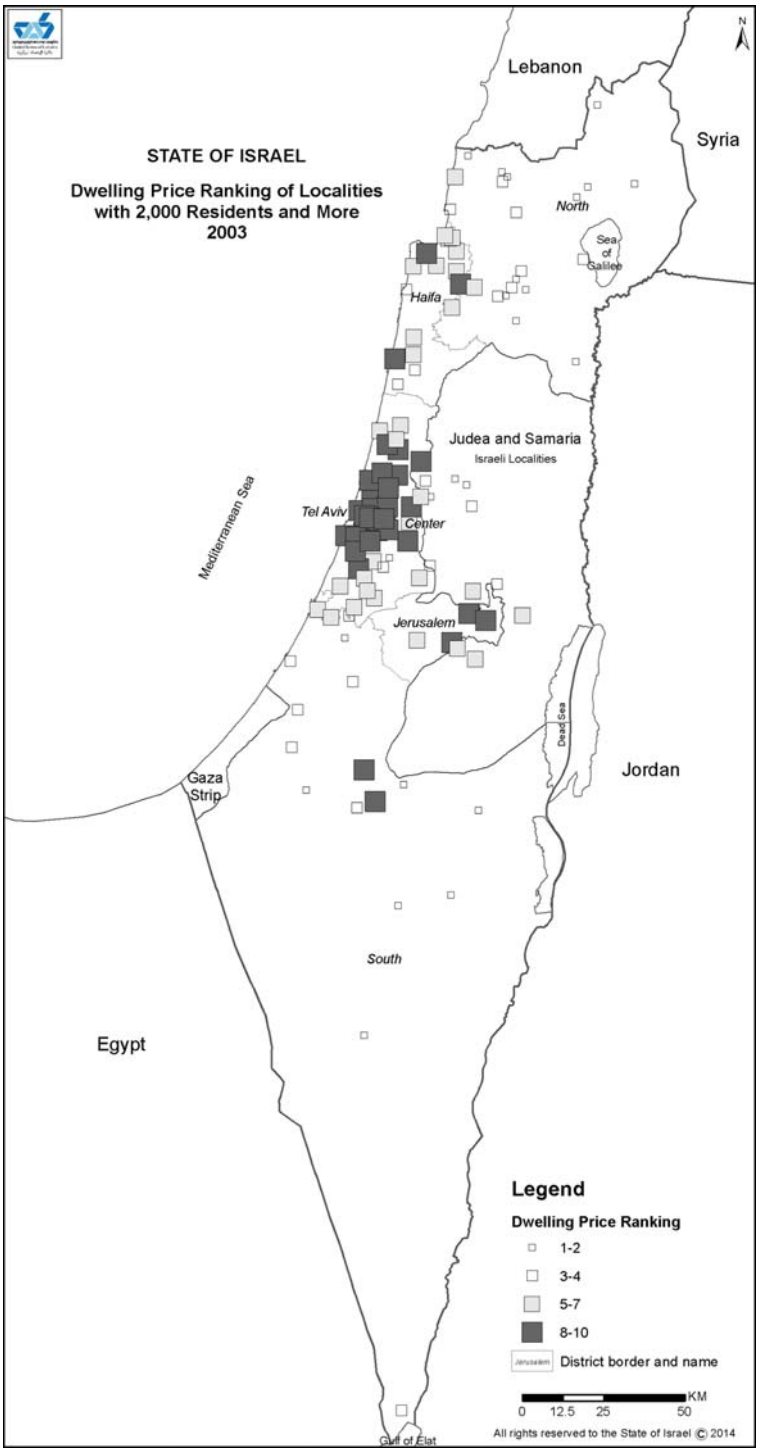
On the basis of these databases, a set of explanatory variables were selected based on the existing literature in this field partly reviewed in Subsection 2.1. Appendix 1 presents and defines the variables that we used in the study, their means, standard deviations and medians.

In order to examine the degree of correlation between the DPL and the selected variables, a correlation analysis was carried out. The Spearman correlation coefficients are presented in Table 2. Note that the Spearman coefficient is used since the distribution of most explanatory variables is skewed.

Of all the variables having a significant correlation with the DPL, there are six variables characterized by a positive correlation with the DPR: the total population in a locality, the rate of cases of property crimes, the number of terror incidents and the three variables indicating the geographic district of a locality – Jerusalem District, Center District and Tel Aviv District.

The degree of correlation between the SEC and the selected variables was also examined. The results of this test are presented in Appendix 2. The correlation analysis on both the DPL (Table 2) and the SEC variables (Appendix 2) revealed similar results.

Map 2. Dwelling Price Ranking of Localities



Map 3. Socioeconomic Cluster of Localities

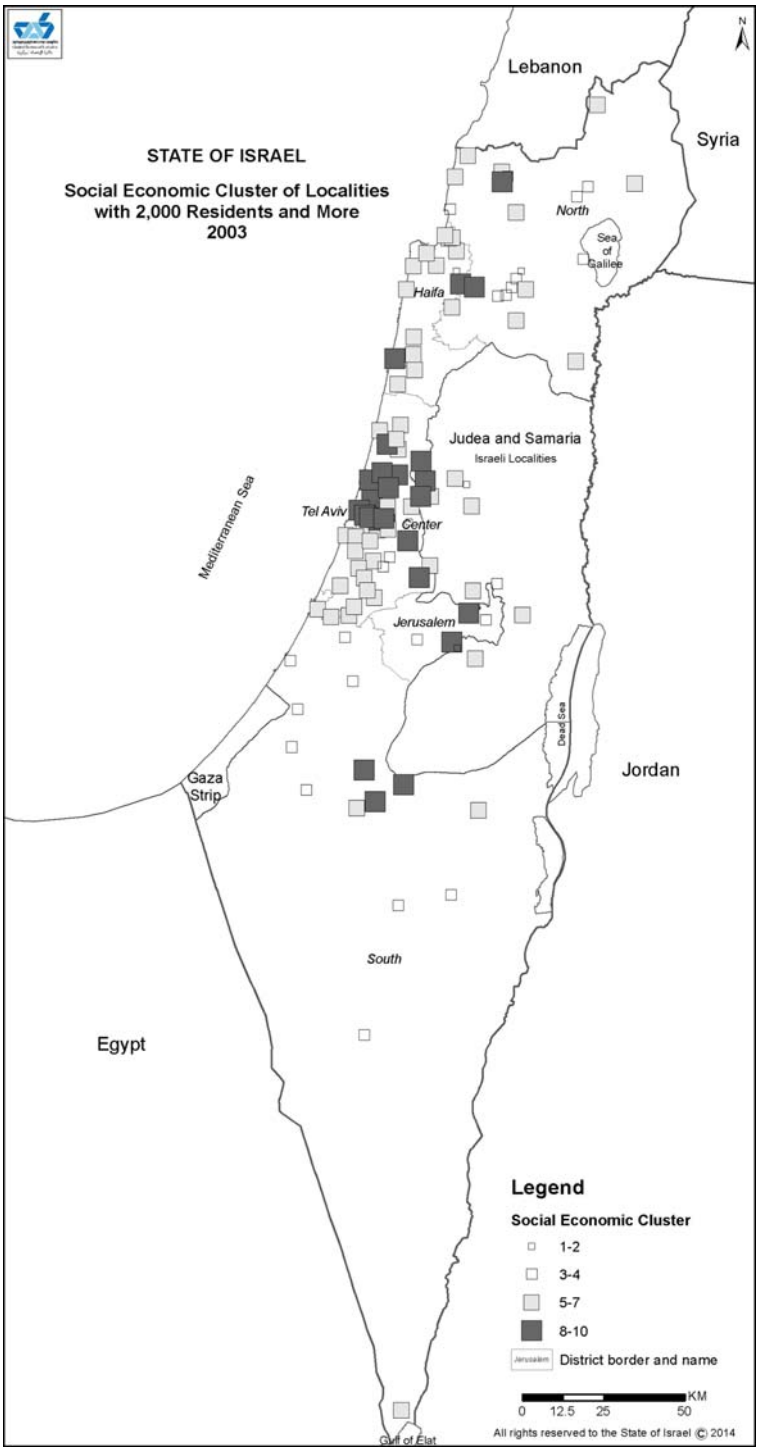


Table 2. Correlations between the DPL and the explanatory variables

Variables	2001		2003	
	Correlation coefficient	Level of significance	Correlation coefficient	Level of significance
Total population	0.169	0.086	0.264	0.005
Percentage of Arab population	−0.493	<0.001	−0.418	<0.001
Percentage of Orthodox population	−0.307	0.002	−0.340	<0.001
Percentage of ultra-Orthodox population	−0.307	0.002	−0.265	0.005
Percentage of immigrants from the former USSR since 1990	−0.423	<0.001	−0.379	<0.001
Percentage of Ethiopian immigrants	−0.205	0.040	−0.068	0.481
Rate of cases of bodily injury crimes	−0.453	<0.001	−0.459	<0.001
Rate of cases of property crimes	0.294	0.002	0.292	0.002
Number of terror incidents	0.204	0.038	0.132	0.166
Districts: Jerusalem	0.124	0.211	0.174	0.066
North	−0.446	<0.001	−0.473	<0.001
Haifa	−0.014	0.885	0.018	0.849
Center	0.346	<0.001	0.410	<0.001
Tel Aviv	0.449	<0.001	0.440	<0.001
South	−0.353	0.001	−0.315	<0.001
Distance from the Tel Aviv district	−0.695	<0.001	−0.721	<0.001

5. Parametric Models and Findings

Two parametric models were estimated: multinomial logistic regression for the SEC variable and the OLS model for the DPL. The models were estimated only for the Jewish sector for the following reasons. The housing market in the Arab sector operates under different conditions than that in the Jewish sector, and some of the explanatory variables are irrelevant to the Arab sector (such as the percentage of new immigrants from the former USSR and Ethiopia). Furthermore, the number of localities in the Arab sector with sufficient number of transactions was inadequate for performing statistical analyses for the Arab sector solely (four localities).

In order to avoid possible multicollinearity, those explanatory variables that were found to be highly correlated with other explanatory variables (Pearson correlation coefficient is more than 0.5) were excluded from the parametric models. These variables were included in the nonparametric analysis presented in Subsection 5.3.

5.1. A Multinomial Logistic Model for the SEC Variable

To estimate the marginal contribution of each of the above factors to the SEC, a regression analysis was carried out. Since the SEC is categorical, a multinomial logistic regression model was estimated, with the dependent variable being the probability of being in cluster i:

$$P(SEC = i) = log\,it(\alpha_i + \beta DPR + \gamma X) \quad (3)$$

Table 3. Multinomial models for the socioeconomic cluster

Variable	2001			2003		
	Estimate	*p*-value	Odds ratio	Estimate	*p*-value	Odds ratio
Intercept for ranking = 10	−6.16	0.000	–	−5.22	0.000	–
Intercept for ranking = 9	−4.05	0.004	–	−3.58	0.005	–
Intercept for ranking = 8	−1.38	0.310	–	−0.44	0.707	–
Intercept for ranking = 7	0.75	0.589	–	2.35	0.059	–
Intercept for ranking = 6	2.42	0.083	–	4.67	0.000	–
Intercept for ranking = 5	5.52	0.000	–	8.32	<0.001	–
Intercept for ranking = 4	9.31	<0.001	–	18.85	<0.001	–
Intercept for ranking = 3	11.22	<0.001	–	21.64	<0.001	–
Intercept for ranking = 2	12.79	<0.001	–	33.82	<0.001	–
DPR	0.61	<0.001	1.85	0.50	<0.001	1.65
Percentage of Orthodox population	−0.04	0.042	0.96	−0.03	0.051	0.97
Percentage of ultra-Orthodox population	−0.09	<0.001	0.91	−0.29	<0.001	0.75
Percentage of Ethiopian immigrants	−0.15	0.318	0.86	−0.59	0.000	0.55
Percentage of immigrants from the former USSR since 1990	−0.12	<0.001	0.89	−0.12	<0.001	0.89
Rate of cases of bodily injury crimes	−0.27	<0.001	0.76	−0.30	<0.001	0.74
Rate of cases of property crimes	−0.31	0.068	0.74	−0.0003	0.43	1.00
North district	1.46	0.015	4.33	0.25	0.66	1.29
South district	1.64	0.032	5.12	0.97	0.209	2.65
Number of terror events	−0.18	0.081	0.84	−0.02	0.839	0.98
Number of observations			98			107
Percent concordant			94.2			94.6

with $\alpha_i, i = 2, \ldots, 10$ being the intercepts of the model for cluster values 2,. . .,10 respectively, where cluster 1 was chosen to be the reference category. In (3), γ denotes a vector of the regression coefficients to be estimated and X the set of explanatory variables.

Table 3 presents the final estimated models for 2001 and 2003, with variables that are significant for at least one of the years (significance level 0.10).

It can be seen that there is a positive correlation between the DPR and the probability of appearing in a higher SEC, given all the other controlled variables.

However, it was found that this influence is partially offset as a result of the effect of minorities, for example the percentage of religious population (both ultra-Orthodox and Orthodox), the percentage of immigrants from Ethiopia, and the percentage of immigrants from the former USSR.

Given all other controlled variables, including DPR, it appears that the location in peripheral districts increases the odds for being in a higher SEC.

The effects of both bodily injuries and property crimes as well as the number of terrorism incidents were found to be significant and negative.

5.2. *OLS Model on DPL*

A regression model was estimated for the continuous DPL variable which served for constructing the DPR.

In order to validate the obtained estimators, we carried out appropriate statistical tests to identify possible multicollinearity, residual dependence and residual normality. Figures 1A and 2A (Appendix 3) demonstrate that the DPL is close to normally distributed, justifying use of the OLS model. Figures 3A and Table 1A (Appendix 3) display the residuals normal probability plots for the estimated OLS models, showing that the residuals' distribution is approximately normal. Statistical test results (such as the Durbin-Watson test and 1st Order Autocorrelation test for residual independence,

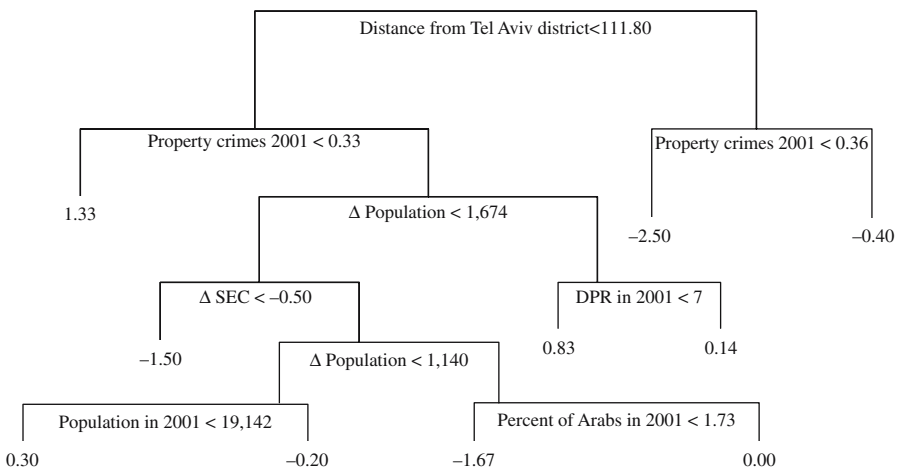


Fig. 1. Regression tree for change in the DPR between 2001 and 2003

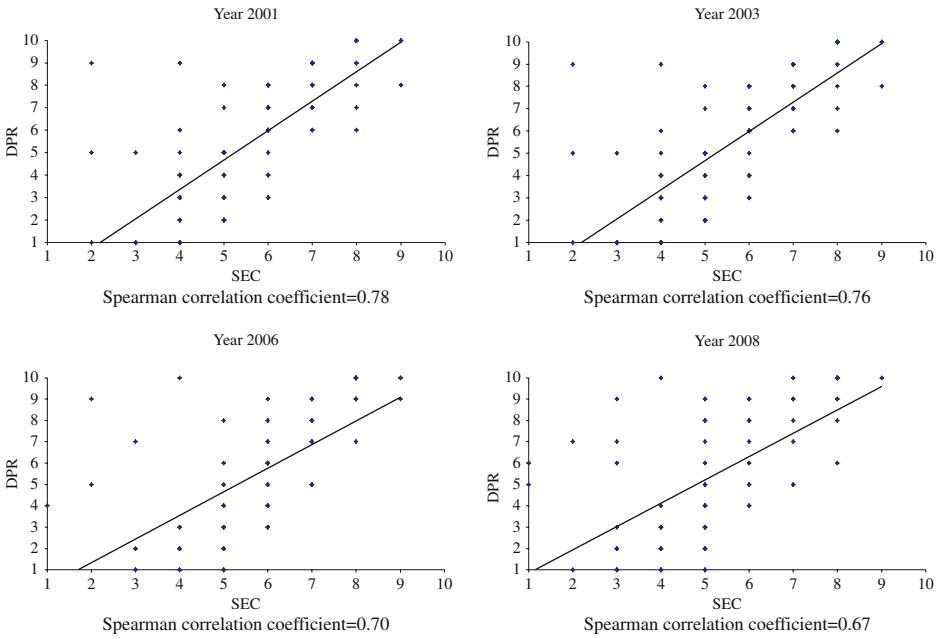


Fig. 2. Socioeconomic cluster vs. dwelling price ranking (Sources: Tax Authority, Central Bureau of Statistics)

and the Shapiro-Wilk and Kolmogorov-Smirnov tests for residual normality) show that at five percent significance level we cannot indicate residual dependence or significant deviation from normality. In addition, results from the correlation analysis do not reveal any evidence of multicollinearity.

The model can be represented as:

$$DPL_k = \alpha + \beta SEC_k + \gamma X_k + \varepsilon \tag{4}$$

In (4), β denotes the regression coefficient for SEC variable in locality k, X_k is a set of explanatory variables for this locality, γ is a vector of coefficients of X_k to be estimated, and ε denotes model residuals with zero expected value and constant variance.

Table 4 shows that most of the effects found to be significant in Model (3) on the SEC variable are also significant in (4) estimated on the DPL and follow the same directions, such as overall effect of crime and the effect of minorities (percentage of Orthodox, percentage of the immigrants from Ethiopia and from the former USSR). The positive and significant estimate of the property crimes variable in 2003 might be explained by "special" correspondence between property crime and dwelling prices, where more expensive residential areas "invite" property crimes. In addition, the negative estimate of the distance from the Tel Aviv district reflects a strong peripherality effect in Israel. However, this influence is nonlinear. A positive sign of the squared term means that the peripherality effect weakens as distance from the center of national economic activity

Table 4. OLS models on "logarithm of the median dwelling price in a locality"

Variable	2001		2003	
	Estimate	*p*-value	Estimate	*p*-value
Intercept	8.819	<0.001	8.647	<0.001
SEC	0.062	<0.001	0.067	<0.001
Percentage of Orthodox population	−0.005	0.003	−0.001	0.443
Percentage of immigrants from the former USSR since 1990	−0.006	0.001	−0.006	0.010
Percentage of Ethiopian immigrants	−0.048	0.000	0.036	0.017
Rate of cases of bodily injury crimes	0.0001	0.975	−0.010	0.096
Rate of cases of property crimes	−0.0006	0.432	0.002	0.052
Distance from Tel Aviv district	−0.005	<0.001	−0.005	<0.001
Distance from Tel Aviv district −squared function	0.00001	0.000	0.00001	0.002
Small locality*	−0.141	0.003	−0.208	<0.001
Total population in a locality (tens of thousands)	0.005	0.029	0.006	0.011
Number of observations		98		107
Adjusted R^2		0.77		0.73

*Localities with 10,000 residents or less (dummy variable)

(Tel Aviv) increases. This occurs due to the existence of additional employment centers in various peripheral towns and the decreasing effect of distance from Tel Aviv in regions that are very far from it.

Estimated Equation (4) for locality k is given by: $DPL_k = \hat{\alpha} + \hat{\beta}SEC_k + \hat{\gamma}X_k$. It follows, that the SEC variable can be expressed as: $SEC_k \approx (DPL_k - \hat{\alpha} - \hat{\gamma}X_k)/\hat{\beta}$. It should be noted that the above approximation for SEC may not be an integer due to continuous characteristics of $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$.

Therefore, this approximation for SEC in a locality k is given by:

$$\overline{SEC}_k = DPR_{corrected} = Round\left[\frac{(DPL_k - \hat{\alpha} - \hat{\gamma}X_k)}{\hat{\beta}}\right] \quad (5)$$

Using (4) and (5), we can examine whether the gap between the SEC and the DPR defined earlier can be at least partly bridged. Table 5 presents the distribution of absolute values of differences between the SEC and the DPR, before and after the correction.

Table 5. Distribution of absolute differences between the SEC and the DPR

	Mean		Min		Max		Percentiles					
							10th		50th		90th	
	2001	2003	2001	2003	2001	2003	2001	2003	2001	2003	2001	2003
Original	1.673	1.776	0	0	7	7	0	0	1	2	3	3
Corrected	0.235	0.234	0	0	1	1	0	0	0	0	1	1

The results presented in Table 5 allow us to conclude that the examined socioeconomic factors that are not currently included in the calculation of the SEC may contribute to a better approximation of the suggested indicator. Furthermore, the earlier obtained Spearman correlation coefficients are also improved to some extent, now being equal to 0.70 and 0.71 for 2001 and 2003, respectively.

5.3. Regression Tree Analysis

In order to draw relevant conclusions on factors influencing the dynamics of the DPR between 2001 and 2003, a nonparametric regression tree was built.

This method of analysis was chosen for the following reasons. First, nonparametric methodology allows for the inclusion of localities in the Arab sector, despite a very small number of available observations and significant differences between the housing markets in Jewish and Arab sectors as mentioned in Section 5. Second, the variables which were removed from the regression models due to multicollinearity can be included in the nonparametric analysis (such as percentage of the Arab population). Given these explanations, the nonparametric regression method is designed to complete and enrich the results obtained from the analysis presented in Subsections 4, 5.1 and 5.2. In this analysis, the dependent variable was defined as the difference between the DPR in 2003 and the DPR in 2001.

To the explanatory variables used in Model (4) we added the differences between the values of these variables in 2003 and 2001. The regression tree method divides observations into homogeneous groups of a dependent variable, given a set of explanatory variables. A detailed description of the regression tree methodology is presented in Breiman et al. (1984).

The algorithm is iterative and works as follows. Initially, from the explanatory variables and their values, the algorithm finds a variable and its values which divides all the observations into two distinct groups, so that the variance of the dependent variable within each group ("leaf") is minimal and the variance between these two groups is maximal (among all possible combinations). This value is fixed as the "split point". The same process is repeated until a specified stopping criterion is fulfilled. At each stage, analysis is performed on the full set of the input variables; therefore, the same explanatory variable can be used several times. This method reveals nonlinear relationships between the dependent and explanatory variables.

The R^2 index was used to test goodness of fit. Let SSW_k denote the estimated variance within the final group k. Since the groups are independent, the total variance within all the groups is calculated by: $SSW = \sum_k SSW_k$. Using the definition of the index R^2, it is given by:

$$R^2 = 1 - \frac{SSW}{SST} \tag{6}$$

where SST denotes the total variance of the dependent variable. The higher the value of the R^2, the better the classification achieved (in terms of the homogeneity of the final "leaves") relative to a previous iteration. In our case, the value of the R^2 index is equal to 0.65.

In Figure 1, the height of the lines between the split points shows the reduction in variance within the group as a result of the division described above, while the numerical value in the final group shows the average increase/decrease in the dependent variable for those leaves. The left branch of each bifurcation corresponds to the "yes" alternative, that is, the condition being fulfilled.

It appears that the dominant factors for change in the DPR are a locality's geographical location and its crime level. For example, in the localities that are situated rather close to the Tel Aviv district (less than 111.8 km) and where the property crime rate was less than 0.33 in 2001, the dwelling price ranking rose by an average of 1.33 (the "leaf" furthest to the left). A decrease in the SEC in a locality caused a consequent decrease in its DPR, given changes in its population and property crime rates in 2001. It also appears that given other controlled variables, an increase in total population is correlated with an increase in the DPR. Additionally, the low percentage of the Arab population in 2001 is correlated with a decrease in the DPR.

6. Conclusions

The current study examined the question of whether dwelling prices in a given locality can serve as an approximation to its socioeconomic level.

The study is based on a number of administrative databases available at a national level. It was found that during the research period (2001 and 2003) there was a strong association between the locality's socioeconomic cluster and the value of its dwellings, with the Spearman correlation coefficients almost identical for these two years.

An analysis of recently obtained SEC data for the years 2006 and 2008 shows that the results obtained for 2001 and 2003 remained consistent, as did the results for 2006 and 2008 (Figure 2). However, a gap has been found between the SEC and the DPR. Our results show that this gap may be explained by location, other social and demographic factors, crime and security characteristics that are exogenous to SEC. In particular, a significant correlation was found between dwelling prices in a specific locality and the percentage of those belonging to defined population groups. It was also found that the size of a locality has a positive correlation with the level of the dwelling prices there. It appears that the effect of the distance from the center of Israel's economic activity is negative, as expected.

It was found that these effects, which reflect other social and demographic characteristics that are not currently included in the SEC calculation, may bridge, at least partly, the gap between the SEC and the DPR.

Overall, we conclude that the ranking based on dwelling prices can serve as a rather good approximation to the socioeconomic level of most urban localities in Israel.

Obviously, this approximation may not always be accurate for some of the localities. Nevertheless, the proposed methodology can provide the required information on socioeconomic profile. This finding is extremely important since the process of SEI and SEC calculation requires a wide variety of variables from various data sources available concurrently for various localities, while the dwelling price ranking allows a rather simple approximation of SEC for different localities for every given year.

For the sake of the methodological consistency and comparability of the SEC series, at the current stage we do not suggest any changes in the set of variables used for the SEC calculations. Rather, we propose the DPR index as an approximation to SEC values for those localities and for years when SEC variables are not available ("intermittent points"). In such cases, using the DPR index, particularly after corrections are made according to Equation (5), can serve as an important working tool for the users of the SEC, such as the Ministry of Finance, Ministry of the Interior, planning authorities and others ensuring a continuum of the index series.

The proposed methodology and the obtained findings are likely to be valid and applicable for different statistical purposes in other countries which possess administrative data on dwelling transactions. For countries that compute socioeconomic indices, the proposed methodology may be used for assessing SEC values for time points and localities for which this index is missing. For countries that do not compute such indices, dwelling price ranking may be used to characterize the socioeconomic profile of a given locality. The suggested approximation may also be used for studying trends in SEC compared with DPR over the years.

Further development and application of an adjustment methodology for SEC imputation is a subject of future research.

Appendix 1. Descriptive Statistics

Name of variable	2001			2003		
	Mean	Std Dev	Median	Mean	Std Dev	Median
Log (DPL)	8.72	0.36	8.64	8.70	0.38	8.68
Total population in a locality	49,287	85.706	21.441	47,299	86,918	21,074
Percentage of Arab population (from the total population in a locality)	11.10	20.56	4.58	10.68	19.98	4.22
Percentage of Orthodox population (from the Jewish population in a locality)	13.56	11.26	11.41	14.90	15.16	11.79
Percentage of ultra-Orthodox population (from the Jewish population in a locality)	13.66	32.20	4.78	12.43	23.09	4.38
Percentage of immigrants from the former USSR since 1990 (from the Jewish population in a locality)	11.73	11.41	7.25	10.86	10.99	6.55
Percentage of Ethiopian immigrants (from the Jewish population in a locality)	0.87	1.47	0.18	0.81	1.43	0.14
Rate of cases of bodily injury crimes (per 1,000 residents)	0.9	0.64	0.78	0.82	0.56	0.69
Rate of cases of property crimes (per 1,000 residents)	40.36	27.40	35.59	38.14	25.81	37.76
Number of terror incidents in a locality	0.43	2.16	0	0.70	3.00	0
Districts: Jerusalem	0.03	0.17	0	0.04	0.18	0
North	0.22	0.42	0	0.20	0.40	0
Haifa	0.15	0.36	0	0.14	0.35	0
Center	0.27	0.45	0	0.28	0.45	0
Tel Aviv	0.11	0.31	0	0.10	0.30	0
South	0.14	0.35	0	0.14	0.35	0
(Dummy variables: 1 if defined district, 0 –otherwise)						
Distance from the Tel Aviv district (km)	55.31	50.76	46.24	54.60	51.12	44.31

Appendix 2. Correlations Between the SEC and the Explanatory Variables

Variables	2001		2003	
	Correlation coefficient	Level of significance	Correlation coefficient	Level of significance
Total population	−0.038	0.702	−0.037	0.702
Percentage of Arab population	−0.445	<0.001	−0.351	0.000
Percentage of Orthodox population	−0.307	0.002	−0.202	0.035
Percentage of ultra-Orthodox population	−0.363	0.000	−0.635	<0.001
Percentage of immigrants from the former USSR since 1990	−0.431	<0.001	−0.359	0.000
Percentage of Ethiopian immigrants	−0.223	0.026	−0.229	0.016
Rate of cases of bodily injury crimes	−0.487	<0.001	−0.434	<0.001
Rate of cases of property crimes	−0.081	0.413	0.055	0.562
Number of terror incidents in a locality	−0.062	0.536	−0.059	0.538
District: Jerusalem	−0.041	0.684	0.014	0.881
North	−0.263	0.007	−0.256	0.006
Haifa	0.071	0.474	0.044	0.647
Center	0.217	0.027	0.233	0.013
Tel Aviv	0.195	0.049	0.181	0.056
South	−0.138	0.163	−0.143	0.134
Distance from the Tel Aviv District	−0.301	0.002	−0.307	0.001

Appendix 3. Validation of OLS Assumptions

3.1. Distribution of the Dependent Variable

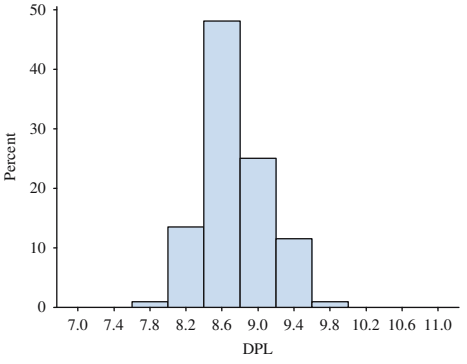


Fig. 1A. Distribution of the DPL 2001

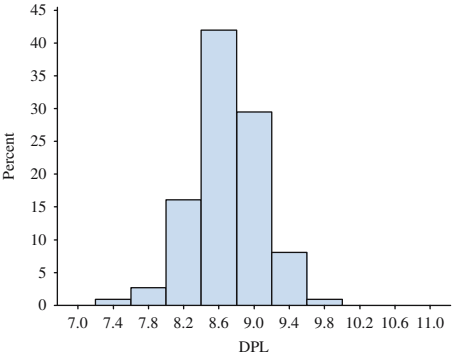


Fig. 2A. Distribution of the DPL 2003

3.2. Residual Normality

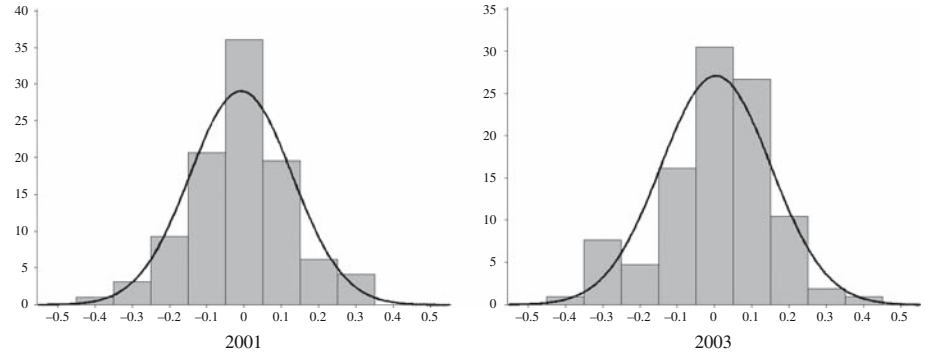


Fig. 3A. Histograms of the model residuals and normal density curves, 2001 and 2003

Table 1A. Test for normality of residuals

	Test (*p*-value)	
Year	Kolmogorov-Smirnov	Shapiro-Wilk
2001	0.66	0.07
2003	0.15	0.10

7. References

Australian Bureau of Statistics. 2006. "An Introduction to Socio-Economic Indexes for Areas (SEIFA)." Information Paper No. 2039.0. Available at: http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/2039.0Main%20Features22006?opendocument&tabname=Summary&prodno=2039.0&issue=2006&num=&view (accessed April 2015).

Blanchard, T.C. 2007. "Conservative Protestant Congregation and Racial Residential Segregation: Evaluating the Closed Community Thesis in Metropolitan and Nonmetropolitan Counties." *American Sociological Review* 72: 416–433. Doi: http://dx.doi.org/10.1177/000312240707200305.

Bourassa, S. C., M. Hoesli, and V.S. Peng. 2003. "Do Housing Submarkets Really Matter." *Journal of Housing Economics* 12: 12–28. Available at: http://www.sciencedirect.com/science/article/pii/S1051137703000032 (accessed April 2015).

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and Regression Trees*. Wadsworth: Belmont.

Buck, A.J., J. Deutsch, J. Hakim, U. Spiegel, and J. Weinblatt. 1991. "A Von Thunen Model of Crime, Casinos and Property Values in New Jersey." *Urban Studies* 28: 673–686. Doi: http://dx.doi.org/10.1080/00420989120080861.

Burck, L. and Y. Kababia. 1996. *Characterization and Ranking of Local Authorities according to the Socio-Economic Level of the Population in 1995*. Publication No. 1039, Central Bureau of Statistics, Jerusalem, (Hebrew).

Burck, L. and Y. Kababia. 1999. *Characterization and Ranking of Local Authorities according to the Socio-Economic Level of the Population in 1999, Based on the 1995 Census of Population and Housing*. Publication No. 1118, Central Bureau of Statistics, Jerusalem. (Hebrew).

Cahaner, L. 2012. "Expansion Processes of the Jewish ultra-Orthodox Population in Haifa." In *Themes in Israel Geography, Special Issue of Horizons in Geography*, edited by J.O. Maos and I. Charney, 70–87, University of Haifa.

Central Bureau of Statistics (CBS). 2000. Characterization and Classification of Geographical Units by the Socio-Economic Level of the Population. 1995 Census of Population and Housing Publications, No. 13, Jerusalem.

Cummings, J. L., D. DiPasquale, and M. E Kahn. 2002. "Measuring the Consequences of Promoting Inner City Homeownership." *Journal of Housing Economics* 11: 330–359. Available at: http://www.cityresearch.com/pubs/dipasquale.pdf (accessed April 2015).

Des Rosiers, F., M. Theriault, Y. Kestens, and P. Villeneuve. 2002. "Landscaping and House Values: An Empirical Investigation." *Journal of Real Estate Research* 23: 139–161.

Dubin, R.A. and A.C. Goodman. 1982. "Valuation of Education and Crime Neighborhood Characteristics Through Hedonic Housing Price." *Population and Environment* 5: 166–181. Doi: http://dx.doi.org/10.1007/BF01257055.

Goodman, A.C. and T.G. Thibodeau. 1998. "Housing Market Segmentation." *Journal of Housing Economics* 7: 121–143. Doi: http://dx.doi.org/10.1006/jhec.1998.0229.

Greenberg, M.R. 1999. "Improving Neighborhood Quality: A Hierarchy of Needs." *Housing Policy Debate* 20: 601–624. Doi: http://dx.doi.org/10.1080/10511482.1999.9521345.

Harris, D.R. 1999. "Property Values Drop When Blacks Move in, Because. . .: Racial and Socioeconomic Determinants of Neighborhood Desirability." *American Sociological Review* 64: 461–479.

Hazam, S. and D. Felsenstein. 2007. "Terror, Fear and Behaviour in the Jerusalem Housing Market." *Urban Studies* 44: 2529–2546. Doi: http://dx.doi.org/10.1080/00420980701558392.

Heikkila, E. 1992. "Describing Urban Structure: A Factor Analysis of Los Angeles." *Review of Urban and Regional Development Studies* 4: 84–101. Doi: http://dx.doi.org/10.1111/j.1467-940X.1992.tb00035.x.

Horton, H.D. and M.E. Thomas. 1998. "Race, Class, and Family Structure: Differences in Housing Values for Black and White Homeowners." *Sociological Inquiry* 68: 114–136. Doi: http://dx.doi.org/10.1111/j.1475-682X.1998.tb00456.x.

Jackson, E., V. Kupke, and P. Rossini. 2007. The Relationship between socio-economic indicators and residential property values in Darwin. Paper presented at the 13th Annual Pacific-Rim Real Estate Society Conference. Fremantle, Western Australia. Available at: http://scholar.google.com.au/citations?view_op=view_citation&hl=en&user=KbV4jccAAAAJ&citation_for_view=KbV4jccAAAAJ:3fE2CS-JIrl8C (accessed April 2015).

Jolliffe, I.T. 2002. *Principal Component Analysis*, 2nd ed. Springer Series in Statistics. New York: Springer.

Kiel, K.A. and J.E. Zabel. 1996. "House Price Differentials in U.S. Cities: Households and Neighborhood Racial Effects." *Journal of Housing Economics* 5: 143–165. Doi: http://dx.doi.org/10.1006/jhec.1996.0008.

Krivo, L.J. 1995. "Immigrant Characteristics and Hispanic-Anglo Housing Inequality." *Demography* 32: 599–615. Doi: http://dx.doi.org/10.2307/2061677.

Lewin-Epstein, N., Y. Elmelech, and M. Semyonov. 1997. "Ethnic Inequality in Home Ownership and the Value of Housing: The Case of Immigrants in Israel." *Social Forces* 75: 1439–1462. Doi: http://dx.doi.org/10.1093/sf/75.4.1439.

Lozano-Gracia, N. and L. Anselin. 2012. "Is the Price Right? Assessing Estimates of Cadastral Values for Bogota, Colombia." *Regional Science Policy & Practice* 4: 495–508. Doi: http://dx.doi.org/10.1111/j.1757-7802.2012.01062.x.

Malpezzi, S., G.H. Chun, and R.K. Green. 1998. "New Place-to Place Housing Price Indexes for U.S. Metropolitan Areas, and Their Determinants." *Real Estate Economics* 26: 235–274. Doi: http://dx.doi.org/10.1111/1540-6229.00745.

McCluskey, W. J., W. G. Deddis, I. G. Lamont, and R. A. Borst. 2000. "The Application of Surface Generated Interpolation Models for the Prediction of Residential Property values." *Journal of Property Investment & Finance* 18: 162–176. Available at: http://eprints.ulster.ac.uk/10588/ (accessed April 2015).

Myers, D. 1990. *Housing Demography: Linking Demographics Structure and Housing Markets*. Madison: University of Wisconsin Press.

Olausson, P.O. and D. Vagero. 1991. "Miscellanea, A Swedish Classification Into Social Classes Based on Census Information and Comparable to The British Classification–A Proposal." *Journal of Official Statistics* 7: 93–103.

Ozanne, L. and T. Thibodeau. 1983. "Explaining Metropolitan Housing Price Differences." *Journal of Urban Economics* 13: 51–66. Doi: http://dx.doi.org/10.1016/0094-1190(83)90045-1.

Potepan, M.J. 1996. "Explaining Intermetropolitan Variation in Housing Prices, Rents and Land Prices." *Real Estate Economics* 24: 219–245. Doi: http://dx.doi.org/10.1111/1540-6229.00688.

Reed, R. 2001. "The Significance of Social Influences and Established Housing Values." *Appraisal Journal, October* 1.

Rose, D. and D. Pevalin. 2001. *The National Statistics Socio-Economic Classification: Unifying Official and Sociological Approaches to the Conceptualization and Measurement of Social Classes*. Institute of Social and Economic Research Working Papers, November 2001–4. Available at: https://www.iser.essex.ac.uk/files/iser_working_papers/2001-04.pdf (accessed April 2015).

Soffer, A. and E. Bystrov. 2006. *Tel Aviv State: A Threat to Israel*. Haifa, Reuven Chaikin Chair in Geostrategy, University of Haifa. Available at: http://geo.haifa.ac.il/ ~ ch-strategy/publications/english/Tel_Aviv_State.pdf (accessed April 2015).

Thaler, R. 1978. "A Note on the Value of Crime Control: Evidence from the Property Market." *Journal of Urban Economics* 5: 137–145. Doi: http://dx.doi.org/10.1016/0094-1190(78)90042-6.

Yates, J. 2002. *A Spatial Analysis of Trends in Housing Markets and Changing Patterns of Household Structure and Income*. Positioning Paper 30 of the Australian Housing and Urban Research Institute. Sydney Research Centre. Available at: http://www.ahuri.edu.au/publications/download/ahuri_60064_fr (accessed April 2015).

Received July 2013
Revised August 2014
Accepted October 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 249–262, http://dx.doi.org/10.1515/JOS-2015-0016

Big Data as a Source for Official Statistics

Piet J.H. Daas[1], *Marco J. Puts*[1], *Bart Buelens*[1], *and Paul A.M. van den Hurk*[1]

More and more data are being produced by an increasing number of electronic devices physically surrounding us and on the internet. The large amount of data and the high frequency at which they are produced have resulted in the introduction of the term 'Big Data'. Because these data reflect many different aspects of our daily lives and because of their abundance and availability, Big Data sources are very interesting from an official statistics point of view. This article discusses the exploration of both opportunities and challenges for official statistics associated with the application of Big Data. Experiences gained with analyses of large amounts of Dutch traffic loop detection records and Dutch social media messages are described to illustrate the topics characteristic of the statistical analysis and use of Big Data.

Key words: Large data sets; traffic data; social media.

1. Introduction

In our modern world, more and more data are generated on the web and produced by sensors in the ever-growing number of electronic devices surrounding us. The amount of data and the frequency at which they are produced have led to the introduction of the term 'Big Data' (Lynch 2008). Big Data sources can generally be described as: "high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making". This definition is a variant of the definition proposed by Gartner (Beyer and Douglas 2012). For more general information on Big Data and their innovative potential, the reader is referred to Manyika et al. (2011).

In addition to generating new commercial opportunities in the private sector, Big Data are potentially a very interesting data source for official statistics, either for use on their own, or in combination with more traditional data sources such as sample surveys and administrative registers (Cheung 2012). However, extracting relevant and reliable information from Big Data sources and incorporating it into the statistical production process is not an easy task (Daas et al. 2012a). Importantly, the statistical point of view has been underexposed in the work that has been "published" on Big Data so far; this work has been published mainly on weblogs and in conference and white papers. The majority of these publications have an IT perspective as they predominantly focus on soft- and hardware issues, and largely fail to address important statistical issues such as coverage, representativity, quality, accuracy and precision. If Big Data are to be used for official

[1] Statistics Netherlands, Division of Process development, IT and methodology P.O. Box 4481, 6401 CZ, Heerlen, The Netherlands. Emails: pjh.daas@cbs.nl (corresponding author), m.puts@cbs.nl, b.buelens@cbs.nl, and pamvandenhurk@gmail.com

statistics, it is essential that these issues are considered and adequately dealt with (Cheung 2012; Daas et al. 2012a; Glasson et al. 2013; Groves 2011).

In this article we provide an overview of the current state of the research on the usage of Big Data for official statistics at Statistics Netherlands and the lessons learned so far. In the next section a description of two Big Data case studies is given, followed by a more general methodological discussion in Section 3. Finally, conclusions are drawn in Section 4.

2. Big Data Case Studies

In this section we report on two Big Data case studies conducted at Statistics Netherlands. These studies serve as examples and allow for a more general formulation of the statistical issues and challenges involved with the application of Big Data in official statistics. All analyses were performed with the open-source software environment R (R Development Core Team 2012) on a Fujitsu Celsius M470-2 workstation with a 64-bit Windows 7 operating system, 32GB of RAM, 512 GB solid state drive and a 1 TB hard disk. Data were imported into R from CSV files which usually each contained one million rows of data. Each file was subsequently processed and analysed. Results were stored as CSV files. This approach was fast and flexible and sufficed for the studies described in this article.

2.1. Analysis of Traffic Loop Detection Data

Traffic loop detection data consist of measurements of traffic intensity. Each loop counts the number of vehicles per minute that pass at that location, and measures speed and vehicle length one. Such data are interesting for traffic and transport statistics and potentially also for statistics on other economic phenomena related to transport. On the particular day studied, data were collected at 12,622 measurement locations on Dutch roads. The data are stored centrally in the National Data Warehouse for Traffic Information (NDW) and managed by a collaboration of participating government organizations (NDW 2012). The National Data Warehouse contains historic traffic data collected from 2010 onwards. To determine the usability of the NDW data for statistics and to get an idea of its peculiar features, we started by studying minute-level data for all locations in the Netherlands for a single day: December 1st, 2011. The data set extracted from the NDW contained 76 million records, one million per CSV file, which were imported into R via the LaF package (Van der Laan 2013). This package supports loading the data in blocks, enabling the processing of enormous amounts of data without fitting all the data into memory.

Data were first aggregated over all loops, resulting in a series of total counts of all vehicles in the Netherlands at minute intervals. The change of this total count through the day is shown in Figure 1A. The overall profile displays clear morning and evening rush hour peaks around 8 am and 5 pm respectively. Importantly, however, there is a huge variation in the numbers of vehicles detected in subsequent minutes. This phenomenon is caused by the fact that – for a substantive number of minutes – data were only available for a subset of all detection loops in the country. This appeared to be caused by some computers failing to submit data to the warehouse at certain time points.

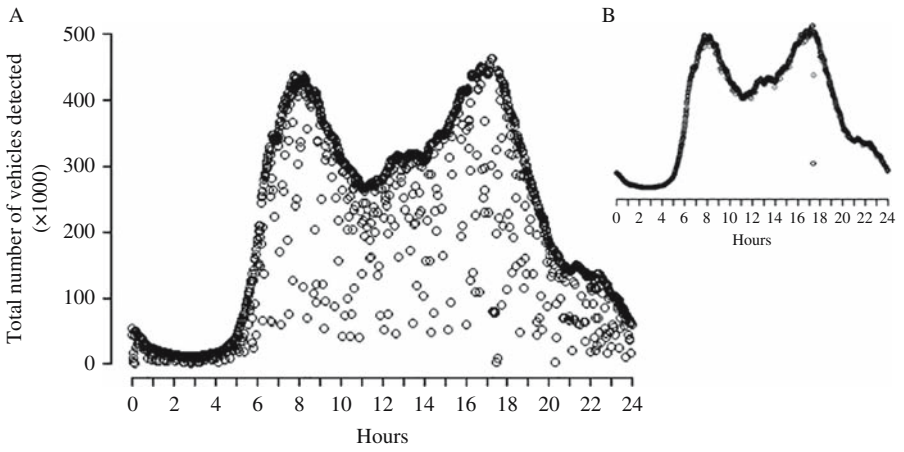


Fig. 1. (A) Total number of vehicles detected per minute in the Netherlands on December 1st, 2011. (B) Results after correcting for missing data.

From a statistical point of view there are various ways to solve such a missing data problem (De Waal et al. 2011).

Because aggregated data were used and this was our first experience with huge amounts of data, we opted for the simplest solution: add (impute) data reported by the same location during a short interval before or after the time point of missing data (if available). More specifically, a sliding, symmetrical five-minute time window to impute data at missing time points for the entire data set was applied. The resulting data pattern is shown in Figure 1B. Except for a period shortly after 5 pm, the majority of the missing data points were adequately replaced with timely data of the same measurement location. As a result of this data-editing procedure a total of nearly 35.8 million vehicle counts were added, which is slightly more than twelve percent of the number of vehicles originally counted, 294.7 million. Alternative model-based approaches can be applied and are preferred when traffic loop data are studied for smaller areas (more on this below).

The edited data set was used to create maps that indicate the number of vehicles for each measurement location for each time point by means of colour coding. Next, by sequencing these maps, a movie was created that displays the changes in vehicle counts for all locations during the day. Thus, this movie (not shown here) illustrates the increases and decreases in traffic intensity in the Netherlands throughout the day studied (Daas et al. 2012b). Unsurprisingly, the traffic intensity between the four major cities in the Netherlands (Amsterdam, Rotterdam, Utrecht and The Hague) was especially high, during all working hours and in the early evening.

Besides the total number of vehicles, the number of vehicles in various length categories was also studied. Because not all detection locations are able to differentiate between different vehicle lengths, only those that are able to do so were used. This subset consisted of 6,002 detection locations, which represented 48 percent of the total number of locations. Vehicles were sorted into three length categories: small (<= 5.6 metre), medium-sized (> 5.6 and <= 12.2 metre), and large (> 12.2 metre). Again, the imputed data set was used. Because the small vehicle category comprised around 75 percent of all vehicles detected, as compared to twelve percent for the medium-sized and 13 percent for the large

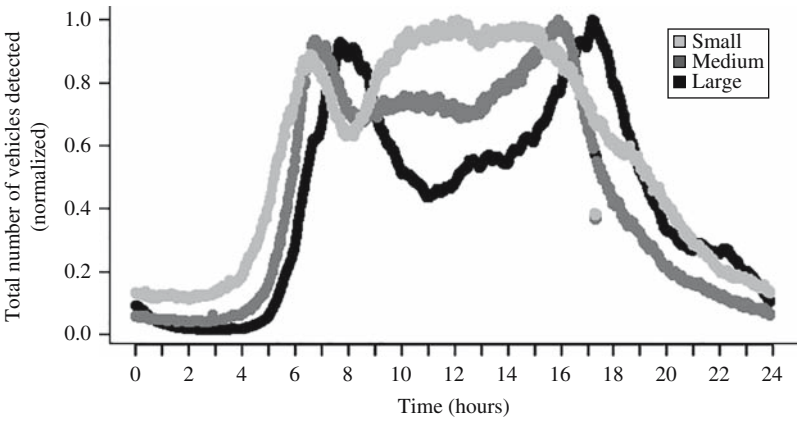


Fig. 2. Normalized numbers of vehicles detected in three length categories on December 1st, 2011 after correcting for missing data. Numbers of small (<= 5.6 meter), medium-sized (> 5.6 and <= 12.2 meter) and large vehicles (> 12.2 meter) are shown. Profiles are normalized by dividing by the maximum value of each series to more clearly reveal the differences. Maximum values are 119,523, 8,673 and 8,599 for small, medium and large vehicles, respectively.

vehicles categories, the normalized results for each category are shown in Figure 2. This figure illustrates the difference in driving behaviour between the three vehicle length categories. The small vehicle category displays clear morning and evening rush-hour peaks at 8 am and 5 pm respectively, in line with the overall profile described above (Figure 1). This finding is not unexpected, as this category of vehicles constitutes the vast majority of all vehicles. The medium-sized vehicles in turn have both an earlier morning and evening rush-hour peak, at 7 am and 4 pm respectively. Finally, the large vehicle category shows a clear morning rush-hour peak around 7 am and more dispersed driving behaviour during the remainder of the day; after 3 pm the number of large vehicles gradually declines without any apparent evening rush-hour peak. Most remarkable is the decrease in the relative number of medium-sized and large vehicles detected at 8 am, that is, during the morning rush-hour peak of the small vehicles. This may be caused by a deliberate attempt of the drivers of the medium-sized and large vehicles to avoid the morning rush-hour peak of the small vehicles or an effect of the more intense traffic (of small vehicles) around that time. Considering these differences, differentiation between vehicles of various lengths when creating a traffic index would not only enable more granular traffic statistics but can also provide more detailed information on transport and phenomena related to economic growth.

In addition to the analysis of traffic intensity at an aggregated level across all detection loops, the traffic intensity profile of a number of individual measurement locations was also studied, for example on highway A4 near Bergen op Zoom. The total number of vehicles detected at this location is shown in Figure 3. Detection at this location displays the same rush-hour peaks as in Figure 1. In addition, the characteristic volatile behaviour of traffic intensity data at the microlevel is shown. Given that this detection location does not suffer from missing data, the changes in the number of vehicles counted each minute are the result of real changes in the number of vehicles passing at this location. However, these rapid fluctuations are not very informative for the production of a traffic index,

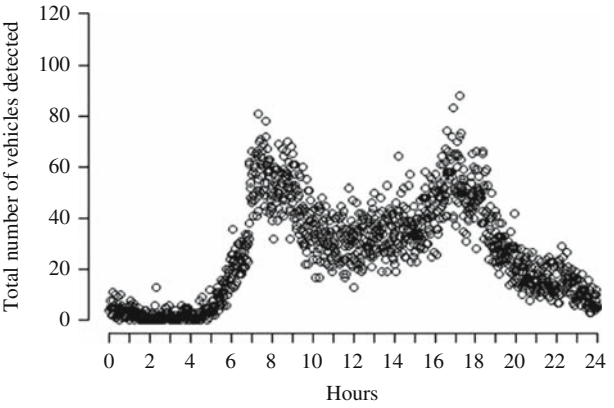


Fig. 3. Total number of vehicles counted by a detection location on highway A4 near Bergen op Zoom.

as interest is focused more upon gradual, long-period, changes, for example, weekly or monthly changes in the number of vehicles (of a certain length class) in a specific region. We are currently studying statistical modelling methods that can deal adequately with these kinds of Poisson-distributed data, such as Bayesian-based signal filters (Manton et al. 1999). These methods need to be applied in a reasonable time to the large amounts of loop detection data. The latter requires high-performance computing techniques (NAS 2013) when applied to the data of all loops in the whole country.

In the analyses in this section, we have assumed that all measurements are without error, except when entire records are missing. The missing records have been imputed using fixed values, not taking into account uncertainty associated with the imputation procedure. Alternatively, a multiple imputation approach could be implemented to account for such uncertainty, which would result in variances and confidence intervals for the aggregates shown above. The aggregates are obtained simply by summing individual loop counts. We have not conducted any form of inference or estimation (except for the imputations). In the future we may do so in order to obtain estimates that are representative of all Dutch highways, including those without traffic loops. A predictive modelling approach would need to be developed, resulting in estimated counts at locations without loops. This would lead to estimated aggregates and variance estimates reflecting the uncertainty of the estimation procedure.

2.2. *Analysis of Social Media Messages*

It is estimated that around 70 percent of the Dutch population actively posts messages on social media (Eurostat 2012). The three million or so Dutch messages generated each day (Daas and Puts 2014) may be an interesting data source for official statistics because they reflect many different aspects of our daily lives. We have studied two aspects of social media messages: content and sentiment. Studies of the content of Dutch Twitter messages – the dominant publicly available social medium in the Netherlands (see below) – revealed that nearly 50 percent of the messages are "pointless babble" (Daas et al. 2012a). In the remainder of the messages, spare-time activities, work, media (TV & radio) and

politics were predominantly discussed. This finding suggests that these messages could be used to extract opinions, attitudes, and sentiments towards these topics, opening up possibilities to collect a considerable amount of interesting information quickly without any response burden. The major problem in analysing social media messages is discriminating the informative from the noninformative ones. Because of the large share of the noninformative "babble" messages, usage of the more serious (informative) messages is negatively affected as many words of interest occur in both types of messages. Text mining approaches to automatically differentiate between both groups of messages have not been very successful so far (Daas et al. 2012a).

Another potential source of information in social media messages is their sentiment. Access to over 1.6 billion public messages written in Dutch from a large number of social media sites was obtained using an infrastructure provided by Coosto (2013). Public messages were sourced from the largest social media sites used by Dutch individuals, such as Twitter, Facebook, Hyves, Google+, and LinkedIn, as well as from numerous public Dutch weblogs and forums. The overall profile of the number of messages created per day revealed that from June 2010 onwards, increasing numbers of messages were generated in the Netherlands on a daily basis. The latter date corresponds to the period during which Coosto started to include huge numbers of Twitter messages in their Hadoop-based distributed database. We therefore used June 2010 as the starting date for our studies, with August 2012 as the end date. Messages could be selected from the database with a query language and a secure web interface. Coosto also determined the sentiment of each message by counting the number of positive and negative words following the general approach described in Golder and Macy (2011). Messages were classified as positive, negative or neutral depending on their overall score. A more detailed description of this part of the work can be found in Daas and Puts (2014).

Since several studies have been performed in English-speaking countries attempting to link the sentiment in social media to consumer confidence (O'Connor et al. 2010; Lansdall-Welfare et al. 2012) we were interested in studying this "relation" for the Netherlands. We looked at the sentiment in messages produced on the various platforms covered by the Coosto data set. The results were intriguing. The development of the sentiment in all Facebook messages produced during the period studied, nearly 170 million (almost ten percent of all messages produced), was found to correlate highly with consumer confidence; $r = 0.84$. Combing the sentiment of all Facebook and Twitter messages, slightly over 1.4 billion (close to 90% of all messages), with a linear model increased the correlation to $r = 0.88$. To reduce the risk of discovering spurious or false correlations, the series were additionally checked for cointegration. Cointegration provides a stronger argument as it checks for a common stochastic drift, indicating that series exhibit fluctuations around a common trend (Engel and Granger 1987). Here, it was found that the sentiment in Facebook and the combination of Facebook and Twitter both cointegrated with consumer confidence, suggesting a strong association between the developments in both series. Remarkably, the sentiment in Twitter messages only correlated less, $r = 0.61$, and did not cointegrate.

Figure 4 displays the survey-based Consumer Confidence series (Statistics Netherlands 2013) and the corresponding Dutch social media sentiment findings for the period studied. Both series relate quite well. This association is remarkable, as the

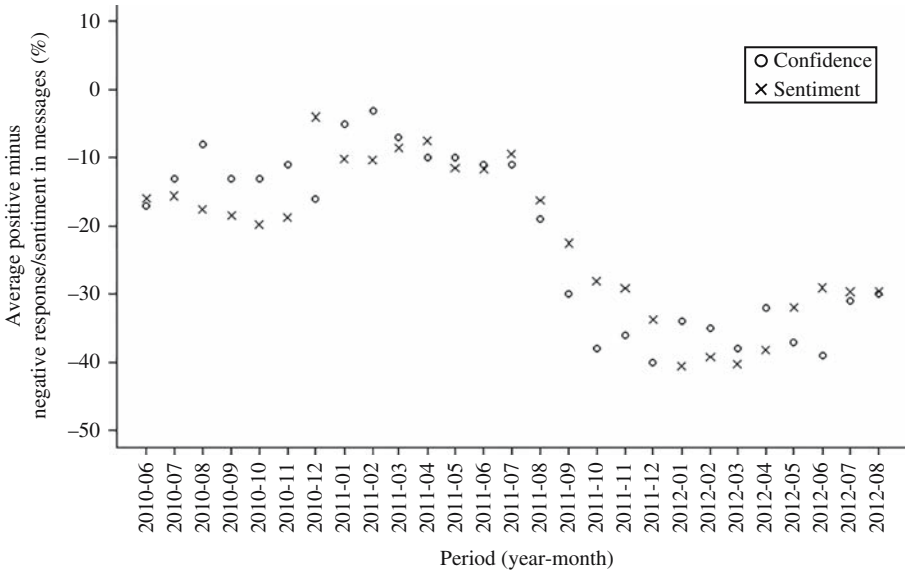


Fig. 4. Comparison of Dutch consumer confidence (○) and the sentiment in Dutch Facebook and Twitter messages on a monthly basis (×). A correlation coefficient of $r = 0.88$ is found for both series.

populations from which the data are obtained are very different. Dutch consumer confidence is obtained from a random sample from the population register, with around 1,000 persons responding each month. Due to sampling variance, the standard errors of the consumer confidence series shown in Figure 4 are on average approximately 2.0. The sentiment in Dutch Facebook and Twitter messages is derived from around 52 million *messages* generated each month. These messages are created by a considerable part of the population, 70 percent according to Eurostat (2012), but i) not all social media messages created in the Netherlands are written in Dutch and ii) different users post varying numbers of messages on various platforms. We have not attempted to estimate the sentiment of the (unknown) subpopulation who does not contribute to social media platforms. Consequently, our social media sentiment series is not subject to sampling, modelling or prediction uncertainty, but may be biased because of differences in the composition of the Dutch population and those active on social media. Previous work by Daas et al. (2012a) also revealed that the number of Twitter messages can vary from 200 per day to not even one message a month for a single person. More recent work has confirmed that the association between both series remains stable over time and that consumer confidence and social media sentiment are related from a Granger-causality perspective (more in Daas and Puts 2014).

3. Discussion

The two case studies described in this article reveal several issues that need to be addressed before Big Data can become a useful and reliable data source for the field of official statistics. These issues, the most important considerations, the way we have dealt with them and the lessons learned are discussed below.

3.1. Data Exploration

Typically, Big Data sets are made available to us, rather than designed by us. As a consequence, their contents and structure need to be understood prior to using the data for statistical analysis (Hassani et al. 2014). This first step is called data exploration, which is aimed at revealing data structure and patterns and, no less important, at assessing the quality of the data as revealed by the presence of errors, anomalies and missing data. Visualisation methods have been proven to be very insightful for such tasks (Fry 2008; Zikopoulos et al. 2012, ch. 7). Recently, certain visualisation methods have been developed that are particularly suited to the exploration of Big Data. Examples are tableplots (Tennekes et al. 2013) to display Big Data with many variables and 3D heatmaps to study variability in multivariate continuous data (Daas et al. 2012b). Sequencing 2D plots into animations is useful to visualise temporal and other aspects of Big Data (Daas et al. 2012b).

3.2. Missing Data

By studying the traffic intensity data on a minute-by-minute level, we discovered that part of the data were missing. If we had analysed the data aggregated at hourly or daily levels, we would have reduced the amount of data studied but would not have noticed that missing data is such a big problem. Since Statistics Netherlands plans to use NDW data to produce reliable traffic and transport intensity statistics at a detailed level, the missing data problem needs to be solved. Missing data is not a problem unique to the traffic loop data set, as other data sources are susceptible to missing data too. For instance, server downtime and network outages can lead to missing social media messages or mobile phone data. However, in the end, the time spent on processing also needs to be reduced to a manageable level to enable the production of frequent statistics. Currently statistical models are being explored that are able to cope adequately with missing data and can be applied to enormous amounts of data in a reasonable amount of time. For such an approach to be successful, the combination of the IT infrastructure available and the ease with which a modelling method can be upscaled needs to be assessed (NAS 2013). We are currently focusing on Bayesian approaches as these are applied to enormous amounts of data in other areas of science and are well suited to capturing various forms of uncertainty. The high-performance computing needs can be met at relatively low cost by using the large amounts of computing power provided by the graphics processing units available on many modern graphics cards (Scott et al. 2013).

3.3. Volatility

The number of vehicles detected by individual loops fluctuates considerably from minute to minute. These fluctuations are caused by real changes in the number of vehicles detected but are not very informative from a statistical point of view as they occur at too high a time resolution. Similarly, sentiment analyses on a daily and weekly basis suffer from a volatility that is not seen at monthly intervals (Daas and Puts 2014; O'Connor et al. 2010). It is therefore recommended to develop statistical methods able to cope with volatile behaviour. Possible methods under consideration are the application of moving averages and advanced filtering techniques (e.g., a Kalman filter or time-series modelling).

3.4. Selectivity

The analyses described in Section 2 apply to traffic intensity on roads equipped with traffic loop sensors, and to the sentiment analysis of people who post Dutch Facebook or Twitter messages on social media websites. It is important to realize that both data sets are created by only a subset of the total population in the Netherlands: only vehicles driving on the major Dutch roads were counted and only the sentiment of a subset of all people in the Netherlands was probed, respectively. The subpopulations from which these Big Data sources were derived are not typical target populations for official statistics. Therefore the data are likely to be selective and not representative of a target population of interest. In addition, both sources contain data resulting from the registration of events. These are vehicles passing and messages sent respectively. Both lack directly available data on the units of interest. Usually, the representativity of Big Data can be assessed through the careful comparison of characteristics of the covered population and the target population. Unfortunately, this may prove problematic for these sources, as hardly any such characteristics are available to conduct such a comparison (Buelens et al. 2014). For instance, vehicles can not be uniquely identified in the traffic loop data as licence plate data are absent. Little is known about the people posting on social media; often only their username is known but not their age or gender. In situations where at least some background information is available, the selectivity issue can be assessed and probably resolved. Alternatively, profiling approaches could be used to extract features to estimate, for instance, the chance that a user is male or female (Flekova and Gurevych 2013). Perhaps this could be achieved through predictive modelling, using a wide variety of algorithms known from statistical learning and data mining techniques (Hastie et al. 2009). These are modelling methods not traditionally used in official statistics. Buelens et al. (2012) explore some possibilities for applications of data mining methods in official statistics. More on this topic can be found in ASA (2014).

3.5. Legal Considerations

Privacy and security are issues that may impede NSIs' use of Big Data. In contrast to the legal basis that permits the use of administrative data sources by a lot of NSIs, the use of privately owned Big Data, such as mobile phone data, needs to be specifically arranged (De Jonge et al. 2012). But even for publicly accessible data, such as price and product information on websites, questions of ownership and purpose of publication can be raised. And even if there are no legal impediments, public perception is a factor that must be taken into account. These concerns have to be taken seriously and tackled one at a time. Fortunately, there are measures that can be taken to overcome at least some of the obstacles, for example, by anonymizing unique identifiers, removing the privacy-sensitive part of a Global Positioning System track (e.g., the first and last 100 metres) or by using informed consent. If a reduction of response burden can be offered, this can be very helpful, also in getting the support of the general public. In the long run, changes in legislation may be considered, to ensure continuous data access for official statistics. But it remains important to stay in line with public opinion, because credibility and public trust are important assets. Within the European Union, changes in European legislation must also be considered. In addition to national laws, European laws or regulations can impede

the collection of data, even if the current Dutch legislation does not present any problem (more in Struijs and Daas 2013).

3.6. Data Management

Long-term stability may also be a problem when using Big Data. Typically, statistics for policy making and evaluation are required for extended periods of time, often covering many years. The Big Data sources encountered so far seem subject to frequent modifications, possibly limiting their long-term use. This suggests a need for more flexible data processing and evaluation strategies, which will have to put more emphasis on ongoing data and metadata management to identify, describe and re-evaluate new sources. Data management is also affected by data ownership, copyright and the purpose for which data are registered. Privately owned Big Data in particular may need special arrangements and will probably also incur costs. The two data sources discussed in this article are examples of each. The organisation that maintains traffic loop data is funded by the Dutch government which means that, as laid down in the Statistics Netherlands Act, they provide us with the data free of charge. For the study of social media data, however, costs were incurred, as the data are collected by a privately owned company.

3.7. High-Performance Computing

Processing enormous amounts of data within a reasonable amount of time requires dedicated and specialized computing infrastructures. Hence it can be expected that the inclusion of Big Data as a source for official statistics will certainly affect the IT environment of NSIs. Our experiences so far, however, reveal that considerable progress can be made even with a limited budget. Having a secure computer environment with many fast processors, large amounts of RAM and fast disk access certainly helps. Several important considerations are described in NAS (2013) and Schutt and O'Neil (2013). Parallel processing is the way to speed things up. For instance, we have found that processing traffic loop data in parallel in R results in a 17x speedup over the original (serial) processing time. We are currently using (multicore) general-purpose computing on graphics processing units and are looking at distributed computing, such as our own secure local cluster.

3.8. New Skills Needed

In order to work with Big Data specific technical expertise is needed, such as knowledge of advanced (high-performance) computing and data engineering. These skills speed up the ease with which Big Data can be incorporated into the statistical process and the way it is analysed. In our office, Big Data is usually processed with R or Python. Besides knowing the language, the most important skill here is knowing how to write a program that is able to access and analyse all the data within a reasonable amount of time. Several of our colleagues have written R packages specifically devoted to these tasks, such as LaF (Van der Laan 2013) and ffbase (De Jonge et al. 2014). Moreover, the models used for Big Data must be able to address the levels of complexity that huge data sets can reveal. This makes many of the standard approaches used in official statistics limited in utility and

performance (NAS 2013). The algorithmic-oriented models developed in fields outside statistics might be more applicable here (Breiman 2001; Hastie et al. 2009).

Perhaps just as important is the attitude of the people involved. Working with Big Data requires an open mindset and the ability not to see all problems *a priori* in terms of sampling theory, as Big Data are more similar to large sets of observational data (Daas and Puts 2014). The term "data scientist" has been coined for researchers with the skills identified above (Schutt and O'Neill 2013). We have solved the initial need by including experimentally trained researchers in our Big Data efforts, as they are more practically oriented and are more accustomed to deriving theory from data. The strict difference between methodology, software engineering and IT hardware expertise commonly used at NSIs is also becoming less well defined. At Statistics Netherlands, a group of data scientists is currently being formed. The work of this group is expected to be beneficial to many of the areas of official statistics, especially when large data sources and complex models are used.

4. Conclusions

The official statistics community can greatly benefit from the possibilities offered by Big Data. However, care is needed when trying to implement these sources in official statistics. The two Big Data case studies described show typical issues including missing data, volatility and selectivity, which all need to be adequately dealt with. For this reason, investment in specific research and skills development is needed. In addition, various new areas of expertise are considered necessary to fully exploit the information contained in Big Data. In particular, knowledge is required from the fields of mining and analysing massive data sets (Rajaraman and Ullman 2011; Hassani et al. 2014), high-performance computing (NAS 2013), and the new emerging discipline commonly referred to as "Data Science" (Schutt and O'Neill 2013). We expect to see some Dutch official statistics derived from Big Data in the coming years. When produced in a methodologically sound manner, official statistics based on Big Data can be cheaper, faster and more detailed than the official statistics known to date. For these endeavours to become successful, it is essential that they are supported by the general public and both Dutch and European legislation.

5. References

ASA. 2014. *Discovery With Data: Leveraging Statistics with Computer Science to Transform Science and Society*. July 2, 2014 version. Available at: http://www.amstat.org/policy/pdfs/BigDataStatisticsJune2014.pdf (accessed July 2014).

Beyer, M.A. and L. Douglas. 2012. *The Importance of 'Big Data': A Definition*. Gartner report, June version, ID Number: G00235055. Available at: http://www.gartner.com/it-glossary/big-data/ (accessed January 2013).

Breiman, L. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16: 99–231. Doi: http://dx.doi.org/10.1214/ss/1009213726.

Buelens, B., H.J. Boonstra, J. van den Brakel, and P. Daas. 2012. *Shifting Paradigms in Official Statistics: from Design-Based to Model-Based to Algorithmic Inference*. Discussion paper 201218, Statistics Netherlands, The Hague/Heerlen.

Buelens, B., P. Daas, J. Burger, M. Puts, and J. van den Brakel. 2014. *Selectivity of Big Data*. Discussion paper 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Cheung, P. 2012. *Big Data, Official Statistics and Social Science Research: Emerging Data Challenges*. Presentation at the December 19th World Bank meeting, Washington. Available at: http://www.worldbank.org/wb/Big-data-pc-2012-12-12.pdf (accessed January 2013).

Coosto. 2013. Main page. Available at: http://www.coosto.com/uk/ (accessed August 2013).

Daas, P.J.H. and M.J.H. Puts. 2014. *Social Media Sentiment and Consumer Confidence*. Paper for the Workshop on using Big Data for Forecasting and Statistics, April 7–8, Frankfurt, Germany. Available at: https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.pdf (accessed April 2015).

Daas, P.J.H., M. Roos, M. van de Ven, and J. Neroni. 2012a. *Twitter as a Potential Data Source for Statistics*. Discussion paper 201221, The Hague/Heerlen: Statistics Netherlands.

Daas, P., M. Tennekes, E. de Jonge, A. Priem, B. Buelens, M. van Pelt, and P. van den Hurk. 2012b. *Data Science and the Future of Statistics*. Presentation at the first Data Science NL meetup, Utrecht University, Utrecht. Available at: http://www.slideshare.net/pietdaas/data-science-and-the-future-of-statistics (accessed December 2012).

De Jonge, E., M. van Pelt, and M. Roos. 2012. *Time Patterns, Geospatial Clustering and Mobility Statistics Based on Mobile Phone Network Data*. Discussion paper 201214, The Hague/Heerlen: Statistics Netherlands.

De Jonge, E., J. Wijffels, and J. van der Laan. 2014. "ffbase: Basic Statistical Functions for Package ff. R package version 0.11.3." Available at: http://cran.r-project.org/web/packages/ffbase/index.html (accessed April 2015).

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.

Engle, R.F. and C.W.J. Granger. 1987. "Co-Integration and Error Correction: Representation, Estimation, and Testing." *Econometrica* 55: 251–276.

Eurostat. 2012. *Internet Access and Use*. Eurostat newsrelease 185/2012, December 18, 2012. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/4-18122012-AP/EN/4-18122012-AP-EN.PDF (accessed January 2013).

Flekova, L. and I. Gurevych. 2013. *Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media*. Paper for the evaluation lab on uncovering plagiarism, authorship, and social software misuse at Conference and Labs Evaluation Forum 2013, September 23–26, Valencia, Spain.

Fry, B. 2008. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. Sebastopol, CA: O'Reilly Media Inc.

Glasson, M., J. Trepanier, V. Patruno, P. Daas, M. Skaliotis, and A. Khan. 2013. *What does "Big Data" mean for Official Statistics?* Paper for the High-Level Group for the Modernization of Statistical Production and Services, March 10.

Golder, S.A. and M.W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures." *Science* 30: 1878–1881. Doi: http://dx.doi.org/10.1126/science.1202775.

Groves, R.M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly* 75: 861–871. Doi: http://dx.doi.org/10.1093/poq/nfr057.

Hassani, H., G. Saporta, and E. Sirimal Silvia. 2014. "Data Mining and Official Statistics: The Past, the Present and the Future." *Big Data* 2: 1–10. Doi: http://dx.doi.org/10.1089/big.2013.0038.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer Science + Business Media, LLC.

Lansdall-Welfare, T., V. Lampos, and N. Cristianini. 2012. "Nowcasting the Mood of the Nation." *Significance* 9: 26–28. Available at: http://www.significancemagazine.org/details/magazine/2468761/Nowcasting-the-mood-of-the-nation.html (accessed January 2013).

Lynch, C. 2008. "Big Data: How Do Your Data Grow?" *Nature* 455: 28–29. Doi: http://dx.doi.org/10.1038/455028a.

Manton, J.H., V. Krishnamurthy, and R.J. Elliott. 1999. "Discrete Time Filters for Double Stochastic Poisson Processes and Other Exponential Noise Models." *International Journal of Adaptive Control and Signal Processing* 13: 393–416.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. Report of the McKinsey Global Institute, McKinsey & Company.

NAS. 2013. *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press.

NDW. 2012. *The Database Explained. Brochure of the National Data Warehouse for Traffic Information, March*. Available at: http://www.ndw.nu/download_files.php?action=download_file&file_hash=209140a807e959f06646b0311f79de26 (accessed December 2012).

O'Connor, B., R. Balasubramanyan, B.R. Routledge, and N.A. Smith. 2010. *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series*. Carnegie Mellon University, Research Showcase. Available at: www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf (accessed April 2015).

R Development Core Team. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rajaraman, A. and J.D. Ullman. 2011. *Mining of Massive Datasets*. Cambridge: Cambridge University Press.

Schutt, R. and C. O'Neil. 2013. *Doing Data Science: Straight Talk from the Frontline*. Sebastopol, CA: O'Reilly Media.

Scott, S.L., A.W. Blocker, F.V. Bonassi, H.A. Chipman, E.I. George, and R.E. McCulloch. 2013. *Bayes and Big Data: The Consensus Monte Carlo Algorithm*. Bayes 250. Available at: http://www.rob-mcculloch.org/some_papers_and_talks/papers/working/consensus-mc.pdf (accessed April 2015).

Statistics Netherlands. 2013. *Consumer Confidence Survey*. Available at: http://www.cbs.nl/en-GB/menu/methoden/dataverzameling/consumenten-conjunctuur-onderzoek-cco.htm (accessed April 2013).

Struijs, P. and P.J.H. Daas. 2013. *Big Data, Big Impact?* Paper for the Seminar on Statistical Data Collection, September 25–27, Geneva. Switzerland.

Tennekes, M., E. de Jonge, and P.J.H. Daas. 2013. "Visualizing and Inspecting Large Datasets with Tableplots." *Journal of Data Science* 11: 43–58.

Van der Laan, J. 2013. *LaF: Fast Access to Large ASCII files*. R package version 0.5.

Zikopoulos, P., D. deRoos, K. Parasuraman, T. Deutsch, D. Corrigan, and J. Giles. 2012. *Harness the Power of Big Data*. New York: McGraw-Hill.

Received August 2013
Revised August 2014
Accepted September 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 249–262, http://dx.doi.org/10.1515/JOS-2015-0016

Big Data as a Source for Official Statistics

Piet J.H. Daas[1], *Marco J. Puts*[1], *Bart Buelens*[1], *and Paul A.M. van den Hurk*[1]

More and more data are being produced by an increasing number of electronic devices physically surrounding us and on the internet. The large amount of data and the high frequency at which they are produced have resulted in the introduction of the term 'Big Data'. Because these data reflect many different aspects of our daily lives and because of their abundance and availability, Big Data sources are very interesting from an official statistics point of view. This article discusses the exploration of both opportunities and challenges for official statistics associated with the application of Big Data. Experiences gained with analyses of large amounts of Dutch traffic loop detection records and Dutch social media messages are described to illustrate the topics characteristic of the statistical analysis and use of Big Data.

Key words: Large data sets; traffic data; social media.

1. Introduction

In our modern world, more and more data are generated on the web and produced by sensors in the ever-growing number of electronic devices surrounding us. The amount of data and the frequency at which they are produced have led to the introduction of the term 'Big Data' (Lynch 2008). Big Data sources can generally be described as: "high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making". This definition is a variant of the definition proposed by Gartner (Beyer and Douglas 2012). For more general information on Big Data and their innovative potential, the reader is referred to Manyika et al. (2011).

In addition to generating new commercial opportunities in the private sector, Big Data are potentially a very interesting data source for official statistics, either for use on their own, or in combination with more traditional data sources such as sample surveys and administrative registers (Cheung 2012). However, extracting relevant and reliable information from Big Data sources and incorporating it into the statistical production process is not an easy task (Daas et al. 2012a). Importantly, the statistical point of view has been underexposed in the work that has been "published" on Big Data so far; this work has been published mainly on weblogs and in conference and white papers. The majority of these publications have an IT perspective as they predominantly focus on soft- and hardware issues, and largely fail to address important statistical issues such as coverage, representativity, quality, accuracy and precision. If Big Data are to be used for official

[1] Statistics Netherlands, Division of Process development, IT and methodology P.O. Box 4481, 6401 CZ, Heerlen, The Netherlands. Emails: pjh.daas@cbs.nl (corresponding author), m.puts@cbs.nl, b.buelens@cbs.nl, and pamvandenhurk@gmail.com

statistics, it is essential that these issues are considered and adequately dealt with (Cheung 2012; Daas et al. 2012a; Glasson et al. 2013; Groves 2011).

In this article we provide an overview of the current state of the research on the usage of Big Data for official statistics at Statistics Netherlands and the lessons learned so far. In the next section a description of two Big Data case studies is given, followed by a more general methodological discussion in Section 3. Finally, conclusions are drawn in Section 4.

2. Big Data Case Studies

In this section we report on two Big Data case studies conducted at Statistics Netherlands. These studies serve as examples and allow for a more general formulation of the statistical issues and challenges involved with the application of Big Data in official statistics. All analyses were performed with the open-source software environment R (R Development Core Team 2012) on a Fujitsu Celsius M470-2 workstation with a 64-bit Windows 7 operating system, 32GB of RAM, 512 GB solid state drive and a 1 TB hard disk. Data were imported into R from CSV files which usually each contained one million rows of data. Each file was subsequently processed and analysed. Results were stored as CSV files. This approach was fast and flexible and sufficed for the studies described in this article.

2.1. Analysis of Traffic Loop Detection Data

Traffic loop detection data consist of measurements of traffic intensity. Each loop counts the number of vehicles per minute that pass at that location, and measures speed and vehicle length one. Such data are interesting for traffic and transport statistics and potentially also for statistics on other economic phenomena related to transport. On the particular day studied, data were collected at 12,622 measurement locations on Dutch roads. The data are stored centrally in the National Data Warehouse for Traffic Information (NDW) and managed by a collaboration of participating government organizations (NDW 2012). The National Data Warehouse contains historic traffic data collected from 2010 onwards. To determine the usability of the NDW data for statistics and to get an idea of its peculiar features, we started by studying minute-level data for all locations in the Netherlands for a single day: December 1st, 2011. The data set extracted from the NDW contained 76 million records, one million per CSV file, which were imported into R via the LaF package (Van der Laan 2013). This package supports loading the data in blocks, enabling the processing of enormous amounts of data without fitting all the data into memory.

Data were first aggregated over all loops, resulting in a series of total counts of all vehicles in the Netherlands at minute intervals. The change of this total count through the day is shown in Figure 1A. The overall profile displays clear morning and evening rush hour peaks around 8 am and 5 pm respectively. Importantly, however, there is a huge variation in the numbers of vehicles detected in subsequent minutes. This phenomenon is caused by the fact that – for a substantive number of minutes – data were only available for a subset of all detection loops in the country. This appeared to be caused by some computers failing to submit data to the warehouse at certain time points.

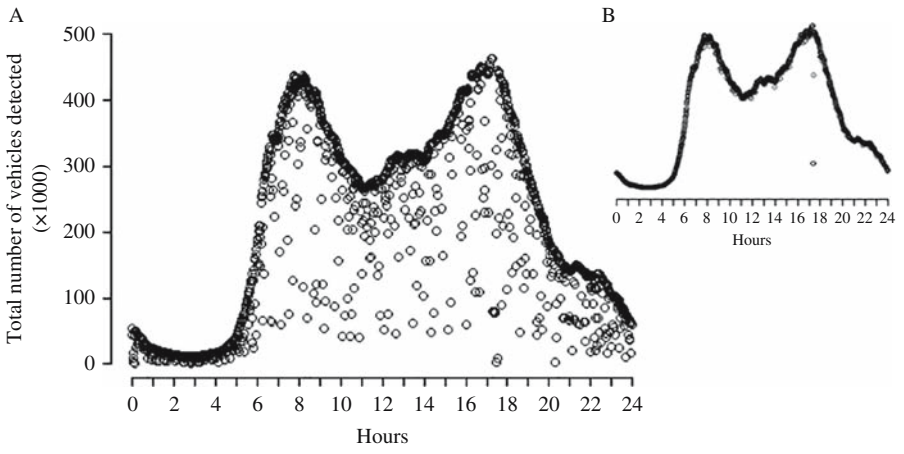


Fig. 1. (A) Total number of vehicles detected per minute in the Netherlands on December 1st, 2011. (B) Results after correcting for missing data.

From a statistical point of view there are various ways to solve such a missing data problem (De Waal et al. 2011).

Because aggregated data were used and this was our first experience with huge amounts of data, we opted for the simplest solution: add (impute) data reported by the same location during a short interval before or after the time point of missing data (if available). More specifically, a sliding, symmetrical five-minute time window to impute data at missing time points for the entire data set was applied. The resulting data pattern is shown in Figure 1B. Except for a period shortly after 5 pm, the majority of the missing data points were adequately replaced with timely data of the same measurement location. As a result of this data-editing procedure a total of nearly 35.8 million vehicle counts were added, which is slightly more than twelve percent of the number of vehicles originally counted, 294.7 million. Alternative model-based approaches can be applied and are preferred when traffic loop data are studied for smaller areas (more on this below).

The edited data set was used to create maps that indicate the number of vehicles for each measurement location for each time point by means of colour coding. Next, by sequencing these maps, a movie was created that displays the changes in vehicle counts for all locations during the day. Thus, this movie (not shown here) illustrates the increases and decreases in traffic intensity in the Netherlands throughout the day studied (Daas et al. 2012b). Unsurprisingly, the traffic intensity between the four major cities in the Netherlands (Amsterdam, Rotterdam, Utrecht and The Hague) was especially high, during all working hours and in the early evening.

Besides the total number of vehicles, the number of vehicles in various length categories was also studied. Because not all detection locations are able to differentiate between different vehicle lengths, only those that are able to do so were used. This subset consisted of 6,002 detection locations, which represented 48 percent of the total number of locations. Vehicles were sorted into three length categories: small (<= 5.6 metre), medium-sized (> 5.6 and <= 12.2 metre), and large (> 12.2 metre). Again, the imputed data set was used. Because the small vehicle category comprised around 75 percent of all vehicles detected, as compared to twelve percent for the medium-sized and 13 percent for the large

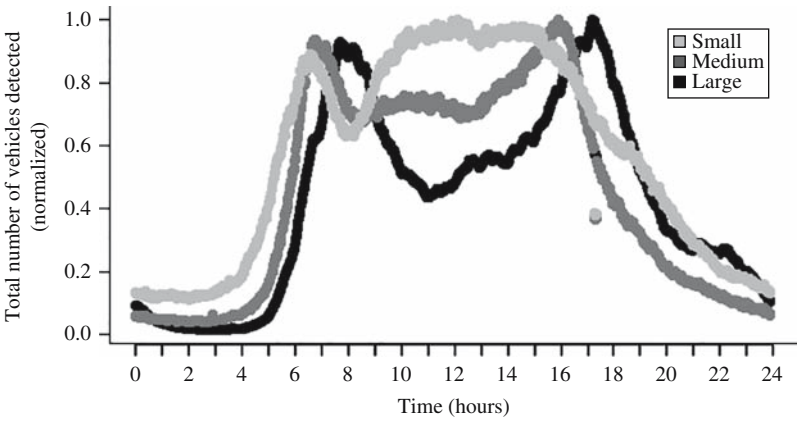


Fig. 2. Normalized numbers of vehicles detected in three length categories on December 1st, 2011 after correcting for missing data. Numbers of small (<= 5.6 meter), medium-sized (> 5.6 and <= 12.2 meter) and large vehicles (> 12.2 meter) are shown. Profiles are normalized by dividing by the maximum value of each series to more clearly reveal the differences. Maximum values are 119,523, 8,673 and 8,599 for small, medium and large vehicles, respectively.

vehicles categories, the normalized results for each category are shown in Figure 2. This figure illustrates the difference in driving behaviour between the three vehicle length categories. The small vehicle category displays clear morning and evening rush-hour peaks at 8 am and 5 pm respectively, in line with the overall profile described above (Figure 1). This finding is not unexpected, as this category of vehicles constitutes the vast majority of all vehicles. The medium-sized vehicles in turn have both an earlier morning and evening rush-hour peak, at 7 am and 4 pm respectively. Finally, the large vehicle category shows a clear morning rush-hour peak around 7 am and more dispersed driving behaviour during the remainder of the day; after 3 pm the number of large vehicles gradually declines without any apparent evening rush-hour peak. Most remarkable is the decrease in the relative number of medium-sized and large vehicles detected at 8 am, that is, during the morning rush-hour peak of the small vehicles. This may be caused by a deliberate attempt of the drivers of the medium-sized and large vehicles to avoid the morning rush-hour peak of the small vehicles or an effect of the more intense traffic (of small vehicles) around that time. Considering these differences, differentiation between vehicles of various lengths when creating a traffic index would not only enable more granular traffic statistics but can also provide more detailed information on transport and phenomena related to economic growth.

In addition to the analysis of traffic intensity at an aggregated level across all detection loops, the traffic intensity profile of a number of individual measurement locations was also studied, for example on highway A4 near Bergen op Zoom. The total number of vehicles detected at this location is shown in Figure 3. Detection at this location displays the same rush-hour peaks as in Figure 1. In addition, the characteristic volatile behaviour of traffic intensity data at the microlevel is shown. Given that this detection location does not suffer from missing data, the changes in the number of vehicles counted each minute are the result of real changes in the number of vehicles passing at this location. However, these rapid fluctuations are not very informative for the production of a traffic index,

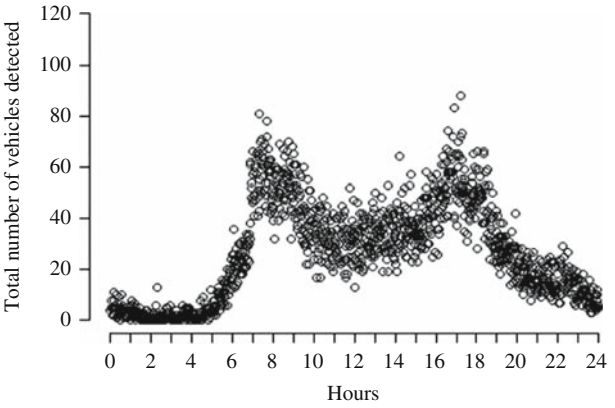


Fig. 3. Total number of vehicles counted by a detection location on highway A4 near Bergen op Zoom.

as interest is focused more upon gradual, long-period, changes, for example, weekly or monthly changes in the number of vehicles (of a certain length class) in a specific region. We are currently studying statistical modelling methods that can deal adequately with these kinds of Poisson-distributed data, such as Bayesian-based signal filters (Manton et al. 1999). These methods need to be applied in a reasonable time to the large amounts of loop detection data. The latter requires high-performance computing techniques (NAS 2013) when applied to the data of all loops in the whole country.

In the analyses in this section, we have assumed that all measurements are without error, except when entire records are missing. The missing records have been imputed using fixed values, not taking into account uncertainty associated with the imputation procedure. Alternatively, a multiple imputation approach could be implemented to account for such uncertainty, which would result in variances and confidence intervals for the aggregates shown above. The aggregates are obtained simply by summing individual loop counts. We have not conducted any form of inference or estimation (except for the imputations). In the future we may do so in order to obtain estimates that are representative of all Dutch highways, including those without traffic loops. A predictive modelling approach would need to be developed, resulting in estimated counts at locations without loops. This would lead to estimated aggregates and variance estimates reflecting the uncertainty of the estimation procedure.

2.2. *Analysis of Social Media Messages*

It is estimated that around 70 percent of the Dutch population actively posts messages on social media (Eurostat 2012). The three million or so Dutch messages generated each day (Daas and Puts 2014) may be an interesting data source for official statistics because they reflect many different aspects of our daily lives. We have studied two aspects of social media messages: content and sentiment. Studies of the content of Dutch Twitter messages – the dominant publicly available social medium in the Netherlands (see below) – revealed that nearly 50 percent of the messages are "pointless babble" (Daas et al. 2012a). In the remainder of the messages, spare-time activities, work, media (TV & radio) and

politics were predominantly discussed. This finding suggests that these messages could be used to extract opinions, attitudes, and sentiments towards these topics, opening up possibilities to collect a considerable amount of interesting information quickly without any response burden. The major problem in analysing social media messages is discriminating the informative from the noninformative ones. Because of the large share of the noninformative "babble" messages, usage of the more serious (informative) messages is negatively affected as many words of interest occur in both types of messages. Text mining approaches to automatically differentiate between both groups of messages have not been very successful so far (Daas et al. 2012a).

Another potential source of information in social media messages is their sentiment. Access to over 1.6 billion public messages written in Dutch from a large number of social media sites was obtained using an infrastructure provided by Coosto (2013). Public messages were sourced from the largest social media sites used by Dutch individuals, such as Twitter, Facebook, Hyves, Google+, and LinkedIn, as well as from numerous public Dutch weblogs and forums. The overall profile of the number of messages created per day revealed that from June 2010 onwards, increasing numbers of messages were generated in the Netherlands on a daily basis. The latter date corresponds to the period during which Coosto started to include huge numbers of Twitter messages in their Hadoop-based distributed database. We therefore used June 2010 as the starting date for our studies, with August 2012 as the end date. Messages could be selected from the database with a query language and a secure web interface. Coosto also determined the sentiment of each message by counting the number of positive and negative words following the general approach described in Golder and Macy (2011). Messages were classified as positive, negative or neutral depending on their overall score. A more detailed description of this part of the work can be found in Daas and Puts (2014).

Since several studies have been performed in English-speaking countries attempting to link the sentiment in social media to consumer confidence (O'Connor et al. 2010; Lansdall-Welfare et al. 2012) we were interested in studying this "relation" for the Netherlands. We looked at the sentiment in messages produced on the various platforms covered by the Coosto data set. The results were intriguing. The development of the sentiment in all Facebook messages produced during the period studied, nearly 170 million (almost ten percent of all messages produced), was found to correlate highly with consumer confidence; $r = 0.84$. Combing the sentiment of all Facebook and Twitter messages, slightly over 1.4 billion (close to 90% of all messages), with a linear model increased the correlation to $r = 0.88$. To reduce the risk of discovering spurious or false correlations, the series were additionally checked for cointegration. Cointegration provides a stronger argument as it checks for a common stochastic drift, indicating that series exhibit fluctuations around a common trend (Engel and Granger 1987). Here, it was found that the sentiment in Facebook and the combination of Facebook and Twitter both cointegrated with consumer confidence, suggesting a strong association between the developments in both series. Remarkably, the sentiment in Twitter messages only correlated less, $r = 0.61$, and did not cointegrate.

Figure 4 displays the survey-based Consumer Confidence series (Statistics Netherlands 2013) and the corresponding Dutch social media sentiment findings for the period studied. Both series relate quite well. This association is remarkable, as the

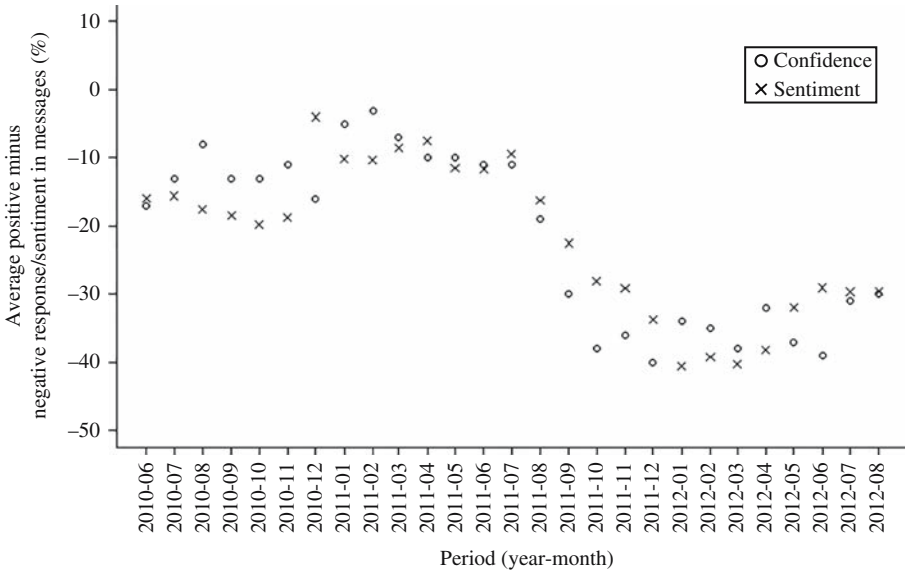


Fig. 4. Comparison of Dutch consumer confidence (○) and the sentiment in Dutch Facebook and Twitter messages on a monthly basis (×). A correlation coefficient of $r = 0.88$ is found for both series.

populations from which the data are obtained are very different. Dutch consumer confidence is obtained from a random sample from the population register, with around 1,000 persons responding each month. Due to sampling variance, the standard errors of the consumer confidence series shown in Figure 4 are on average approximately 2.0. The sentiment in Dutch Facebook and Twitter messages is derived from around 52 million *messages* generated each month. These messages are created by a considerable part of the population, 70 percent according to Eurostat (2012), but i) not all social media messages created in the Netherlands are written in Dutch and ii) different users post varying numbers of messages on various platforms. We have not attempted to estimate the sentiment of the (unknown) subpopulation who does not contribute to social media platforms. Consequently, our social media sentiment series is not subject to sampling, modelling or prediction uncertainty, but may be biased because of differences in the composition of the Dutch population and those active on social media. Previous work by Daas et al. (2012a) also revealed that the number of Twitter messages can vary from 200 per day to not even one message a month for a single person. More recent work has confirmed that the association between both series remains stable over time and that consumer confidence and social media sentiment are related from a Granger-causality perspective (more in Daas and Puts 2014).

3. Discussion

The two case studies described in this article reveal several issues that need to be addressed before Big Data can become a useful and reliable data source for the field of official statistics. These issues, the most important considerations, the way we have dealt with them and the lessons learned are discussed below.

3.1. Data Exploration

Typically, Big Data sets are made available to us, rather than designed by us. As a consequence, their contents and structure need to be understood prior to using the data for statistical analysis (Hassani et al. 2014). This first step is called data exploration, which is aimed at revealing data structure and patterns and, no less important, at assessing the quality of the data as revealed by the presence of errors, anomalies and missing data. Visualisation methods have been proven to be very insightful for such tasks (Fry 2008; Zikopoulos et al. 2012, ch. 7). Recently, certain visualisation methods have been developed that are particularly suited to the exploration of Big Data. Examples are tableplots (Tennekes et al. 2013) to display Big Data with many variables and 3D heatmaps to study variability in multivariate continuous data (Daas et al. 2012b). Sequencing 2D plots into animations is useful to visualise temporal and other aspects of Big Data (Daas et al. 2012b).

3.2. Missing Data

By studying the traffic intensity data on a minute-by-minute level, we discovered that part of the data were missing. If we had analysed the data aggregated at hourly or daily levels, we would have reduced the amount of data studied but would not have noticed that missing data is such a big problem. Since Statistics Netherlands plans to use NDW data to produce reliable traffic and transport intensity statistics at a detailed level, the missing data problem needs to be solved. Missing data is not a problem unique to the traffic loop data set, as other data sources are susceptible to missing data too. For instance, server downtime and network outages can lead to missing social media messages or mobile phone data. However, in the end, the time spent on processing also needs to be reduced to a manageable level to enable the production of frequent statistics. Currently statistical models are being explored that are able to cope adequately with missing data and can be applied to enormous amounts of data in a reasonable amount of time. For such an approach to be successful, the combination of the IT infrastructure available and the ease with which a modelling method can be upscaled needs to be assessed (NAS 2013). We are currently focusing on Bayesian approaches as these are applied to enormous amounts of data in other areas of science and are well suited to capturing various forms of uncertainty. The high-performance computing needs can be met at relatively low cost by using the large amounts of computing power provided by the graphics processing units available on many modern graphics cards (Scott et al. 2013).

3.3. Volatility

The number of vehicles detected by individual loops fluctuates considerably from minute to minute. These fluctuations are caused by real changes in the number of vehicles detected but are not very informative from a statistical point of view as they occur at too high a time resolution. Similarly, sentiment analyses on a daily and weekly basis suffer from a volatility that is not seen at monthly intervals (Daas and Puts 2014; O'Connor et al. 2010). It is therefore recommended to develop statistical methods able to cope with volatile behaviour. Possible methods under consideration are the application of moving averages and advanced filtering techniques (e.g., a Kalman filter or time-series modelling).

3.4. Selectivity

The analyses described in Section 2 apply to traffic intensity on roads equipped with traffic loop sensors, and to the sentiment analysis of people who post Dutch Facebook or Twitter messages on social media websites. It is important to realize that both data sets are created by only a subset of the total population in the Netherlands: only vehicles driving on the major Dutch roads were counted and only the sentiment of a subset of all people in the Netherlands was probed, respectively. The subpopulations from which these Big Data sources were derived are not typical target populations for official statistics. Therefore the data are likely to be selective and not representative of a target population of interest. In addition, both sources contain data resulting from the registration of events. These are vehicles passing and messages sent respectively. Both lack directly available data on the units of interest. Usually, the representativity of Big Data can be assessed through the careful comparison of characteristics of the covered population and the target population. Unfortunately, this may prove problematic for these sources, as hardly any such characteristics are available to conduct such a comparison (Buelens et al. 2014). For instance, vehicles can not be uniquely identified in the traffic loop data as licence plate data are absent. Little is known about the people posting on social media; often only their username is known but not their age or gender. In situations where at least some background information is available, the selectivity issue can be assessed and probably resolved. Alternatively, profiling approaches could be used to extract features to estimate, for instance, the chance that a user is male or female (Flekova and Gurevych 2013). Perhaps this could be achieved through predictive modelling, using a wide variety of algorithms known from statistical learning and data mining techniques (Hastie et al. 2009). These are modelling methods not traditionally used in official statistics. Buelens et al. (2012) explore some possibilities for applications of data mining methods in official statistics. More on this topic can be found in ASA (2014).

3.5. Legal Considerations

Privacy and security are issues that may impede NSIs' use of Big Data. In contrast to the legal basis that permits the use of administrative data sources by a lot of NSIs, the use of privately owned Big Data, such as mobile phone data, needs to be specifically arranged (De Jonge et al. 2012). But even for publicly accessible data, such as price and product information on websites, questions of ownership and purpose of publication can be raised. And even if there are no legal impediments, public perception is a factor that must be taken into account. These concerns have to be taken seriously and tackled one at a time. Fortunately, there are measures that can be taken to overcome at least some of the obstacles, for example, by anonymizing unique identifiers, removing the privacy-sensitive part of a Global Positioning System track (e.g., the first and last 100 metres) or by using informed consent. If a reduction of response burden can be offered, this can be very helpful, also in getting the support of the general public. In the long run, changes in legislation may be considered, to ensure continuous data access for official statistics. But it remains important to stay in line with public opinion, because credibility and public trust are important assets. Within the European Union, changes in European legislation must also be considered. In addition to national laws, European laws or regulations can impede

the collection of data, even if the current Dutch legislation does not present any problem (more in Struijs and Daas 2013).

3.6. Data Management

Long-term stability may also be a problem when using Big Data. Typically, statistics for policy making and evaluation are required for extended periods of time, often covering many years. The Big Data sources encountered so far seem subject to frequent modifications, possibly limiting their long-term use. This suggests a need for more flexible data processing and evaluation strategies, which will have to put more emphasis on ongoing data and metadata management to identify, describe and re-evaluate new sources. Data management is also affected by data ownership, copyright and the purpose for which data are registered. Privately owned Big Data in particular may need special arrangements and will probably also incur costs. The two data sources discussed in this article are examples of each. The organisation that maintains traffic loop data is funded by the Dutch government which means that, as laid down in the Statistics Netherlands Act, they provide us with the data free of charge. For the study of social media data, however, costs were incurred, as the data are collected by a privately owned company.

3.7. High-Performance Computing

Processing enormous amounts of data within a reasonable amount of time requires dedicated and specialized computing infrastructures. Hence it can be expected that the inclusion of Big Data as a source for official statistics will certainly affect the IT environment of NSIs. Our experiences so far, however, reveal that considerable progress can be made even with a limited budget. Having a secure computer environment with many fast processors, large amounts of RAM and fast disk access certainly helps. Several important considerations are described in NAS (2013) and Schutt and O'Neil (2013). Parallel processing is the way to speed things up. For instance, we have found that processing traffic loop data in parallel in R results in a 17x speedup over the original (serial) processing time. We are currently using (multicore) general-purpose computing on graphics processing units and are looking at distributed computing, such as our own secure local cluster.

3.8. New Skills Needed

In order to work with Big Data specific technical expertise is needed, such as knowledge of advanced (high-performance) computing and data engineering. These skills speed up the ease with which Big Data can be incorporated into the statistical process and the way it is analysed. In our office, Big Data is usually processed with R or Python. Besides knowing the language, the most important skill here is knowing how to write a program that is able to access and analyse all the data within a reasonable amount of time. Several of our colleagues have written R packages specifically devoted to these tasks, such as LaF (Van der Laan 2013) and ffbase (De Jonge et al. 2014). Moreover, the models used for Big Data must be able to address the levels of complexity that huge data sets can reveal. This makes many of the standard approaches used in official statistics limited in utility and

performance (NAS 2013). The algorithmic-oriented models developed in fields outside statistics might be more applicable here (Breiman 2001; Hastie et al. 2009).

Perhaps just as important is the attitude of the people involved. Working with Big Data requires an open mindset and the ability not to see all problems *a priori* in terms of sampling theory, as Big Data are more similar to large sets of observational data (Daas and Puts 2014). The term "data scientist" has been coined for researchers with the skills identified above (Schutt and O'Neill 2013). We have solved the initial need by including experimentally trained researchers in our Big Data efforts, as they are more practically oriented and are more accustomed to deriving theory from data. The strict difference between methodology, software engineering and IT hardware expertise commonly used at NSIs is also becoming less well defined. At Statistics Netherlands, a group of data scientists is currently being formed. The work of this group is expected to be beneficial to many of the areas of official statistics, especially when large data sources and complex models are used.

4. Conclusions

The official statistics community can greatly benefit from the possibilities offered by Big Data. However, care is needed when trying to implement these sources in official statistics. The two Big Data case studies described show typical issues including missing data, volatility and selectivity, which all need to be adequately dealt with. For this reason, investment in specific research and skills development is needed. In addition, various new areas of expertise are considered necessary to fully exploit the information contained in Big Data. In particular, knowledge is required from the fields of mining and analysing massive data sets (Rajaraman and Ullman 2011; Hassani et al. 2014), high-performance computing (NAS 2013), and the new emerging discipline commonly referred to as "Data Science" (Schutt and O'Neill 2013). We expect to see some Dutch official statistics derived from Big Data in the coming years. When produced in a methodologically sound manner, official statistics based on Big Data can be cheaper, faster and more detailed than the official statistics known to date. For these endeavours to become successful, it is essential that they are supported by the general public and both Dutch and European legislation.

5. References

ASA. 2014. *Discovery With Data: Leveraging Statistics with Computer Science to Transform Science and Society*. July 2, 2014 version. Available at: http://www.amstat.org/policy/pdfs/BigDataStatisticsJune2014.pdf (accessed July 2014).

Beyer, M.A. and L. Douglas. 2012. *The Importance of 'Big Data': A Definition*. Gartner report, June version, ID Number: G00235055. Available at: http://www.gartner.com/it-glossary/big-data/ (accessed January 2013).

Breiman, L. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16: 99–231. Doi: http://dx.doi.org/10.1214/ss/1009213726.

Buelens, B., H.J. Boonstra, J. van den Brakel, and P. Daas. 2012. *Shifting Paradigms in Official Statistics: from Design-Based to Model-Based to Algorithmic Inference*. Discussion paper 201218, Statistics Netherlands, The Hague/Heerlen.

Buelens, B., P. Daas, J. Burger, M. Puts, and J. van den Brakel. 2014. *Selectivity of Big Data*. Discussion paper 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Cheung, P. 2012. *Big Data, Official Statistics and Social Science Research: Emerging Data Challenges*. Presentation at the December 19th World Bank meeting, Washington. Available at: http://www.worldbank.org/wb/Big-data-pc-2012-12-12.pdf (accessed January 2013).

Coosto. 2013. Main page. Available at: http://www.coosto.com/uk/ (accessed August 2013).

Daas, P.J.H. and M.J.H. Puts. 2014. *Social Media Sentiment and Consumer Confidence*. Paper for the Workshop on using Big Data for Forecasting and Statistics, April 7–8, Frankfurt, Germany. Available at: https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.pdf (accessed April 2015).

Daas, P.J.H., M. Roos, M. van de Ven, and J. Neroni. 2012a. *Twitter as a Potential Data Source for Statistics*. Discussion paper 201221, The Hague/Heerlen: Statistics Netherlands.

Daas, P., M. Tennekes, E. de Jonge, A. Priem, B. Buelens, M. van Pelt, and P. van den Hurk. 2012b. *Data Science and the Future of Statistics*. Presentation at the first Data Science NL meetup, Utrecht University, Utrecht. Available at: http://www.slideshare.net/pietdaas/data-science-and-the-future-of-statistics (accessed December 2012).

De Jonge, E., M. van Pelt, and M. Roos. 2012. *Time Patterns, Geospatial Clustering and Mobility Statistics Based on Mobile Phone Network Data*. Discussion paper 201214, The Hague/Heerlen: Statistics Netherlands.

De Jonge, E., J. Wijffels, and J. van der Laan. 2014. "ffbase: Basic Statistical Functions for Package ff. R package version 0.11.3." Available at: http://cran.r-project.org/web/packages/ffbase/index.html (accessed April 2015).

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.

Engle, R.F. and C.W.J. Granger. 1987. "Co-Integration and Error Correction: Representation, Estimation, and Testing." *Econometrica* 55: 251–276.

Eurostat. 2012. *Internet Access and Use*. Eurostat newsrelease 185/2012, December 18, 2012. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/4-18122012-AP/EN/4-18122012-AP-EN.PDF (accessed January 2013).

Flekova, L. and I. Gurevych. 2013. *Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media*. Paper for the evaluation lab on uncovering plagiarism, authorship, and social software misuse at Conference and Labs Evaluation Forum 2013, September 23–26, Valencia, Spain.

Fry, B. 2008. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. Sebastopol, CA: O'Reilly Media Inc.

Glasson, M., J. Trepanier, V. Patruno, P. Daas, M. Skaliotis, and A. Khan. 2013. *What does "Big Data" mean for Official Statistics?* Paper for the High-Level Group for the Modernization of Statistical Production and Services, March 10.

Golder, S.A. and M.W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures." *Science* 30: 1878–1881. Doi: http://dx.doi.org/10.1126/science.1202775.

Groves, R.M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly* 75: 861–871. Doi: http://dx.doi.org/10.1093/poq/nfr057.

Hassani, H., G. Saporta, and E. Sirimal Silvia. 2014. "Data Mining and Official Statistics: The Past, the Present and the Future." *Big Data* 2: 1–10. Doi: http://dx.doi.org/10.1089/big.2013.0038.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer Science + Business Media, LLC.

Lansdall-Welfare, T., V. Lampos, and N. Cristianini. 2012. "Nowcasting the Mood of the Nation." *Significance* 9: 26–28. Available at: http://www.significancemagazine.org/details/magazine/2468761/Nowcasting-the-mood-of-the-nation.html (accessed January 2013).

Lynch, C. 2008. "Big Data: How Do Your Data Grow?" *Nature* 455: 28–29. Doi: http://dx.doi.org/10.1038/455028a.

Manton, J.H., V. Krishnamurthy, and R.J. Elliott. 1999. "Discrete Time Filters for Double Stochastic Poisson Processes and Other Exponential Noise Models." *International Journal of Adaptive Control and Signal Processing* 13: 393–416.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. Report of the McKinsey Global Institute, McKinsey & Company.

NAS. 2013. *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press.

NDW. 2012. *The Database Explained. Brochure of the National Data Warehouse for Traffic Information, March*. Available at: http://www.ndw.nu/download_files.php?action=download_file&file_hash=209140a807e959f06646b0311f79de26 (accessed December 2012).

O'Connor, B., R. Balasubramanyan, B.R. Routledge, and N.A. Smith. 2010. *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series*. Carnegie Mellon University, Research Showcase. Available at: www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf (accessed April 2015).

R Development Core Team. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rajaraman, A. and J.D. Ullman. 2011. *Mining of Massive Datasets*. Cambridge: Cambridge University Press.

Schutt, R. and C. O'Neil. 2013. *Doing Data Science: Straight Talk from the Frontline*. Sebastopol, CA: O'Reilly Media.

Scott, S.L., A.W. Blocker, F.V. Bonassi, H.A. Chipman, E.I. George, and R.E. McCulloch. 2013. *Bayes and Big Data: The Consensus Monte Carlo Algorithm*. Bayes 250. Available at: http://www.rob-mcculloch.org/some_papers_and_talks/papers/working/consensus-mc.pdf (accessed April 2015).

Statistics Netherlands. 2013. *Consumer Confidence Survey*. Available at: http://www.cbs.nl/en-GB/menu/methoden/dataverzameling/consumenten-conjunctuur-onderzoek-cco.htm (accessed April 2013).

Struijs, P. and P.J.H. Daas. 2013. *Big Data, Big Impact?* Paper for the Seminar on Statistical Data Collection, September 25–27, Geneva. Switzerland.

Tennekes, M., E. de Jonge, and P.J.H. Daas. 2013. "Visualizing and Inspecting Large Datasets with Tableplots." *Journal of Data Science* 11: 43–58.

Van der Laan, J. 2013. *LaF: Fast Access to Large ASCII files*. R package version 0.5.

Zikopoulos, P., D. deRoos, K. Parasuraman, T. Deutsch, D. Corrigan, and J. Giles. 2012. *Harness the Power of Big Data*. New York: McGraw-Hill.

Received August 2013
Revised August 2014
Accepted September 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 283–304, http://dx.doi.org/10.1515/JOS-2015-0018

Sentiments and Perceptions of Business Respondents on Social Media: an Exploratory Analysis

Vanessa Torres van Grinsven[1] *and Ger Snijkers*[2]

The perceptions and sentiments of business respondents are considered important for statistical bureaus. As perceptions and sentiments are related to the behavior of the people expressing them, gaining insights into the perceptions and sentiments of business respondents is of interest to understand business survey response. In this article we present an exploratory analysis of expressions in the social media regarding Statistics Netherlands. In recent years, social media have become an important infrastructure for communication flows and thus an essential network in our social structure. Within that network participants are actively involved in expressing sentiments and perceptions. The results of our analysis provide insights into the perceptions and sentiments that business respondents have of this national statistical institute and specifically its business surveys. They point towards the specific causes that have led to a positive or a negative sentiment. Based on these results, recommendations aimed at influencing the perceptions and sentiments will be discussed, with the ultimate goal of stimulating survey participation. We also suggest recommendations regarding social media studies on sentiments and perceptions of survey respondents.

Key words: Business survey communication; survey participation; response motivation; expressions; social media.

1. Introduction

In the Netherlands, Statistics Netherlands (hereinafter SN) is responsible for publishing official statistics to be used in practice for policy making and scientific research. Data on a large variety of topics are collected and processed for the production of statistics. To obtain its data, SN established a policy on data collection a decade ago; the most recent update to this policy was carried out in 2011 (Snijkers et al. 2011). Although surveys are still an important way of collecting data, secondary data need to be used before a survey is considered. Secondary data include administrative data, and in the future will very likely also include big data (Groves 2011; Daas et al. 2013). This data collection strategy applies to both social and business statistics. One major driver in the implementation of this strategy was and still is the reduction of response burden, that is, the compliance costs for businesses. Over the last two decades, SN has reduced its actual response burden by about

[1] Faculty of Social Sciences, Utrecht University, Padualaan 14, 3584 CH, Utrecht; Statistics Netherlands, CBS-weg 11, 6412 EX, Heerlen, Netherlands. Email: torresvangrinsven@gmail.com
[2] Statistics Netherlands, CBS-weg 11, 6412 EX, Heerlen, Netherlands. Email: gsks@cbs.nl
Acknowledgment: We would like to thank Marleen Verbruggen and Piet Daas of Statistics Netherlands for their contributions. Without their support this study would not have been possible. We would also like to thank the editors and the referees for their helpful comments.

70 percent (Snijkers 2008). In addition to its responsibility for official national statistics, SN also has the task of producing European (community) statistics.

Some of the official business surveys request financial data, like the annual Structural Business Survey (SBS), but others request other kinds of data, for example, movements (the Traffic and Transport Surveys), ICT use within a company (the ICT survey), and international trade (International Trade Survey). Some businesses are sampled for multiple surveys, and for recurring surveys they receive a questionnaire for a number of waves, because of their importance for the statistics. On a yearly basis, about a million questionnaires are sent to businesses.

We know that some businesses are not fond of these surveys, but most of them comply. An important reason for compliance is the fact that most of these surveys are mandatory (Torres van Grinsven et al. 2014; Snijkers et al. 2013). In the Netherlands, sentiments about official business surveys sent out by SN have been expressed in traditional media, for instance in newspapers, and in publications by business organizations, among others. For example, in a 2006 publication of the Dutch Employers' Union in the provinces of Brabant and Zeeland (Brabants Zeeuwse Werkgeversvereniging) it is stated for example that for surveys "the costs outweigh the added value" (van Vroenhoven 2006, 23). Some businesses express their views on surveys by sending letters and e-mails to Statistics Netherlands, or contacting its help desk by phone. In the past, politicians have also expressed their sentiments on official business surveys.

Research on the sentiments of business respondents towards official business surveys is not new. At SN, Customer Satisfaction Surveys of respondents have been conducted to study these views (for an analysis of these data, see, Giesen 2012). An overview of (data on) these sentiments that concern the official business surveys conducted by Statistics Netherlands has been presented by Snijkers et al. (2007). These studies will be discussed in the concluding section of the present article.

Collecting and analyzing these sentiments has been cumbersome because the data needed to be collected from various sources. The gathering of the appropriate data that measured these sentiments was not self-evident for any of these sources. However, with the expanding usage of social media, data on perceptions and sentiments have become more readily available. Recent research estimates that around 60 percent of the Dutch population actively posts messages on social media (Daas and Puts 2014). The huge amount of Dutch messages (Coosto 2014) may thus be an interesting data source for the analysis of perceptions and sentiments of business survey respondents in the Netherlands, and might be an appropriate replacement for the data formerly used.

In this article we present the results of an analysis of expressions in social media by business respondents about SN, its business surveys and questionnaires. This analysis can be characterized as exploratory, as it is not used to test hypotheses, but to explore how many expressions have been posted on social media and the content of these messages. For this study, all available data (expressions in social media) within a specific period of time have been used, as we will discuss in Subsection 2.1. As such, this study can be characterized as an analysis of "organic" data, as opposed to "designed" data created by survey research (Groves 2011). Researching perceptions and sentiments through a survey, for example, is inherently different from the analysis we have done on social media data. The expressions in social media data can be genuinely defined as "texts": they are "words

and images that have become recorded without the intervention of a researcher" (Silverman 2000, 825). In a survey or an interview, the researchers' preconceptions always strongly influence the categories of topics that are revealed. When the data analyzed are "texts" as defined by Silverman, it is much more likely that original participants' categories will be discovered.

This analysis indicates sentiments and perceptions respondents have of SN, or, in another words, meanings these respondents attribute to SN and its actions. In addition, it also provides insights into sources that cause these perceptions and sentiments. Analyzing these expressions in depth unveils both major irritations and positive attitudes, and this knowledge can in turn be used to improve the design of communication strategies and surveys in such a way that the perception businesses and business respondents have of a national statistical institute (hereinafter NSI) will be influenced in a positive way.

The sentiments and perceptions of business respondents are relevant to business surveys because it is proposed that these are related to survey participation, as well as the respondent's behavior when completing a questionnaire, via perceived response burden and the motivation to respond (as is discussed by Giesen 2012; Torres van Grinsven et al. 2014; Snijkers and Jones 2013; Willimack and Snijkers 2013; Haraldsen et al. 2013). Consequently, this affects the quality of the resulting survey data (see e.g., Wenemark et al. 2011; Haraldsen 2013) as well as the cost efficiency of survey data collection (Snijkers and Jones 2013). Blumer (1973) noted that human behavior results from a vast interpretive process in which people, both individually and collectively, guide themselves by defining and evaluating the objects, events and situations they encounter. This is another way of saying that business survey response behavior is affected by business respondents' interpretation of the NSI and the survey request – or the perception they have of the NSI and the survey request. Social media are used more and more to express these perceptions and sentiments. The internet and social media have developed into a new, vast communicative infrastructure and cultural forum (Jensen and Helles 2011), in which social actors as communicators become sources of information themselves (Jensen 2012) and are actively involved in expressing perceptions and images. Moreover, business survey respondents may be active on social media, which makes social media an interesting data source to explore when researching the perceptions and sentiments of business survey respondents, especially if one is interested in a better understanding of these respondents' behavior.

This study has two explorative research questions. The first question is: what can we find in the social media, and is the study of messages on social media useful for SN in understanding business respondents' views on SN and its surveys? Second, we aim at gaining insights into the content of these expressions: the perceptions or images they reflect, the sentiments expressed, and the causes for both positive and negative sentiments.

That is, first, we explored social media as a new kind of data source to find out if researching social media messages can lead to useful results; and second, we explored these expressions on social media to find out what sentiments and perceptions they reflect, and what the causes of these sentiments and perceptions are.

Based on the first research question, the following research objectives are discussed in this article: what is the number of expressions about SN and its surveys and questionnaires on social media and on web fora? Are there fluctuations in these numbers over time?

The aim here is to see whether there is any connection with the dispatching of questionnaires or other events. The second research question led to the following research objectives: what are the sentiments, are they negative or positive in nature? Which perceptions (ideas or images) do business respondents have of SN, its surveys and questionnaires as shown in these expressions? What aspects of the survey do these expressions relate to, and which aspects do people complain or write positively about?

2. Data and Method

2.1. Data

The data source used for this study is a database named Coosto, a social media monitor operating in the Netherlands. In this database, virtually all public posts in the Dutch language on Dutch social media, web fora and weblogs have been structurally collected and stored since January 2009 (Coosto 2014). Since August 2010 this has included all posts on Twitter, so-called "tweets". Currently, this database contains more than a billion entries, and each day about 3.2 million entries are added. At the time of data extraction (August 2012), more than 390,000 different social media channels were used, among which the most important were Twitter, Facebook, Hyves, and Google+. Studies of the content of Dutch Twitter messages – the dominant social medium in the Netherlands – revealed that nearly 50 percent of the messages were composed of "pointless babble" (Daas et al. 2012). This makes the main problem in social media research discriminating the informative from the noninformative messages. The large share of the noninformative "babble" messages negatively affects the use of the more serious informative messages (Daas et al. 2013).

We addressed this problem by devising a restrictive search in the Coosto database, which resulted in a selection of posts to be analyzed. The selection was based on posts for the period January 2009–August 2012. Keywords for the search included a combination of different denotations of the NSI, and a set of survey-related keywords, like "survey", "questionnaire", "letter", "response", "fine", "obligation", "mandatory", "administrative burden", the name of major surveys, and so on. The resulting data set, however, still included a lot of "noise" and irrelevant records (see, Daas and Puts 2014, 26). We then restricted the posts to those strictly relevant to business respondents and business surveys. Double entries that resulted from the overlapping original sets of queries were taken care of. Entries from large news sites and Twitter and Facebook news accounts were left out, as our interest was on more personal sentiments and dialogues, stemming from business respondents themselves. The few posts or tweets by SN or the Dutch Ministry of Economic affairs were left out as well. Posts that clearly referred to household surveys (114 posts) and posts that did not clearly speak about business surveys (50 posts) were left out as well, as we were only interested in sentiments and perceptions of business respondents towards business surveys in particular. This resulted in 477 posts that were clearly about business surveys and written by business respondents; these posts were analyzed in this study. These procedures can be characterized as the selection of posts based on the relevance to our research objectives. The resulting data set of 477 posts can be defined as the population of public posts with regard to our research questions for the period

January 2009–August 2012; furthermore, in our study no sample was drawn from this population and instead we analyzed the whole population.

2.2. *Methods of Analysis*

The selected posts were analyzed using a sequential two-step mixed-method design. First, a word count or "lexical analysis" was carried out, followed by a qualitative thematic analysis.

In a lexical analysis (also called "word counts" or "concordance analyses") a word list is created which can be seen as *concentrated* or distilled data (Tesch 1990, 138–139). This enables the exploration and objective identification of central themes in large bodies of text. The words from the word list are therefore clustered into meaningful categories of words with shared semantic fields in a process analogous to the development of a coding scheme for the interpretative qualitative analysis of text. Interpretive researcher input is thus required to a certain extent in certain steps of the analysis, as the "lemmas" or categories in the classification of words with a similar meaning are constructed by the researcher. Nevertheless, using this technique makes the analysis more inductive than a purely qualitative analysis, as the researchers construct the categories after the identification of the word and production of the word list.

Lexical analysis is based on an innovative approach to using software originally designed for "corpus linguistics analysis" (CL) (e.g., Adolphs et al. 2004; Seale et al. 2006). Corpus linguistics is the analysis of large collections of stored, naturally occurring texts, and is typically used to examine discourses, that is, to examine texts as the representation of a certain world view or perception. This type of analysis has been used as an effective approach for quantitative analyses of large volumes of texts in the traditional media (Tesch 1990; Leech 1992; Gabrielatos and Baker 2008), as well as postings on social media like Twitter and web fora (e.g., Seale et al. 2006). Lexical analysis has also been proposed as a suitable method for analyzing qualitative textual data (e.g., Ryan and Weisner 1996; Jehn and Doucet 1996, 1997), as it is more inductive than conventional qualitative approaches. It seems, however, to be especially suited to the conjoint qualitative (thematic) and quantitative analysis of large bodies of texts (e.g., Seale et al. 2006; Gabrielatos and Baker 2008).

One key concept in lexical analysis is the notion of *collocates*. Collocates are two or more words that regularly co-occur. In this study we focus on the collocates of the NSI, which we define as all words appearing in the selected posts (as all selected and analyzed posts contain a reference to the NSI). All the words examined and the resulting word lists are thus collocates of the NSI. Many of those are an evaluative expression. In the context of lexical analysis, the examination of collocates can not only provide a "semantic analysis of a word" (Sinclair 1991, 1156-116) but also contributes to revealing its attributed meaning (e.g., Nattinger and DeCarrico 1992).

An analysis of collocates reveals the attitude or perception expressed (Gabrielatos and Baker 2008), in this case about the NSI. These attitudes and perceptions are subjective in nature. The aspect of subjectivity is taken into account in lexical analysis by making clear that the frequent use of particular collocates may result in particular *meaning attributes* being associated with the NSI that may be subjective and are not necessarily elements of

the NSI's nature (Gabrielatos and Baker 2008). Moreover, words that at a first glance may seem descriptive can also be used in an evaluative way (be it positive or negative). This is an important notion, because in this study we are not interested in objective descriptions of the NSI, its surveys, the survey questionnaires, and the official statistics, but in the subjective perceptions and sentiments that the "speakers" have of these. Collocates thus provide information on the most frequent ideas associated with an entity or phenomenon; for example, our lexical analysis shows that the NSI is associated, among other things, with technical issues and "having to".

For the lexical analysis, the software program Concordance was used (see, www.concordancesoftware.co.uk/). The Concordance tool allows researchers, among others, to count words, cluster words into categories, and view and sort collocates. This software tool allows a more inductive approach to the formulation of coding categories than some other text analysis programs (e.g., Pennebaker et al. 2001) that rely on prespecified categories. In this project, the Concordance software was preferred over another frequently used tool, the WordSmith Tool, as the latter is primarily used to calculate keywords in texts on the basis of comparing the corpus with a reference corpus, which we did not have in this study.

In a second step, following the lexical analysis, we carried out a thematic (discourse) analysis. In this step the sentiment of the posts (positive/negative/neutral) as well as the themes or topics present in the posts were coded (e.g., Ryan and Bernard 2003; Braun and Clarke 2006). In line with Silverman's (2000) definition of a "text", discourse analysis focuses on how different versions of the world are produced through the use of a discourse. Accordingly, we were interested in the representations of SN as displayed in the postings we analyzed.

We made the decision to make our interpretation and coding scheme as objective as possible. Concerning the sentiment of the posts, the requirement for a post to be coded as negative was that the post contained a clear and objectively definable word or sentence as indicator of a negative evaluation of SN or a survey or something else "sent out" by SN (e.g., a survey, a reminder, a telephone call). For example, "failure", "bad", "☹", "Aargh", and so on, are indicators for negative sentiments. The same holds for the positive sentiment, and includes indicators like "☺", "good news".

The complete set of posts was coded three times by one of the researchers. Between the second and the third coding round, the second researcher coded a subsample of the set, after which differences were discussed and the coding adapted. As with the collocates, themes could be descriptive as well as have evaluative meanings.

In the following section we describe the results of the analysis.

3. Results

3.1. Exploring the Social Media

In terms of the number of posts in the three-and-a-half year period and the number of different authors, the amount of the communication related to the NSI and business surveys on social media is not great (Table 1). Relative to the total numbers, very few posts about official business surveys are to be found in social media and on web forums:

Table 1. Number of posts, words and authors in the analyzed dataset

	Number
Posts	477
of which retweet or other kind of "repost"	91 (19%)
Words	19,257
Without stop words*	3,513
Authors	378

* Stop words are usually frequent words like "the" that are not meaningful.

477 posts are relevant. Furthermore, of these eligible posts a large number was "retweeted" or otherwise re-sent: 19 percent of the total number of posts is a "repost" of a former post by someone else, or of a news item. The vast majority of these posts are posted by different authors: on average an author posted 1.3 posts.

When looking at these posts over time, at first sight no clear structural annual fluctuations are visible, nor is there any relationship with survey contacts (Figure 1), for example there is no relation to when advance letters were sent out. There is one peak that stands out especially and which we could find an explanation for, namely January 2012. This peak was caused by a press release including a social media post by a Dutch ministry, stating that the administrative burden imposed on entrepreneurs by the NSI had decreased,

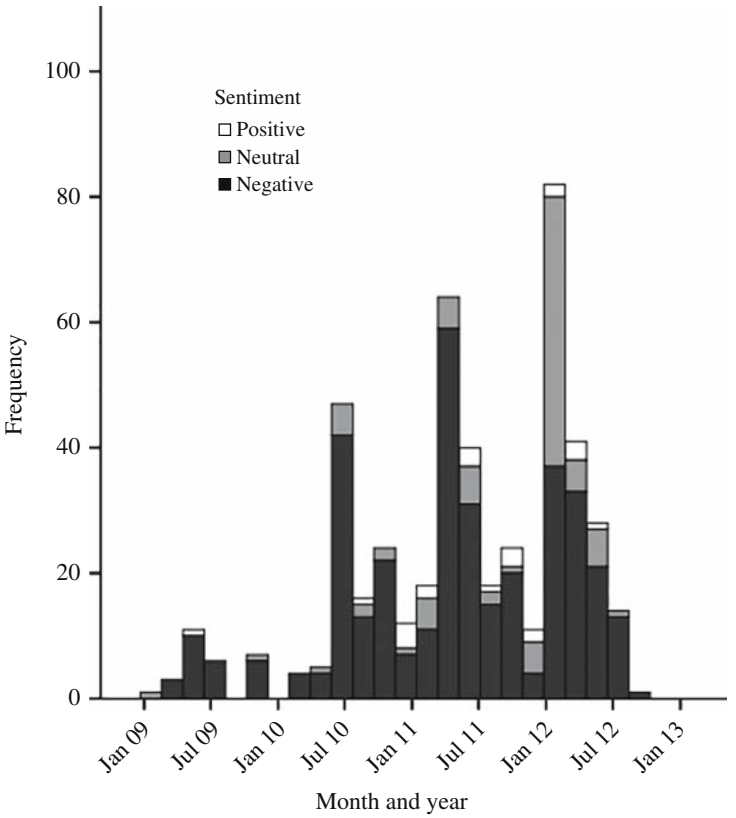


Fig. 1. Number of posts over time

which was neutral in its presentation. This press release resulted in a large number of posts, including a large number of reposts by many different authors. Figure 1 also shows the sentiments of posts over time. The peak in January 2012 mainly consisted of neutral posts (43), followed by 26 negative posts, and only two positive posts. Most of these posts were coded as neutral as they were a straight repost of the original social media post by the Dutch ministry, without adding any positive or negative evaluations.

The largest number of posts in our selection were placed on Twitter (383), followed by Facebook (22 posts). It is important to note that possibly many posts on Facebook are not public, while we only had access to public posts. Conversely, tweets are all public by default.

3.2. *Results of the Lexical Analysis*

The results of the word count or "lexical analysis" are shown in Tables 2 and 3. Table 2 shows the prevalence of words that were among the query words in the Coosto database search. Table 3 shows meaningful words occurring most frequently that were not query words. This table conveys a picture of evaluative meanings that are associated with the NSI in the posts under study, that is, the collocates, and thus of the perceptions the authors of the posts have of the NSI. It is not surprising that in these posts the NSI is associated with "entrepreneurs", "filling in" questionnaires, "data", "statistics", and so on, as these are aspects that are part of the objective, factual role of the NSI in society. These thus can be seen as descriptive aspects that are not the core of our interest.

However, we also see the NSI being associated with other aspects that are not necessarily part of its factual definition. These are thus evaluative meanings attributed to the NSI, which are indicators of the perception and sentiments of the authors. It is these perceptions and sentiments we are interested in. In the following, we list these, clustered for a number of themes, giving examples only for the cases that complement the results of the thematic analysis (as will be discussed in the next subsection):

Technological issues ("software and hardware tools") and "**failure**" (in the two first examples together in one post):

> *"What a #failure. The mandatory survey that you have to fill in as an entrepreneur doesn't function with that browser."*
>
> *"Grrrr. . . . Obliged to share information with the NSI but have been trying for already 2 months just to open the corresponding programme. #fail"*

Table 2. Most frequently occurring query words in the analyzed dataset

Item	Number
NSI	607
Questionnaire	347
Obligation	229
Letter	85
Fine	56
SBS	50
Threatening letter	18

Table 3. Most occurring meaningful words or categories of words in the analyzed dataset (excluding the query words of Table 2)

Item	Number	Item	Number
Fill in, supply, provide	335	"Again"	58
Negation	329	Annual report	50
Software or hardware tools	215	Research	49
Time period	152	Receive (a request)	48
Entrepreneurs	146	Failure	47
Data, information, figures	137	Statistics	34
Must, have to	136	"Time"	31
Sanctions	136	Threat, coercion	30
(of which "fine"*)	56	Report	29
Internet, internet tools	103	Netherlands	28
Government	86	Accountancy	24
Decrease, less	66	Economics	22
Question	66	"Cost", "costs"	20
		Administration	20

* Quotation marks indicate a precise word as opposed to a category of words with similar meaning.

"It is possible to file a complaint but the term to receive an answer is 6 weeks. What do you mean, I'm an organization that is busy only with herself. #NSI#fail."

*"Right. You receive an NSI survey that is obligatory, and then it does **not** function. Big #fail."*

Coercion ("threat and coercion", "sanctions") is highly represented in the data.

We see as well that **negations** are overabundant in the posts, especially as opposed to "yes" or variations of that word (91 occurrences). Detailed scrutiny reveals that they concern a variety of issues of which we give two examples here. In the next section we will explore the issues related to negative feelings more deeply.

*"This is sick. If you do **not** return your data in time to the NSI you're fined to up to X €! Nutcase" (negation in relation with sanctions)*

*"You **cannot** fill in surveys of the NSI with browser X. What a fuss." (negation in relation with technological issues)*

"**Time**" and denotations of **time periods**, such as "month", "year" or "week" are frequently used by business respondents on social media in combination with the NSI. This gives an impression of how important time is for businesses. Among the time periods we see the months January, March, April and July appear. These coincide with the dispatching of survey requests. The references to "time" are two sided: we find posts that talk about a decrease in time spent on the NSI's surveys, but others express experiencing the filling of the questionnaires as a waste of time.

Costs are referred to as costs in time but also as money spent on completing the NSI's questionnaires.

The word "**again**" shows us that the NSI is associated with reiteration. After a detailed scrutiny of these posts, it turns out that the authors of these posts experience contacts with

the NSI as overly frequent. They refer not only to the advance letter and many survey requests, but also to the reminders.

"Pffff again the obligatory exercise for the NSI."

*"Every time again these *** NSI questionnaires. . ."*

Lastly, there are many references to "**decrease"** or **"less"**. These concern posts that talk about a decrease in the administrative burden caused by the NSI.

3.3. Thematic Analysis: Sentiments and Themes

Next, we were interested in the themes present in the posts from a qualitative thematic perspective, and the sentiments expressed: positive, negative, or neutral. The majority of the posts express a negative sentiment (362 or 76 percent), followed by neutral posts (92 or 19 percent). Posts that express a positive sentiment are a minority and only number 23 (5 percent). The fact that there are more negative than neutral or positive posts could be due to the fact that people might be more prone to post something on the web when they have negative feelings than when they have neutral or positive feelings. These percentages are thus not representative of the feelings of the whole population of Dutch business respondents. However, by carefully examining the content of the posts we can infer reasons for these feelings and perceptions. In this section we present the themes found in the posts by sentiment. As Tables 4, 5, and 6 show, the content of the posts clearly diverges for the three kinds of sentiments. This divergence indicates that the respective themes show causes of positive, negative, or neutral feelings. The tables show which themes are related to these sentiments.

Table 4. Major themes in negative posts

Theme	Number	Percentage*
Questionnaires**	285	79
Statutory obligation	186	51.4
Technical problems	*102*	*28*
Unfamiliarity	*101*	*28*
Letter	*81*	*22*
Fine, sanctions	*74*	*20*
Coercive tone	*51*	*14*
Waste of time, costs in time	*42*	*11*
Difficult questionnaire	*35*	*10*
Many questionnaires	*29*	*8*
Unnecessary regulation	*28*	*8*
Tenacity	*24*	*7*
Long questionnaire	*25*	*7*
Lack of communication	*17*	*5*

* Percentage (out of the 362 negative posts) of occurrence of the theme. As the table shows, most posts contain several themes.

** The themes not printed in italics are themes which overlap with search words and are present in the posts with all three sentiments. These are therefore considered to be less significant.

Table 5. Major themes in neutral posts

Theme	Number	Percentage*
Questionnaires**	45	49
Administrative burden (decrease)	*39*	*42*
Statutory obligation	38	41
Unfamiliarity	*31*	*34*
Less costs in time	*9*	*10*
Letter	*9*	*10*
Fine, sanctions	*7*	*8*
Less costs in money	*6*	*7*

* Percentage (out of the 92 neutral posts) of occurrence of the theme. As the table shows, most posts contain several themes.

** The themes not printed in italics are themes which overlap with search words and are present in the posts with all three sentiments. These are therefore considered to be less significant.

In Tables 4, 5 and 6, the percentages indicate what percentages of posts contain the respective theme. Several themes can be present in one post at the same time. Themes not printed in italics are considered to be less important. The reason for this is that these themes contain search words *and* are present in the posts with all three sentiments. It is obvious that when you look for a certain theme, this theme will be present in your search results. On the other hand, if these search words are clearly differentially represented in the postings with different sentiments, then they can be an indicator of what influences the sentiment. The theme "failure of the NSI" is also present in the posts, but is not presented in this section as it was already covered above in the results of the lexical analysis.

3.3.1. Themes in Negative Posts

Table 4 shows the major topics that are identified in posts with a negative sentiment. Below we sum up the main causes of negative sentiments that are revealed and quote some exemplary posts. As these examples show, most posts contain several themes at once.

A significant portion of the posts that express a negative sentiment refers to **technical problems**, namely problems with software and/or hardware (e.g., the questionnaires cannot be completed in certain internet browsers). In particular, the combination of these technical problems with the fact that responding is mandatory appears to have resulted in negative sentiments.

Table 6. Major themes in positive posts

Theme	Number	Percentage*
Questionnaires**	18	78
Statutory obligation	8	35
Simplification questionnaires	*6*	*26*
Positive value of statistics	*4*	*17*
Administrative burden (decrease)	*3*	*13*
Decrease in amount of questionnaires	*3*	*13*

* Percentage (out of the 23 positive posts) of occurrence of the theme. As the table shows, most posts contain several themes.

** The themes not printed in italics are themes which overlap with search words, are blended and are present in the posts with all three sentiments. These are therefore considered to be less significant.

"How tragic: the site of the survey of the NSI – which I'm legally obliged to fill in – only works with internet browser X. #bunglers # government"

"I received a survey from the NSI which I'm legally obliged to complete, but can't be downloaded on X [hardware]. #inwhatyeararetheylivingin?"

A lot of negative posts show that respondents are **unfamiliar** with the NSI, its role in society, the legal obligation to comply and the reasons for receiving the questionnaire. This unfamiliarity seems to make respondents insecure about their position. This seems to cause negative sentiments in combination especially with the receipt of a letter in which one is "threatened" with a fine.

"Oops, the NSI! What a nasty threatening letter. Since when am I legally obliged to hand in statistical data? #sanctions (civil servants open until 17h)."

"Why am I LEGALLY MANDATED to supply company data to the NSI and otherwise the fine will be up to xxx €? #daretoask"

As the posts also show, the **letters** that mention the possibility of **fines** are experienced as intimidating. Moreover, the **coercive tone** is expressed not to have a positive influence on the motivation and response behavior.

"@X1 @X2 Last month I also had such a letter. Badly formulated and compelling, while they don't even explain why #nsi"

"Another of those coercive letters by the NSI about obligatorily filling in the survey on X. Legal sanctions, fine, hell and damnation"

"Just filled in an NSI survey about my company. A task to seriously procrastinate. Probably the threatening letter has a role in that. . #dig my heels in."

The lack of **communication** with the NSI may additionally cause negative sentiments when the help desk is hard to reach for business respondents who have a question.

"#NSI survey, how much time would that cost? Could that be 2 days being unreachable by phone & more than 10 minutes waiting time? And then I still have to fill it in . . . .;("

"But the online NSI questionnaire doesn't function. Telephone waiting time 15 minutes! A shame!"

"Ooooh yes, I am obliged to return the questionnaire by the 11th, but questions by email are only answered after 10 working days because of busyness !!! #fail#NSI"

Taking together all the above, the combination of the fine, the coercive tone, the tenacity, the deadline, the lack of communication, and the technical problems seem to additionally cause negative sentiments, and reinforce these.

"Great. The NSI OBLIGES me to fill in a questionnaire, or else. . . fine. When I want to do this, I get an error message. So then, tell me how I should do this?"

"The NSI is stalking me with some survey and is threatening with sanctions, but in the meanwhile their own online survey questionnaire isn't functioning: #notdoingwell."

Completing official statistics' questionnaires is expressed to be experienced as a **waste of time**. Presumably the reason for this is that entrepreneurs prefer to spend their time on profit-oriented activities, and that they are unaware of the background of the NSI and its surveys.

> *"What a f** survey from the #NSI, costs so much time to fill in, as like if I have nothing better to do."*

> *"NOOOOOO, NSI has made up a new survey and we have to fill that in. #Redtape. Want to work instead of filling in NSI surveys."*

> *"Obliged to complete that survey!. . . . . WASTE OF MY TIME!!!!!!!!!!!!!!!!! With your vague letters!"*

In addition, **characteristics of the questionnaires**, like questionnaires that are hard to complete, long questionnaires, and the fact that businesses receive many questionnaires (for various surveys or as part of a recurring survey) are also found as a cause of negative sentiments.

> *"Filled in an NSI survey on the internet. Jesus, what a user-unfriendly survey was that pffff *completely irritated*"*

> *"@x. It also cost me a full afternoon and a pot of diazepam to fill in that #NSI bunch of misery. I ate the threatening letter ☺"*

> *"#NSI is giving me the itches. On an average I receive 10 surveys a month . . ."*

Some negative posts refer to the **large number of legal requirements** entrepreneurs have to respect. The legal obligation to respond to the NSI's business surveys is seen as one of these, and experienced as an "unnecessary control mechanism".

> *"Filling in the obligatory NSI questionnaire costs hours with those technical problems. Nonsense rules; #NSI #government #wasted tax money # fail."*

> *"I've had my own company for 22 years now and have wasted at least 60 percent of my time complying with rules of the municipality, province, state and some more of these scumbags. And don't forget filling in surveys of the NSI among others. I'll become a communist as well very soon. I don't think it will make a huge difference in this country."*

3.3.2. Themes in Neutral Posts

Even though the neutral posts do not clearly express a positive or negative sentiment, they identify themes that denote business respondents' concern or interest. Besides, they also reflect the perceptions of the authors of the posts. Most of these posts are reposts or retweets, indicating that at least the authors somehow find it worthwhile to make others aware of the message. The main themes in neutral posts, apart from the search words as shown in Table 5, refer to administrative burden (39), and unfamiliarity with the NSI (31), indicating that these issues have the attention of business respondents.

The posts below are examples of neutral posts. We decided to code these as neutral as no clear and objectively definable word(s), symbol(s) or sentence as indicator of a negative

evaluation of Statistics Netherlands or its surveys are present in these posts, like "#fail", "bad", "☹", "Aargh", and so on.

Administrative burden:

"Administrative burden on entrepreneurs by the NSI decreases."

Unfamiliarity:

"Ooops I didn't know that as a company you're legally obliged to supply data if the NSI asks for that, so just decided to do that . . ."

"If I don't cooperate with a survey, there will be consequences?? Since when is it compulsory to fill in a survey?"

"Today I received a letter by the NSI. I must COMPULSORILY cooperate in a study into the business environment. Is that really possible just like that?"

Less costs in time:

"NSI surveys are costing entrepreneurs less time: the administrative burden for entrepreneurs . . ."

Letter:

"Letter from the NSI: the government considers the providing of data that important that they have made it legally obligatory"

3.3.3. Themes in Positive Posts

Themes in positive posts are shown in Table 6. These include the simplification of the questionnaires, the acknowledgment of the value of statistics produced by the NSI, the decrease of the administrative burden, and the observation that the number of questionnaires is being reduced. Examples of these posts are:

Simplification:

"Then there's good news. The NSI reduces and simplifies the surveys for one-man companies and SMEs."

Positive value of statistics:

"Just completely filled in a long questionnaire on request by the NSI. That's good for the accuracy of the statistics. #you'rewelcome."

Decrease administrative burden:

"#goodnews: Administrative burden for entrepreneurs by the NSI decreases."

Decrease in amount of questionnaires:

"Good news for one-man companies, less surveys by the NSI to one-man companies."

4. Conclusion and Discussion

4.1. Exploring and Analyzing Social Media Data

The exploratory study presented in this article is – to the best of our knowledge – one of the first on the use of social media data aimed at exploring the perception and sentiments of business survey respondents. Based on our findings and experience we can suggest the following recommendations toward conducting social media studies on survey respondents.

As discussed in the introduction, the large share of the noninformative "babble" messages on social media negatively affects the use of the more serious informative messages (Daas et al. 2013). It is thus very important to make an appropriate selection and to define the relevant keywords and synonyms (as used by the target population) to use for the query. When the target population is not well defined, posts from units that are out of scope may be selected, resulting in overcoverage of the posts. This is important as the number of data records will be vast. On the other hand, when keywords are missed, the data set may not cover all eligible posts, resulting in undercoverage of posts. As Stieglitz and Dang-Xuan (2012, 1283) state: "to attain a high level of data completeness, relevant keywords representing the topic of interest have to be carefully and systematically chosen in advance".

The characteristics of social media data also have implications for the analysis methods that should be used. As discussed in Subsection 2.2, we applied a two-step mixed-method design, including both a word count and a thematic analysis. As for the thematic analysis, posts were coded several times by one of the authors and a sample was coded by both authors. This was done to achieve intra- and intercoding reliability.

The two methods used in this two-step mixed-method approach can help triangulate their respective findings (e.g., Baker et al. 2008, 295). Triangulation is a method used by qualitative or mixed-methods researchers to check and establish validity in their studies by analyzing a research question from multiple perspectives (Guion et al. 2011). However, our two chosen methods are also complementary: with the lexical analysis we were able to identify themes we would not have found with the thematic analysis and vice versa. The lexical analysis in the first step aided the identification of themes and coding of the posts, as it showed quantitative evidence of words and patterns being used repeatedly (e.g., Hardt-Mautner 1995; Baker et al. 2008), and aided in the identification of areas of interest (e.g., Mautner 2007). In this way it helps to safeguard against over- or underinterpretation. The thematic analysis in the second step has the advantage that one can look beyond the semantic level, and themes can be coded at the latent level. Also, so-called wider themes (Gabrielatos and Baker 2008) or metathemes (Ryan and Bernard 2003) can be identified. Subtle, implicit meanings cannot be easily analyzed through lexical analysis (Wodak 2007), but can be identified with qualitative thematic coding. One example in the case of this study would be the unfamiliarity with the NSI frequently referred to in the postings. We were only able to identify this theme when we did the thematic coding. If we had only done the lexical analysis we would not have identified this theme.

The representativeness of social media data is still an issue, but recent findings are positive. Daas et al. (2013) discuss opportunities and challenges associated with using social media data from the Coosto database for official statistics based on a case study conducted at Statistics Netherlands. They found that the monthly sentiment for the period June 2010 to August 2012 derived from Dutch social media messages taken from the Coosto database correlated very strongly (0.83) with the officially determined monthly Dutch consumer confidence. In addition, it also correlated with the sentiment for the subindicator of the attitude towards the economic climate (0.88). This high correlation is remarkable, as the populations from which the data are obtained are different: both official indicators are based on a sample survey in which 1,500 people are interviewed each month.

4.2. *Summary and Discussion of the Analysis Results*

We found that the discussions on social media with regard to the NSI, its surveys and questionnaires are very small in number. The number is small both relative to the total number of public posts and relative to the total number of questionnaires dispatched every year. Furthermore, we did not find important annually reoccurring increases in communication activities related to the dispatching of questionnaires.

The topics discussed show a variety of themes, varying over the associated sentiments: negative, positive, and neutral. In negative posts the main themes are: technical failures; unfamiliarity with the NSI and its role; letters that are perceived as too coercive; the idea that filling in a questionnaire is a waste of time and that these surveys are "unnecessary mandatory regulations"; characteristics of the questionnaire and the inaccessibility of the NSI for business respondents with questions, both by telephone and email. In a number of posts more than one theme was mentioned, indicating that especially a combination of these aspects may lead to negative feelings. The combination of the technical problems with the waiting time to contact the NSI, the legal obligation to comply and the strict deadlines to return the completed questionnaire, is not helpful in establishing a positive perception of the NSI.

A positive image is associated with the simplification of questionnaires, the positive value of official statistics and the (perception of) a decrease of the response burden.

Neutral posts indicate that entrepreneurs somehow show an interest in the reduction of response burden, simply by retweeting messages about this topic. The neutral posts also show, like the negative posts, that some business respondents are unfamiliar with the NSI, from which we can conclude that this is an important topic to address.

Considering the attention that is given to the imposed response burden by SN, by politicians and in various publications by, for example, business organizations, commercial banks, as well as the government, we had expected that this topic would also be discussed quite often on social media. However, this assumption is not corroborated by our exploratory analysis. This may have several reasons. The findings may indicate that a vast majority of entrepreneurs is not as interested in posting on this topic on social media as we thought. Possibly they are too busy running their businesses to spend much time communicating on social media about these issues. It is also possible that the topic in itself is not as important to them as we thought. This might

also be a reason why we did not find important annually reoccurring increases in communication activities related to the dispatching of questionnaires. However, we still need to put some critical remarks with these data and the analyses. We only had access to public posts; private posts are not included in our database. This may lead to an underestimation of the number of messages in our source.

Nevertheless, we can validate our conclusions. We can conclude that in the overall picture, the findings (as to the expressed sentiments) are in line with findings from previous qualitative and quantitative studies on sentiments of business respondents towards SN and its surveys. Giesen (2012) analyzed the results of the Customer Satisfaction Survey conducted by SN in 2006. In this survey, a sample of respondents of the Dutch Structural Business Survey (SBS) was contacted for a short CATI interview asking questions about the completion of the SBS questionnaire and the respondents' opinion about SN. As an indicator of the overall attitude towards SN, a question is used which asks respondents to rate their overall satisfaction with SN with a grade between 1 and 10. On average respondents grade SN at 6.5. According to the Dutch system of rating school grades this would be slightly above satisfactory. 13 percent of the respondents give a grade below 6, an unsatisfactory grade. In Snijkers et al. (2007), qualitative data show that negative sentiments are related to: costs; the statutory obligation; the fines; threats and tone of the letters; and lack of knowledge of survey procedures (especially the importance of random sampling). Interest in and usefulness of the survey topic is another important topic: businesses generally did not rate surveys as useful for society or themselves. Also unfamiliarity is found. Customer-friendliness, on the other hand, is associated with positive sentiments.

Following on from these studies, with this analysis we have gained additional quantitative information about perceptions of business respondents of official surveys. More importantly, we have gained information about causes for a positive or negative perception, and aspects that are related to positive or negative sentiments. This analysis is an inductive and empirical observation of existing perceptions and sentiments about the NSI and its surveys and causes for these perceptions and sentiments. The breakdown of the sentiments in percentages may not be representative of the breakdown in percentages of the whole population of business respondents in the Netherlands, but that is not the main and most important finding of the analysis.

The analysis shows which actions taken by the NSI and which features of the communication and survey design are related to the expressed perceptions and sentiments. Therefore, the results give an empirical indication of how the communication and survey design could be adapted to positively influence survey response (Figure 2).

4.3. Recommendations for Survey and Communication Design

Based on the results, we can formulate a set of recommendations that should eliminate the causes of negative sentiments, enhance a positive perception of the NSI and its surveys, and therefore enhance motivation to comply and increase response rates and the quality of the response. These recommendations follow and corroborate the recommendations of previous studies (for example Snijkers et al. 2007; Torres van Grinsven et al. 2011, 2012).

They also are in line with internationally identified factors affecting the business survey response process (Snijkers et al. 2013).

As costs in time and money are important for entrepreneurs, it is important to reduce participation costs but also to increase the *(perceived) value*, or the (perceived) benefit, of the surveys for the businesses. This is in line with social exchange theory (Homans 1958), which when applied to survey behavior asserts that the actions of respondents in answering a questionnaire are motivated by the personal benefit these actions are expected to bring, or usually do bring. Whether a given behavior occurs is a function of the perceived costs of engaging in that activity and the expected rewards (Poon et al. 1999). Social exchange theory has been applied extensively to improve survey participation in the field of household surveys (e.g., Dillman 1978). Singer (2012) proposes a variant of this theory – the benefit-cost theory, in which the argument is that people choose to act when the benefits of doing so outweigh the costs in their subjective calculation.

Along the same line, it is very important to *facilitate and simplify* the response tasks to reduce (perceived) costs. Initial perceptions of high costs and negative experiences cause a negative perception and therefore a higher perceived burden (Giesen 2012), especially in combination with a coercive tone, the strict response deadline, and response chasing. The NSI should also make clear to business respondents that it has made efforts to facilitate the response process. Furthermore, our analysis shows that simplification of questionnaires enforces positive sentiments, which may lift the public image of an NSI.

Besides, the *unfamiliarity* with the NSI, its role as surveyor of data and producer and publisher of statistics needs to be marketed by developing a *sound and coherent survey communication strategy* (see also Snijkers 2009).

A final set of recommendations is related to the use of social media as a communication channel to be used by NSIs. Social media can be used to disseminate statistics, but can also be used to communicate with respondents and as a channel in a communication strategy to enhance a positive perception (see, Figure 2).

Current theories within linguistics claim that people are unconsciously primed to infer meanings due to the cumulative effect of all of their previous encounters with a word, that is, the collocates of that word (e.g., Hoey 2005). This means that if people in the media

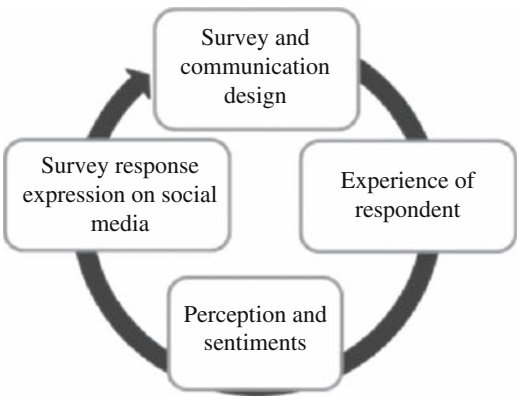


Fig. 2. Communication strategy and sentiments of respondents

consistently come across the NSI in combination with burden, they are primed to associate the NSI with burden. Therefore it is important for statistical bureaus to consciously engage in well-considered communicative practices to influence the perception that people and business respondents have of the statistical bureau, and evaluate the effectiveness of these practices. Social media, together with traditional media, can be used to actively and pervasively inform the public and businesses about the NSI, its role and importance in the modern information society, and consequently influence the perception and image business respondents and the public in general have of an NSI. Social media can also be used to monitor the effectiveness of these communication activities.

5. References

Adolphs, S., B. Brown, R. Carter, P. Crawford, and O. Sahota. 2004. "Applying Corpus Linguistics in a Health Care Context." *Journal of Applied Linguistics* 1: 9–28.

Baker, P., C. Gabrielatos, M. Khosravinik, M. Krzyzanowski, T. McEnery, and R. Wodak. 2008. "A Useful Methodologic Synergy? Combining Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum Seekers in the UK Press." *Discourse and Society* 19: 273–306. Doi: http://dx.doi.org/10.1177/0957926508088962.

Blumer, H. 1973. *Symbolic Interactionism: Perspectives and method*. Prentice-Hall, Englewood Cliffs: New Jersey.

Braun, V. and V. Clarke. 2006. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3: 77–101. Doi: http://dx.doi.org/10.1191/1478088706qp063oa.

Coosto. 2014. *The Facts Webpage*. Available at: http://www.coosto.nl/home/about/feiten and in English at http://www.coosto.co.uk/home/about/facts (accessed March 2014).

Daas, P.J.H. and M.J. Puts. 2014. "Social Media Sentiment and Consumer Confidence. Statistics Paper Series, No. 5. European Centeral Bank." Available at: http://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.pdf (accessed April 16, 2015).

Daas, P.J.H. and M.J. Puts. 2014a. "New and Emerging Methods: Big Data as a Source of Statistical Information." *The Survey Statistician* 69. Available at: http://isi.cbs.nl/iass/N69.pdf (accessed April 16, 2015).

Daas, P.J.H., M.J. Puts, B. Buelens, and P.A.M. van den Hurk. 2013. "Big Data and Official Statistics." In *Proceedings of the NTTS (New Techniques and Technologies for Statistics) 2013, March 5–7, Brussels*. Available at: http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_76.pdf (accessed April 16, 2015).

Daas, P.J.H., M. Roos, M. van de Ven, and J. Neroni. 2012. "Twitter as a Potential Data Source for Statistics." Discussion paper 201221. The Hague/Heerlen: Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/04B7DD23-5443-4F98-B466-1C67AAA19527/0/ 201221x10pub.pdf (accessed April 16, 2015).

Dillman, D. 1978. *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley and sons.

Gabrielatos, C. and P. Baker. 2008. "Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press

1996–2005." *Journal of English Linguistics* 36: 5–38. Doi: http://dx.doi.org/10.1177/0075424207311247.

Giesen, D. 2012. "Exploring Causes and Effects of Perceived Response Burden." Paper presented at the Fourth International Conference on Establishment Surveys (ICES IV), Montreal, 11–14 June, 2012. Available at: http://www.amstat.org/meetings/ices/2012/papers/302171.pdf (accessed April 16, 2015).

Groves, R.M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly* 75: 861–871. Doi: http://dx.doi.org/10.1093/poq/nfr057.

Guion, L.A., D.C. Diehl, and D. McDonald. 2011. *Triangulation: Establishing the Validity of Qualitative Studies*. Available at: http://edis.ifas.ufl.edu/fy394 (accessed June 1, 2014).

Haraldsen, G. 2013. "Quality Issues in Business Surveys." In *Designing and Conducting Business Surveys*, edited by G. Snijkers, H. Haraldsen, J. Jones, and D. Willimack, 83–125. Hoboken, NJ: Wiley.

Haraldsen, G., J. Jones, D. Giesen, and L.-C. Zhang. 2013. "Understanding and Coping with Response Burden." In *Designing and Conducting Business Surveys*, edited by G. Snijkers, H. Haraldsen, J. Jones, and D. Willimack, 219–252. Hoboken: Wiley.

Hardt-Mautner, G. 1995. *Only Connect*. Critical Discourse Analysis and Corpus Linguistics. UCREL Technical Paper 6. Lancaster: Lancaster University. Available at: http://ucrel.lancs.ac.uk/papers/techpaper/vol6.pdf (accessed April 16, 2015).

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Homans, G.C. 1958. "Social Behavior as Exchange." *Americal Journal of Sociology* 63: 597–606.

Jehn, K.A. and L. Doucet. 1996. "Developing Categories from Interview Data: Text Analysis and Multidimensional Scaling. Part 1." *Field Methods* 8: 15–16.

Jehn, K.A. and L. Doucet. 1997. "Developing Categories for Interview Data: Consequences of Different Coding and Analysis Strategies in Understanding." *Field Methods* 9: 1–7.

Jensen, K.B. 2012. *A Handbook of Media and Communication Research. Qualitative and Quantitative Methodologies*, 2nd ed. New York: Routledge.

Jensen, K.B. and R. Helles. 2011. "The Internet as a Cultural Forum: Implications for Research." *New Media and Society* 13: 517–533. Doi: http://dx.doi.org/10.1177/1461444810373531.

Leech, G. 1992. "Corpora and Theories of Linguistic Performance." In *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991*, edited by J. Svartvik, 105–122. Berlin: Mouton de Gruyter.

Mautner, G. 2007. "Mining Large Corpora for Social Information: The Case of *Elderly*." *Language in Society* 36: 51–72. Doi: http://dx.doi.org/10.1017/S0047404507070030.

Nattinger, J.R. and J.S. DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Pennebaker, J.W., M.E. Francis, and R.J. Booth. 2001. *Linguistic Inquiry and Word Count (LWIC): A Computerized Text Analysis Program*. Mahwah, NJ: Erlbaum Publishers.

Poon, P., G. Albaum, and F. Evangelista. 1999. "An Empirical Test of Alternative Theories of Survey Response Behaviour." *International Journal of Market Research* 41: 1–18.

Ryan, G.W. and H.R. Bernard. 2003. "Techniques to Identify Themes." *Field Methods* 15: 85–109. Doi: http://dx.doi.org/10.1177/1525822X02239569.

Ryan, G.R. and T. Weisner. 1996. "Analyzing Words in Brief Descriptions: Fathers and Mothers Describe Their Children." *Field Methods* 8: 13–16.

Seale, C., S. Ziebland, and J. Charteris-Black. 2006. "Gender, Cancer Experience and Internet Use: A Comparative Word Analysis of Interviews and Online Cancer Support Groups." *Social Science and Medicine* 62: 2577–2590. Doi: http://dx.doi.org/10.1016/j.socscimed.2005.11.016.

Silverman, D. 2000. "Analyzing Talk and Text." In *The Handbook of Qualitative Research*, edited by N.K. Denzin and Y.S. Lincoln, 821–834. Thousand Oaks, CA: Sage.

Sinclair, J. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Singer, E. 2012. "Toward a Benefit-Cost Theory of Survey Participation: Evidence, Further Tests, and Implications." *Journal of Official Statistics* 27: 379–392.

Snijkers, G. 2008. "Getting Data for Business Statistics: A Response Model." Paper presented at the 4th European Conference on Quality in Official Statistics, Rome. Available at: http://q2008.istat.it/sessions/25.html (accessed April 16, 2015).

Snijkers, G. 2009. "Getting Data for (Business) Statistics: What's new? What's next?" Paper presented at the 2009 NTTS Conference (New Techniques and Technologies for Statistics). Brussels. Available at: http://ec.europa.eu/eurostat/documents/1001617/4398389/S5P2-GETTING-DATA-FOR-STATISTICS-SNIJKERS.pdf (accessed April 16, 2015).

Snijkers, G., B. Berkenbosch, and M. Luppes. 2007. "Understanding the Decision to Participate in a Business Survey." In Proceedings of the Third International Conference on Establishment Surveys (ICES-III). 18–21 June, 2007. Alexandria, VA: American Statistical Association, 1048–1059. Available at: https://www.amstat.org/meetings/ices/2007/proceedings/TOC.pdf.

Snijkers, G., R. Göttgens, and H. Hermans. 2011. "Data Collection and Data Sharing at Statistics Netherlands: Yesterday, Today, Tomorrow." Paper presented at the 59th plenary session of the Conference of European Statisticians, 14–16 June, 2011, Geneva. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2011/20.e.pdf (accessed April 16, 2015).

Snijkers, G., G. Haraldsen, J. Jones, and D.K. Willimack. 2013. *Designing and Conducting Business Surveys*. Hoboken, NJ: Wiley.

Snijkers, G. and J. Jones. 2013. "Business Survey Communication." In *Designing and Conducting Business Surveys*, edited by G. Snijkers, H. Haraldsen, J. Jones, and D. Willimack, 359–430. Hoboken, NJ: Wiley.

Stieglitz, S. and L. Dang-Xuan. 2012. "Social Media and Political Communication: A Social Media Analytics Framework." *Social Network Analysis and Mining* 3: 1277–1291. Doi: http://dx.doi.org/10.1007/s13278-012-0079-3.

Tesch, R. 1990. *Qualitative Research: Analysis Types and Software Tools*. New York: Falmar Press.

Torres van Grinsven, V., I. Bolko, and M. Bavdaž. 2014. In Search of Motivation for Business Survey Response Task. *Journal of Official Statistics* 30: 579–606. Doi: http://dx.doi.org/10.2478/JOS-2014-0039.

Torres van Grinsven, V., I. Bolko, M. Bavdaž, and S. Biffignandi. 2011. "Motivation in Business Surveys." BLUE-ETS Conference on business' burden and motivation in official surveys, Statistics Netherlands, March 22–23 edited by D. Giesen and M. Bavdaž. Available at: http://www.cbs.nl/NR/rdonlyres/23FD3DF5-6696-4A04-B8EF-1FAACEAD995C/0/2011proceedingsblueets.pdf (accessed April 16, 2015).

Van Vroenhoven, J. 2006. "Storend!" In *Humor om te huilen: Zwartboek doorgeslagen regelgeving*, 22–23. Tilburg: Brabants-Zeeuwse Werkgeversvereniging (BZW).

Wenemark, M., A. Persson, H. Noorlind Brage, T. Svensson, and M. Kristenson. 2011. "Applying Motivation Theory to Achieve Increased Response Rates, Respondent Satisfaction and Data Quality." *Journal of Official Statistics* 27: 393–414.

Willimack, D. and G. Snijkers. 2013. "The Business Context and Its Implications for the Survey Response Process." In *Designing and Conducting Business Surveys*, edited by G. Snijkers, H. Haraldsen, J. Jones, and D. Willimack, 39–82. Hoboken, NJ: Wiley.

Wodak, R. 2007. "Pragmatics and Critical Discourse Analysis: A Cross-Disciplinary Inquiry." *Journal of Pragmatics and Cognition* 15: 203–227. Doi: http://dx.doi.org/10.1075/pc.15.1.13wod.

Received August 2013
Revised January 2015
Accepted January 2015

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 305–324, http://dx.doi.org/10.1515/JOS-2015-0019

Measuring Disclosure Risk and Data Utility for Flexible Table Generators

Natalie Shlomo[1], *Laszlo Antal*[1], *and Mark Elliot*[1]

Statistical agencies are making increased use of the internet to disseminate census tabular outputs through web-based flexible table-generating servers that allow users to define and generate their own tables. The key questions in the development of these servers are: (1) what data should be used to generate the tables, and (2) what statistical disclosure control (SDC) method should be applied. To generate flexible tables, the server has to be able to measure the disclosure risk in the final output table, apply the SDC method and then iteratively reassess the disclosure risk. SDC methods may be applied either to the underlying data used to generate the tables and/or to the final output table that is generated from original data. Besides assessing disclosure risk, the server should provide a measure of data utility by comparing the perturbed table to the original table. In this article, we examine aspects of the design and development of a flexible table-generating server for census tables and demonstrate a disclosure risk-data utility analysis for comparing SDC methods. We propose measures for disclosure risk and data utility that are based on information theory.

Key words: Statistical disclosure control; census tabular data; entropy; Hellinger distance.

1. Introduction

Driven by demand from policy makers and researchers for specialized and tailored census frequency tables, many statistical agencies are considering the development of a web-based software platform where users can generate tables of interest from underlying census microdata through a user-friendly interface. This platform is called a "flexible table-generating server". Users access the server via the internet and generate their preferred set of tables from predefined variables or categories using drop-down lists. These tables can then be downloaded to the personal computers of the users. The United States Census Bureau and the Australian Bureau of Statistics have developed such servers on their websites to disseminate census frequency tables.

When generating flexible tables, the server should be able to provide a measure of disclosure risk for the original table, apply a statistical disclosure control (SDC) method and then reassess disclosure risk and the impact on data utility following the SDC method. These steps must be carried out "on the fly" within the server for each generated output table. SDC is a set of statistical practices which aim to ensure that no individual population

[1] University of Manchester, Social Statistics, Humanities Bridgeford Street, Manchester M13 9PL, United Kingdom. Emails: natalie.shlomo@manchester.ac.uk, laszlo.antal@postgrad.manchester.ac.uk, and mark.elliot@manchester.ac.uk

Acknowledgments: The project is funded by the EU 7th framework infrastructure research grant: 262608, Data Without Boundaries (DwB) and the ONS-ESRC funded PhD studentship (Ref. ES/J500161/1).

unit can be reidentified from anonymised data nor any new information learnt about any specific individual (with certainty). SDC is an active research area. For reviews of this area, see Willenborg and de Waal (2001), Doyle et al. (2001), Duncan et al. (2011) and Hundepool et al. (2012).

There are two main types of disclosure risks in census frequency tables: identity disclosure, where small cell counts may lead to the identification of an individual in the population, and attribute disclosure, where new information may be learnt about an individual or group of individuals. Attribute disclosure in frequency tables occurs when rows or columns of a table contain (real) zeroes and only one or two cells are nonzero. This enables an "intruder" to first make an identification based on a margin total and subsequently reveal new information according to other variables spanning the table. Another type of disclosure risk that needs to be guarded against is disclosure by differencing. The differencing of tables generated through the server can lead to residual tables that are more susceptible to the above disclosure risks and even to the reconstruction of individual records. This is typically dealt with by applying perturbative methods of SDC, which raises the level of uncertainty of true counts in the tables and hence of the difference between counts across tables. After the table is protected, a data utility measure must also be calculated by comparing the perturbed table to the original table.

The need to measure disclosure risk "on the fly" for census frequency tables produced via a flexible table-generating server motivated the research and development of a new global disclosure risk measure. Until now, disclosure risk measures for tabular data have been defined at the cell level and not for the entire table. We propose a new disclosure risk measure based on information theory as shown in Antal et al. (2014) and also relate this theory to a data utility measure.

The key issues when developing a web-based flexible table generating server addressed in this article are: (1) what underlying data should be used in the background for generating the output tables, and (2) at what stage should the SDC method be applied. In addition, the article provides a comparison study of some common SDC methods which may be used to protect census tables within a flexible table-generating server and demonstrates how statistical agencies should undertake a disclosure risk-data utility analysis to inform decisions about SDC methods and their parameterization. In general, SDC methods employed by statistical agencies are often motivated by country-specific agendas and policy sensitivities and it is difficult to develop a universal best practice. However, one important distinction when considering SDC methods for flexible table-generating servers is that the outputs are defined by users and the amount of disclosure risk may vary in each output.

Section 2 presents aspects to consider in the design of a flexible-table generating server, including the underlying data for generating output tables and the stage when SDC methods may be applied. In Section 3, some common SDC methods for census frequency tables are described. Section 4 introduces a new global disclosure risk measure based on information theory and a related data utility measure that can be calculated "on the fly" for each output table generated in the server. In Section 5, a comparison study is carried out on generated census output tables from a flexible table-generating server. The comparison study will be informed by a disclosure risk-data utility analysis on the generated tables perturbed by the SDC methods described in Section 3 based on the

measures outlined in Section 4. A discussion and concluding remarks are presented in Section 6.

2. Designing a Flexible Table-Generating Server

In this section, we describe the design of an online flexible table-generating server and discuss the following issues: the underlying data that may be used as input to the server, the stage at which SDC methods can be applied, and preliminary SDC rules to determine *a priori* whether the requested table can be generated or not.

2.1. Underlying Input Data to the Server

The underlying data to use as input for a web-based flexible table-generating server can be based on the original microdata or disclosure-controlled microdata. The input data is largely determined by the source and content of the data as well as the SDC method that will be applied to the final output tables (if any). Microdata arising from social surveys with small sampling fractions have a lower disclosure risk than microdata arising from censuses containing whole population counts, and therefore are more appropriate for use in their original form. Output tables generated from survey microdata where only weighted counts are released are generally considered to be of low disclosure risk with no further need for an application of SDC methods. Census (and administrative data) containing whole populations and particularly those containing sensitive data, such as health statistics or business microdata, are more problematic. In microdata containing the whole population, individuals (or businesses) can easily be identified leading to the disclosure of attributes. In this case, the underlying input data should be protected prior to the generation of tables.

For a flexible table-generating server of census tables, one method for producing the underlying input data is to aggregate the microdata into a very large multi-dimensional frequency table, called a hypercube, where no data of individuals can be disseminated below the level of a cell value in the hypercube. For example, users may only be able to disseminate frequency counts of age in 5-year age bands and not counts for single years. This approach was taken by Eurostat for the dissemination of census tables from European Member States. A flexible table-generating server for European census tables is being developed through the European Census Hub Project. Each Member State is required to produce a set of predefined hypercubes containing their country's census counts: 19 hypercubes at the geography level of LAU2 and over 100 hypercubes at the geography level of NUTS2, cross-classified with as many as six other census variables in each hypercube. NUTS2 is a European subregional geography and LAU2 are small municipalities or equivalent. Researchers are able to use the considerable number of multidimensional hypercubes and their wealth of census data made available through the European Census Hub to generate tables of interest beyond what would have been available previously using standard table-extraction software. The flexible table-generating server will allow comparative tables across Member States and the combining of census data from multiple Member States. The hypercubes have the additional advantage that they provide some limited protection against disclosure risk since no data below the level of the cell values of the hypercube can be disseminated.

However, the hypercubes themselves still have considerable disclosure risk since they are very large and sparse with many zero and small cell counts. Therefore, there will still be the need to apply an SDC method to protect output tables generated from the flexible table-generating server.

2.2. *Application of SDC Methods*

SDC methods for protecting output tables generated from a flexible table-generating server can be applied either on the underlying input data so that all tables generated are deemed safe for dissemination (the pretabular SDC approach), or applied directly to the final output table generated from the original data (the post-tabular SDC approach) or a combination of both. Although sometimes neater and less resource intensive when data is from a single source, the pretabular SDC approach is problematic for the dissemination of European Census data for two reasons. Firstly, all Member States would have to agree on a common SDC method in order to provide consistent hypercubes across all Member States. For example, if one Member State employs a rounding method whilst another Member State employs cell suppression, there will be significant quality issues in a table that is generated based on both Member States' data. Secondly, when aggregating data which have been separately disclosure controlled, the effects of the SDC methods are compounded and the data may be overprotected. For example, aggregating cells that have already been rounded not only overprotects the data but also exacerbates the data utility impact by providing counts that are no longer rounded to the nearest base. With the second approach of protecting only the final tabular output, SDC methods are not compounded in this way. We investigate the pretabular and post-tabular approaches in the comparison study presented in Section 5.

2.3. *Preliminary SDC Rules*

The design of a web-based flexible table-generating server typically involves many *ad hoc* preliminary SDC rules which determine *a priori* if generated tables can be released or not. These SDC rules may include:

- Limiting the number of dimensions in the output tables.
- Ensuring consistent and nested categories of variables to avoid disclosure by differencing.
- Ensuring minimum population thresholds.
- Ensuring that the percentage of small cells is below a maximum threshold.
- Ensuring average cell size above a minimum threshold.

The steps in a flexible table-generating server are:

(1) Determine whether the table can be released according to the preliminary SDC rules.
(2) Calculate a disclosure risk measure to determine if an SDC method should be applied to the final output table.
(3) Apply the SDC method.

(4) Recalculate the disclosure risk measure to determine if the table is safe to generate; if yes proceed to Step 5, otherwise do not release the table.
(5) Output the final table with a measure of data utility.

According to the steps of a flexible table-generating server, it is clear that analytical expressions of disclosure risk and data utility that can be calculated "on the fly" within the server are necessary.

3. Statistical Disclosure Control Methods

In this section, we describe some common SDC methods which have been used to protect census frequency tables: a pretabular SDC method of record swapping is used in the United States and the United Kingdom, a post-tabular method of random rounding is used in New Zealand and Canada, and a post-tabular probabilistic perturbation mechanism has recently been implemented in Australia.

3.1. Record Swapping

Record swapping is based on the exchange of values of variable(s) between similar pairs of population units (often households). In order to minimize bias, pairs of population units are determined within strata defined by control variables. For example, when swapping households, control variables may include: a large geographical area, household size, and the age-sex distribution of individuals in the households. In addition, record swapping can be targeted to high-risk population units found in small cells of census tables. In a census context, geographical variables related to place of residence are often swapped. Swapping place of residence has the following properties: (1) it minimizes bias based on the assumption that place of residence is independent of other census target variables conditional on the control variables; (2) it provides more protection for census tables since place of residence is a highly visible variable which can be used to identify individuals; (3) it preserves marginal distributions within a larger geographical area. For more information on record swapping, see Dalenius and Reiss (1982), Fienberg and McIntyre (2005), and Shlomo (2007).

3.2. Semi-Controlled Random Rounding

A post-tabular method of SDC for census frequency tables is unbiased random rounding. Let *Floor(x)* be the largest multiple bk of the base b such that $bk < x$ for any value of x. In this case, $res(x) = x\text{-}Floor(x)$. For an unbiased rounding procedure, x is rounded up to $Floor(x) + b$ with probability $res(x)/b$ and rounded down to $Floor(x)$ with probability $(1 - (res(x)/b))$. If x is already a multiple of b, it remains unchanged.

In general, each cell is rounded independently in the table, that is, a random uniform number u between 0 and 1 is generated for each cell. If $u \leq (res(x)/b)$ then the entry is rounded up, otherwise it is rounded down. This ensures an unbiased rounding scheme, that is, the expectation of the rounding perturbation is zero. However, the realization of this stochastic process on a finite number of cells in a table will not ensure that the sum of the perturbations will exactly equal zero. To place some control in the random rounding procedure, we use a semi-controlled random rounding algorithm for selecting entries to round up or down as follows: first the expected number of entries of a given $res(x)$ that are

to be rounded up is predetermined (for the entire table or for each row/column of the table). The expected number is rounded to the nearest integer. Based on this expected number, a random sample of entries is selected (without replacement) and rounded up. The other entries are rounded down. This procedure ensures that rounded internal cells aggregate to the controlled rounded total.

Due to the large number of perturbations under random rounding, margins are typically rounded separately from internal cells and tables are not additive. When using semicontrolled random rounding this alleviates some of the problems of nonadditivity since one of the margins and the overall total will be preserved. Another problem with random rounding is the consistency of the rounding across same cells that are generated in different tables. It is important to ensure that the cell value is rounded consistently, otherwise the true cell count can be learnt by generating many tables containing the same cell and observing the perturbation patterns. Fraser and Wooton (2005) propose the use of *microdata keys* which can solve the consistency problem. First, a random number (which they call a key) is defined for each record in the microdata. When building a census frequency table, records in the microdata are combined to form a cell defined by the spanning variables of the table. When these records are combined to a cell, their keys are also aggregated. This aggregated key serves as the seed for the rounding and therefore same cells will always have the same seed and result in consistent rounding.

Further research is needed to ensure both the additivity and consistency properties for random rounding. For simple tables of the type that would be generated in a flexible table-generating server, controlled rounding algorithms can be applied to ensure additivity on remaining totals without distorting the unbiasedness of the rounding (see Willenborg and De Waal 2001).

3.3. *Stochastic Perturbation*

A more general method than random rounding is stochastic perturbation, which involves perturbing the internal cells of a table using a probability transition matrix and is similar to the postrandomisation method that is used to perturb categorical variables in microdata (see Gouweleeuw et al. 1998). In this case, it is the cell counts in a table that are perturbed. More details can be found in Fraser and Wooton (2005) and Shlomo and Young (2008).

Let $\mathbf{P}$ be a $(L+1)\times(L+1)$ transition matrix containing conditional probabilities: $p_{ij} = P(\textit{perturbed cell value is } j \,|\, \textit{original cell value is } i)$ for cell values from 0 to L, where L is a cap on the cell values and any cell value above the cap will have the same perturbation probabilities. Let $\mathbf{t}$ be the vector of frequencies of the cell values where the last component would contain the number of cells above cap L and let $\mathbf{v}$ be the vector of relative frequencies: $\mathbf{v} = \mathbf{t}/\mathrm{K}$ where K is the number of cells in the table. In each cell of the table, the cell value i is changed or not changed according to the prescribed transition probabilities in matrix $\mathbf{P}$ and the result of a draw of a random multinomial variate u with parameters $p_{ij}j = 0, 1, \ldots, L$. If the jth value is selected, value i is moved to value j. When $i = j$, no change occurs.

Placing the condition of invariance on the probability transition matrix $\mathbf{P}$ (i.e., $\mathbf{tP} = \mathbf{t}$) means that the marginal distribution of the cell values are approximately preserved under

the perturbation. As described in the random rounding procedure in Subsection 3.2, in order to obtain the exact marginal distribution a similar strategy for selecting cell values to change can be carried out. For each cell value *i,* the expected number of cells that need to be changed to a different value *j* is calculated according to the probabilities in the transition matrix. We then randomly select (without replacement) the expected number of cells *i* and carry out the change to *j*.

To preserve exact additivity in the table, an iterative proportional fitting algorithm can be used to fit the margins of the table after the perturbation according to the original margins. This results in cell values that are not integers. Exact additivity with integer counts can be achieved for simple tables by controlled rounding to base 1 using Tau-Argus, for example (Salazar-Gonzalez et al. 2005). Cell values can also be rounded to their nearest integers resulting in "close" additivity because of the invariance property of the transition matrix. Finally, the use of microdata keys as described in Subsection 3.2 can also be adapted to this SDC method to ensure the consistent perturbation of same cells across different tables by fixing the seed for the perturbation.

4. Information Theory-Based Disclosure Risk and Data Utility Measures

For each output table generated, the flexible table-generating server must provide analytical expressions of disclosure risk and data utility that can be calculated "on the fly" within the server. As mentioned in Section 1, one of the major causes of disclosure risk in census frequency tables is attribute disclosure caused by rows/columns that have many zero cells and only one or two populated cells. A row/column with a uniform distribution of cell counts would have little attribute disclosure risk, whilst a degenerate distribution of cell counts would have high attribute disclosure risk. Moreover, a row/column with large counts would have less risk of reidentification compared to a row/column with small counts.

There is no single global-level disclosure risk measure for census frequency tables that measures attribute disclosure and identity disclosure. In planning for the 2011 UK Census, the Office for National Statistics assessed attribute disclosure by producing many census tables and calculating the proportion of those columns/rows where only one or two cells were populated and the rest of the cells were zero. They also provided a measure based on the proportion of small cells across the tables. These measures do not provide an accurate quantification of the disclosure risk for a specific table. To obtain an analytical expression of disclosure risk for the entire table (or row/columns), it is natural to use information theory, specifically the entropy.

4.1. An Information Theory Disclosure Risk Measure

As described in Antal et al. (2014), a disclosure risk measure for a census frequency table should have the following properties: (a) small cell values have higher disclosure risk than large values; (b) uniformly distributed frequencies imply low disclosure risk; (c) the more zero cells in the census table, the higher the disclosure risk; (d) the risk measure should be bounded by 0 and 1. Using information theory, we develop an analytical expression of disclosure risk that meets these properties.

Information theory is covered comprehensively in Cover and Thomas (2006). One of the most important measures is the entropy. Let X be a discrete random variable having a

distribution $P = (p_1, p_2, \ldots, p_K)$. The entropy is defined as:

$$H(X) = H(P) = -\sum_{i=1}^{K} p_i \cdot \log p_i$$

If $p_i = 0$ for a category i, the respective term in the sum will be considered 0, since $\lim_{x \to 0} x \log x = 0$. It follows that $H(P) \geq 0$, since $-p_i \cdot \log p_i \geq 0$ with $H(P) = 0$ *iff* the probability mass is concentrated on one point. Therefore, the smaller the entropy *H(P)*, the more likely that attribute disclosure can occur. Under the uniform distribution $U_K = ((1/K), (1/K), \ldots, (1/K))$, we obtain the maximum entropy: $H(U_K) = \log K$ and minimum attribute disclosure risk.

The entropy of the frequency vector in a table of size K, $F = (F_1, F_2, \ldots, F_K)$ where $\sum_{i=1}^{K} F_i = N$ is:

$$H(P) = H\left(\frac{F}{N}\right) = -\sum_{i=1}^{K} \frac{F_i}{N} \cdot \log \frac{F_i}{N} = \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N} \tag{1}$$

To produce a disclosure risk measure between 0 and 1, we define the risk measure as:

$$1 - \frac{H\left(\frac{F}{N}\right)}{\log K}. \tag{2}$$

The disclosure risk measure in (2) ensures property (b) since the term will tend to zero as the frequency distribution is more uniform, and ensures property (d) since the measure is bounded between 0 and 1. However, the disclosure risk measure does not take into account the magnitude of the cells counts or the number of zero cells in the table (or row/column of the table) and does not preserve properties (a) and (c). Therefore, an extended disclosure risk measure is proposed in (3) and is defined as a weighted average of three different terms, each term being a measure between 0 and 1.

$$\begin{aligned} R(F, w_1, w_2) = {} & w_1 \cdot \left[\frac{|A|}{K}\right] + w_2 \cdot \left[1 - \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N \cdot \log K}\right] \\ & - (1 - w_1 - w_2) \cdot \left[\frac{1}{\sqrt{N}} \cdot \log \frac{1}{e\sqrt{N}}\right] \end{aligned} \tag{3}$$

where A is the set of zeroes in the table and $|A|$ the number of zeros in the set, K, N and F as defined above and w_1, w_2 are arbitrary weights: $0 \leq w_1 + w_2 \leq 1$.

The first measure in (3) is the proportion of zeros which is relevant for attribute disclosure and property (c). The third measure in (3) allows us to differentiate between tables with different magnitudes and accounts for property (a). As the population size N gets larger in the table, the third measure tends to zero. The weights w_1 and w_2 should be chosen depending on the data protector's choice of how important each of the terms are in contributing to disclosure risk. Alternatively, one can avoid weights altogether by taking the L_2- norm (see Subsection 4.3) of the three terms of the risk measure in (3) as follows: $\left(\left(\left(\sum_{i=1}^{3} |x_i|^2\right)^{1/2}\right) \Big/ \sqrt{3}\right)$ where x_i represents term i, $i = 1, 2, 3$ in (3).

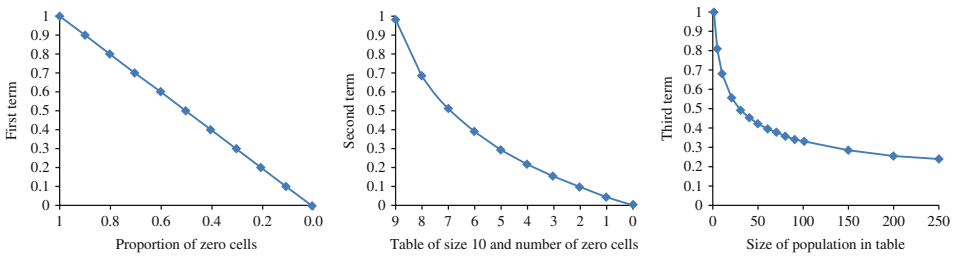


Fig. 1. The three components of the proposed disclosure risk measure in (3)

Figure 1 provides a graphical interpretation of each of the three terms of the proposed disclosure risk measure in (3). The figure on the left shows the first term of the disclosure risk measure as a function of the proportion of zero cells (although a table of all zeros would not be permissible in a flexible table-generating server). The figure in the middle shows the second term based on the entropy in (2) where we demonstrate with a table of ten cells and move from a uniform distribution to a degenerate distribution by accumulating zero cells and spreading the total to the remaining cells. The figure on the right shows the third term of the disclosure risk measure as the size of the population of the table increases.

The final disclosure risk measure (3) is an analytical expression and can be calculated "on the fly" in the flexible table-generating server without the need to see the generated table beforehand. In order to emphasize the risk of identity disclosure arising from small counts (ones and twos), we split the entropy measure as shown in (2) into two parts, small counts up to six and larger counts of seven and more, and provide different weights for each part. For the comparison study in Section 5, the following weights were chosen: $w_1 = 0.1$, $w_{2Part1} = 0.7$, $w_{2Part2} = 0.1$ and $w_3 = 0.1$ where the largest weight is attributed to the entropy term based on small counts. These weights were motivated by the empirical work carried out at the Office for National Statistics on SDC methods for the 2011 UK census tabular outputs, where attribute disclosure and small counts were of the highest concern.

4.2. Modifying the Disclosure Risk Measure After Perturbation

The disclosure risk measure in (3) does not take into account the application of SDC methods and therefore needs to be modified to reflect the uncertainty that is introduced into the counts of the table. Random rounding, for example, eliminates cells of size one and two by introducing more cells of size zero and three in the table, and seemingly increases the risk of attribute disclosure. However, these additional cells of size zero and three are not true counts and the risk of attribute disclosure should decrease. The disclosure risk as measured by the entropy in (2) (and the second term in (3)) does not reflect this uncertainty on whether the cell count is a true value or not. Therefore, we introduce an additional property for the disclosure risk measure following on from those defined in Subsection 4.1: (e) the disclosure risk measure following the application of an SDC method must be less than the original disclosure risk measure. In order to ensure property (e), we propose to modify the first two terms of the disclosure risk measure in (3) after the application of an SDC method as follows:

Modifying the First Term in (3):

The first term in (3) based on the proportion of zero cells can be generalized to compare the number of zero cells in the original and perturbed table. From (3), A is the set of zero cells in the original table and $|A|$ is the number of zero cells in the set. Similarly, let B be the set of zero cells in the perturbed table and $|B|$ the number of zero cells in the set. Denote $A \cup B$ as the union of the sets of zero cells and $A \cap B$ as the intersection of the sets of zero cells in the original and perturbed table. The revised first term in (3), which takes into account that nonzero cells may have been perturbed into zero cells and vice versa, is defined as: $(|A|/K)^{|A \cup B|/|A \cap B|}$. If there are no zero cells in the original table and hence $A \cap B = 0$, then the first term in (3) will remain equal to 0 following perturbation. For example, assume in a table there is a fraction of 0.10 zero cells and following perturbation a fraction of 0.20 zero cells and all original zero cells remain as zero in the perturbed table. In this case, the power term will be 2 and the risk measure following perturbation is reduced to 0.01 from the original 0.10. The modification of the first term in (3) is always less than the original term if nonzero cells are perturbed to zero cells and vice versa, and thus property (e) is ensured.

Modifying the Second Term in (3):

Assume that the possible values in the table are: $0, 1, 2, \ldots, L$ and the frequency of frequencies of these values is denoted by: $(n_0, n_1, n_2, \ldots, n_L)$. The table is perturbed according to a probability transition matrix (for example, the probability transition matrix $\mathbf{P}$ defined in Subsection 3.3). Let the frequency of frequencies of the perturbed values be denoted by: $(n'_0, n'_1, n'_2, \ldots, n'_L)$. For an observed perturbed value j, $j = 0, 1, \ldots, L$, the expected total from the cells of value j can be estimated by the proportion of the original values of j that are not changed: $(j \cdot n_j) \cdot p_{jj}$ and the proportion of other values i, $i \neq j$ that are changed to value j: $\sum_{i \neq j} (i \cdot n_i) \cdot p_{ij}$, so the expected total from cells of value j after perturbation is: $\sum_{i=0}^{L} (i \cdot n_i) \cdot p_{ij}$.

To reflect the uncertainty of the counts in the perturbed table, we replace the observed perturbed cells of value j by the expected total from cells of value j distributed evenly across all cells having the perturbed value j: $\left(\left(\sum_{i=0}^{L} (i \cdot n_i) \cdot p_{ij} \right) / (n'_j) \right)$. As an example, assume the SDC method of random rounding to base 3. We replace the zero cells in the perturbed table with: $[0 \cdot n_0 + 1 \cdot n_1 \cdot (2/3) + 2 \cdot n_2 \cdot (1/3)]/n'_0$. This reflects the fact that zero cells in the perturbed table are not true zeroes; rather, a proportion of them arise from the perturbation of cells of values one and two to zero cells under the probability mechanism, and it is unknown which zero cells are true zero cells and which zero cells are a result of the perturbation. Similarly, for the perturbed cell values of size three, we replace these with the term: $[1 \cdot n_1 \cdot (1/3) + 2 \cdot n_2 \cdot (2/3) + 3 \cdot n_3 + 4 \cdot n_4 \cdot (2/3) + 5 \cdot n_5 \cdot (1/3)]/n'_3$.

For the pretabular method of record swapping, we use a probability transition matrix applied at the cell level of the table for calculating the expectations as explained above, although it is possible that a perturbed table will be equal to the original table if the swapping variable is not involved in generating the table. The expected total from cells of value j in the table after record swapping is: $\sum_{i=0}^{L} (i \cdot n_i) \cdot p_{ij}$, where p_{ij} is a probability

transition matrix with the swap rate on the diagonal and all off-diagonals have equal probability constrained to the sum of the row probabilities being equal to 1. This means that we assume that every cell in the table can be perturbed according to the swap rate and reflects the assumption that an intruder would not know which variables were swapped.

The modification of the entropy term in (2) replaces observed perturbed counts with their expectations according to the probability transition matrix. In particular, true zero cells which did not contribute to the entropy in the original table are now replaced by their expected values. This should lead to a more even distribution of cell counts in the calculation of the entropy and to a general reduction in the disclosure risk measure in (2) following perturbation. As a final adjustment and to further guarantee property (e), we multiply the resulting entropy-based disclosure risk measures in (2) by a multiplier based on the average of the diagonal probabilities of the probability transition matrix. This multiplier reflects a global level of uncertainty introduced into the perturbed cell counts.

4.3. An Information Theory Data Utility Measure

To assess the distance between two distributions, we use the L_2-norm which, when applied to the difference of two vectors, preserves the properties of a distance metric (non-negativity, coincidence axiom, symmetry and triangle inequality). Measuring the distance infers that the smaller the distance, the more information is left in the table. For an arbitrary vector $x = (x_1, x_2, \ldots, x_K)$ the L_2-norm of x is defined as:

$$\| x \|_2 = \left(\sum_{i=1}^{K} |x_i|^2 \right)^{1/2}.$$

Let $P = (p_1, p_2, \ldots, p_K)$ be the original probability distribution of cell counts and $Q = (q_1, q_2, \ldots, q_K)$ the perturbed probability distribution of cell counts. Define: $\sqrt{P} = (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_K})$ and $\sqrt{Q} = (\sqrt{q_1}, \sqrt{q_2}, \ldots, \sqrt{q_K})$. These are not (necessarily) probability distributions but have the property that as vectors, their L_2- norms are 1.

The Hellinger Distance is defined as the L_2 -norm:

$$HD(P, Q) = \frac{1}{\sqrt{2}} \cdot \left\| \sqrt{P} - \sqrt{Q} \right\|_2$$

and is bounded by 0 and 1.

In the case of frequency distributions from census tables, where $F = (F_1, F_2, \ldots, F_K)$ is the vector of original counts and $G = (G_1, G_2, \ldots, G_K)$ is the vector of perturbed counts, and $\sum_{i=1}^{K} F_i = N$ and $\sum_{i=1}^{K} G_i = M$, the Hellinger distance is defined as:

$$HD(F, G) = \frac{1}{\sqrt{2}} \cdot \left\| \sqrt{F} - \sqrt{G} \right\|_2 = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(\sqrt{F_i} - \sqrt{G_i} \right)^2} \tag{4}$$

The Hellinger distance is grounded in Information Theory and takes into account the magnitude of the cells since the difference between square roots of two "large" numbers is smaller than the difference between square roots of two "small" numbers, even if these pairs have the same absolute difference. Naturally, while the lower bound remains zero,

the upper bound of this distance metric changes:

$$HD(F,G) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(\sqrt{F_i} - \sqrt{G_i}\right)^2} = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(F_i + G_i - 2 \cdot \sqrt{F_i \cdot G_i}\right)}$$

$$= \frac{1}{\sqrt{2}} \cdot \sqrt{N + M - 2 \cdot \sum_{i=1}^{K} \sqrt{F_i \cdot G_i}} \leq \sqrt{\frac{N+M}{2}}.$$

Since the SDC methods described in Section 3 produce approximately the same overall population total N due to controlled methods of perturbation, the Hellinger distance is bounded by 0 and $\sqrt{N}$. For the comparison study in Section 5, we use the expression of $1 - (HD(F,G)/\sqrt{N})$ as the data utility measure, which is bounded between 0 and 1, 0 representing low utility and 1 representing high utility.

5. A Comparison Study

In this section we present a flexible table-generating server for census tables where we proceed with the European Census Hub approach of defining a large hypercube as the underlying data input to the server. We compare the application of SDC methods described in Section 3 to four generated output tables and examine the properties of the disclosure risk and data utility measures presented in Section 4.

5.1. Preparing the Underlying Hypercube and Generating Output Tables

For the comparison study, we generate a hypercube with an underlying population of size 1,500,000 individuals for two NUTS2 regions. The variables defining the hypercube follow one of Eurostat's specifications for a hypercube required by all Member States as follows:

- NUTS2 Region – 2 regions
- Gender – 2 categories
- Banded age groups – 21 categories
- Current employment status – 5 categories
- Occupation – 13 categories
- Educational attainment – 9 categories
- Country of citizenship – 5 categories

From the UK Census 2001, cell proportions from published tables were calculated and cross-classified using iterative proportional fitting. We then multiplied the proportions by our population size of 1,500,000 individuals to produce the final hypercube. The hypercube used in the comparison study has 245,700 cells. The distribution of cell counts is skewed with a large proportion of zero cells as seen in Table 1.

The distribution of cell counts in the hypercube as shown in Table 1 was comparable to the hypercube based on real census data produced by the United Kingdom according to the above specification.

Table 1. Distribution of cell counts in the generated hypercube

Cell value	Number of cells	Percentage of cells
0	226,939	92.4
1	4,028	1.6
2	2,112	0.9
3–5	2,964	1.2
6–8	1,664	0.7
9–10	720	0.3
11 and over	7,273	3.0
Total	245,700	100.0

In the flexible table-generating server of our comparison study, we apply a set of preliminary SDC rules for generating tables and allow a maximum of three dimensions with one additional variable to define the population of the table. Four different-size output tables are generated from the input hypercube as follows (number of categories of each variable are in parenthesis):

(1) Selected population: NUTS2 = 1, table spanned by: Banded age group (21) * Educational Attainment (9) * Occupation (13).
(2) Selected population: NUTS2 = 2, table spanned by: Gender (2) * Banded age group (21) * Country of citizenship (5)
(3) Selected population: Gender = 1, table spanned by: Current activity status (5) * Occupation (13) * Educational attainment (9)
(4) Selected population: Banded age group = 10, table spanned by: NUTS2 (2) * Occupation (13) * Educational attainment (9)

Table 2 contains details of the four generated output tables that are used in the comparison study: the total size of the population, the number of cells and the average cell size in each table as well as the distribution of cell counts.

Table 2. Details of four generated tables to be used in the comparison study

Details	Table 1	Table 2	Table 3	Table 4
Total Population	854,539	645,461	736,355	96,656
Number of cells	2,457	210	585	234
Average cell size	347.8	3,073.6	1,258.7	413.1
Number of	%	%	%	%
Zeroes	1,534 (62.4)	49 (23.3)	275 (47.0)	84 (35.9)
Ones	44 (1.8)	14 (6.7)	16 (2.7)	9 (3.9)
Twos	35 (1.4)	2 (1.0)	9 (1.5)	4 (1.7)
Threes	27 (1.1)	5 (2.4)	3 (0.5)	6 (2.6)
Fours	20 (0.8)	4 (1.9)	9 (1.5)	1 (0.4)
Fives	17 (0.7)	1 (0.5)	5 (0.9)	4 (1.7)
Sixes and over	780 (31.8)	135 (64.3)	268 (45.8)	126 (53.9)

From Table 2, output Table 1 is the largest table with the largest proportion of zero cells. Output Tables 2 and 4 are similar in the number of cells but the size of the population is considerably smaller in output Table 4, resulting in a larger proportion of zero cells and a smaller proportion of cells of value one. Output Table 3 is a midsize table. It is clear from the small cell counts and many zero cells that the generated output tables require the application of SDC methods in the flexible table-generating server.

In the comparison study we provide an example of how a statistical agency might go about assessing different SDC methods for a flexible table-generating server of census tables through disclosure risk and data utility measures. In the pretabular approach of protecting the input hypercube prior to generating tables, we apply three SDC methods as follows:

- Full random rounding of the hypercube to base 3 semicontrolled to the two NUTS2 totals.
- Random record swapping carried out by first constructing microdata of individuals from the hypercube where each cell is duplicated to the number of individuals in the cell. A random sample of five percent of individuals is selected in each NUTS2 region, then randomly paired with individuals in the opposite NUTS2 region and their geographical variables swapped. This produced a total swap rate of ten percent of individuals having their NUTS2 regions swapped. Following the record-swapping procedure, the hypercube is reconstructed.
- Stochastic perturbation on the hypercube is based on an invariant probability transition matrix with controls in the overall totals of the two NUTS2 regions. The perturbation is carried out on cells of values in the range 0–10; all cells above a value of 10 have the same probabilities of perturbation depending on their residual value to base 5. The probability transition matrix for each NUTS2 region used in this study is presented in Table 3.

The pretabular disclosure-controlled hypercubes are used as input to the flexible table-generating server and the four output tables generated under each SDC method. The comparison results also include the case where a post-tabular SDC method of semicontrolled random rounding to base 3 is applied directly to the four output tables that are generated from the original unperturbed hypercube. The SDC methods are compared through the disclosure risk and data utility measures described in Section 4.

5.2. *Results of the Comparison Study*

To compare the pretabular SDC methods applied to the original hypercube (record swapping, semicontrolled random rounding and stochastic perturbation), we first assess the impact of the perturbation on the small cells in the generated output tables. Table 4 presents the number of small cells of size 1 and 2 in the original hypercube and in each of the four output tables defined in Subsection 5.1, and the percentage of those cells that were altered under the SDC methods. Record swapping generally provided the least number of small cells perturbed except for output Table 4, where the swapped variable NUTS2 is used as a spanning variable of the table. Output Table 3 did not include the swapped NUTS2 variable and hence all cells in the table contain original cell counts. Random rounding eliminates all small cells of size 1 and 2 and provides more protection compared

Table 3. Probability transition matrix used to perturb the hypercube in each NUTS2

		Perturbed Counts											Residual of count to base 5 is:				
Cell Value		0	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5
Original Counts	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0.5000	0.1250	0.1250	0.1250	0.1250	0	0	0	0	0	0	0	0	0	0
	2	0	0.1250	0.5000	0.1250	0.1250	0.1250	0	0	0	0	0	0	0	0	0	0
	3	0	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0.0167	0	0	0	0	0	0	0	0
	4	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0	0	0	0
	5	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0	0	0
	6	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0	0
	7	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0
	8	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0
	9	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0
	10	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0
Residual of count	1	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0
to base 5 is:	2	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167
	3	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167
	4	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.0167	0.0167	0.9000	0.0167
	5	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.9000

Table 4. Number of small cells of size 1 and 2 in original hypercube and generated tables, and percentage of those cells that were perturbed

	Original hypercube	Table 1	Table 2	Table 3	Table 4
Number of cells of size 1 and 2	6140	79	16	25	13
Percentage perturbed:					
Record swapping	26.9	15.2	12.5	0	30.8
Stochastic perturbation	33.2	29.1	25.0	36.0	23.1
Random rounding	100	100	100	100	100

to record swapping and the stochastic perturbation. It is well known, however, that random rounding has the risk of being able to reveal original cell values, especially when the sum of rounded cells does not add up to the rounded marginal totals. However, ensuring the consistency of the rounding across same cells in different tables and controlling some of the marginal totals lowers the risk of being able to reveal original cell values.

Table 5 presents the disclosure risk measure in (3) and the Hellinger distance in (4) for the output tables defined in Subsection 5.1 generated on the pretabular disclosure-controlled hypercubes according to the SDC methods: record swapping, semicontrolled random rounding and stochastic perturbation. In addition, we report the measures for the case where the SDC method of semicontrolled random rounding is applied directly to the output tables that were generated from the original hypercube to compare the pretabular and post-tabular approach for this SDC method.

To modify the second term in the disclosure risk measure in (3) following the SDC methods as described in Subsection 4.2, we used the following multipliers: for record swapping, the average diagonal probability of the probability transition matrix is 0.9; for the stochastic perturbation, the average diagonal probability of the probability transition matrix is 0.75 for the small counts and 0.9 for the large counts; for the random rounding to base 3, we use the multiplier of 0.33.

From Table 5, we see that the disclosure risk measures are all smaller for the perturbed tables compared to the original tables, even for the case of record swapping in output Table 3 where the perturbed table is identical to the original table since the perturbed NUTS2 variable was not included as a spanning variable of the table. The utility measures are all high, showing that all SDC methods can provide tables that are fit for purpose for users.

In general, it is clear that the method of record swapping when applied to the input hypercube did little to reduce disclosure risk in the final output tables in the comparison study. However, the disclosure risk measure is always slightly smaller than the disclosure risk measure of the original table to reflect the uncertainty in the table based on the assumption that an intruder cannot be certain which variables were swapped. The data utility measure based on the Hellinger distance for output Table 3 under record swapping is 1.00, since the perturbed table is equal to the original table. The data utility measure under record swapping was low for the two output Tables 1 and 2 where the perturbed

Table 5. Disclosure risk and data utility (Hellinger distance) for the generated tables

	Disclosure risk $R(F, w_1, w_2)$ in (3)	Data utility $1 - (HD(F, G)/\sqrt{N})$ in (4)
Table 1		
Original	0.318	-
Perturbed input		
Record swapping:	0.282	0.988
Semicontrolled random rounding	0.137	0.991
Stochastic perturbation	0.239	0.995
Perturbed output:		
Semicontrolled random rounding	0.135	0.993
Table 2		
Original	0.248	-
Perturbed input:		
Record swapping	0.191	0.972
Semicontrolled random rounding	0.070	0.996
Stochastic perturbation	0.210	0.998
Perturbed output:		
Semicontrolled random rounding	0.072	0.996
Table 3		
Original	0.339	-
Perturbed input:		
Record swapping	0.295	1.000
Semicontrolled random rounding	0.130	0.994
Stochastic perturbation	0.254	0.996
Perturbed Output:		
Semicontrolled random rounding	0.127	0.996
Table 4		
Original	0.298	-
Perturbed input:		
Record swapping	0.271	0.987
Semicontrolled random rounding	0.105	0.991
Stochastic perturbation	0.229	0.994
Perturbed output:		
Semicontrolled random rounding	0.105	0.992

NUTS2 variable was used to select the population for these tables. The data utility measure under record swapping for output Table 4 was slightly higher, since in this case NUTS2 was a variable spanning the table and hence did not change the overall total of the table.

The stochastic perturbation carried out on the input hypercube outperformed record swapping with smaller disclosure risk measures and higher data utility measures (except for output Table 3). The stochastic perturbation has a higher disclosure risk compared to semicontrolled random rounding, since a large percentage of small cells are unchanged by the perturbation, but it has higher data utility.

The semicontrolled random rounding outperformed all other methods with respect to the lowest disclosure risk, since there are no small cells in the tables and attribute disclosure risk is reduced by the introduction of random zeros. However, the data utility measure based on the Hellinger distance was slightly lower compared to the stochastic perturbation method as mentioned above. There was little difference between the disclosure risk measures comparing the pretabular semicontrolled random rounding on the input hypercube to the post-tabular semicontrolled random rounding applied directly to the output tables generated from the original hypercube. However, there is an increase in the data utility measure when applying the post-tabular semicontrolled random rounding, especially for the large output Table 1 and midsize output Table 3.

Figure 2 presents a disclosure risk-data utility map of the four generated tables where RS is record swapping, SP is the stochastic perturbation, RR is the semicontrolled random rounding on the input hypercube and RRP is the semicontrolled random rounding applied directly to the generated output tables. The data utility measure is the Hellinger distance in (4). The upper right-hand quadrant of the map represents high disclosure risk and high utility and the lower left-hand quadrant represents low disclosure risk and low data utility.

The statistical agency needs to decide on a tolerable disclosure risk threshold above which they are not prepared to release a table. As an example, the disclosure risk-data utility map shows that for a tolerable disclosure risk threshold of up to 15 percent, the

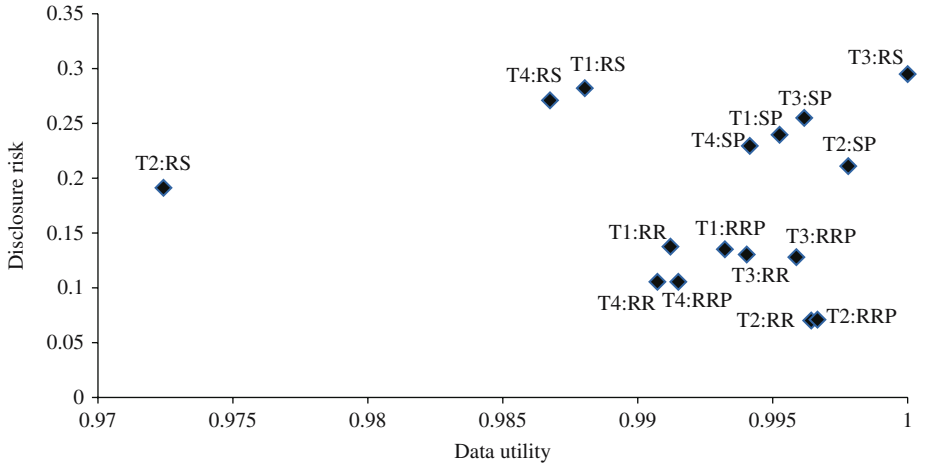


Fig. 2. Disclosure risk – data utility map for generated tables (output Table 1 (T1) to output Table 4 (T4)): RS – record swapping, SP – stochastic perturbation, RR – semicontrolled random rounding on input hypercube, RRP – semicontrolled random rounding on generated tables

output tables where semicontrolled random rounding was applied directly to tables that were generated from the original hypercube have the highest data utility as they are on the farthest right-hand side of the map.

6. Concluding Remarks

In this article, we have compared pretabular SDC methods applied to a large hypercube (record swapping, stochastic perturbation and semicontrolled random rounding) and a semicontrolled random rounding applied directly to output tables generated from the original hypercube. For the pretabular SDC methods, record swapping had little impact on reducing disclosure risk and also had lower data utility. Semicontrolled random rounding offered more protection as all cell values in the table not a multiple of base b are perturbed, and by preserving the consistency of cells across tables, it is more difficult to undo the rounding to reveal original cell values. The stochastic perturbation had the best overall data utility, but entailed higher disclosure risks compared to the semicontrolled random rounding. Finally, we have seen that the post-tabular SDC method of semicontrolled random rounding applied directly to the generated output tables produced nearly the same amount of disclosure risk reduction as the pretabular semicontrolled random rounding applied to the input hypercube, but had a higher level of data utility.

The aim of the comparison study was not primarily to evaluate specific SDC methods or indeed determine their optimum parameterization, but rather to demonstrate how such a disclosure risk and data utility analysis should be carried out by a statistical agency when disseminating census data. To this end, we have proposed new global measures of disclosure risk and data utility based on information theory that are particularly suited for assessing disclosure risk arising from attribute and identity disclosure in census frequency tables and can easily be embedded in a web-based flexible table-generating server. The proposed modifications to the disclosure risk measure following the application of an SDC method show that we can reflect the level of uncertainty added to the tables and therefore reduce the disclosure risk. Further research is needed to refine and improve post-tabular SDC methods whilst preserving additivity and consistency of user-defined tables. More extensive empirical studies are needed that involve real data and the testing of SDC methods across their respective parameter spaces.

Another key aspect of the SDC problem in a flexible table-generating server is the management of users and governance processes. The server can be freely available on the statistical agency's website for all users or restricted via licensing and passwords to only approved users. For the former case, it is clear that SDC rules and methods would have to be highly protective to guard against the fact that users can query the same table multiple times in an attempt to undo SDC methods and reveal original cell counts. Clearly, perturbative SDC methods, preserving the additivity and consistency of same cells across different tables, and high thresholds for dissemination would be required. For the latter case, less protection would be needed, allowing for higher-quality data, but protocols would then need to be in place to handle multiple overlapping queries from the same user, the management of users and their expectations.

7. References

Antal, L., N. Shlomo, and M. Elliot. 2014. "Measuring Disclosure Risk with Entropy in Population Based Frequency Tables." In *PSD'2014 Privacy in Statistical Databases*, edited by J. Domingo-Ferrer, 62–78. Berlin: Springer.

Cover, T.M. and J.A. Thomas. 2006. *Elements of Information Theory*, 2nd ed. New York: Wiley.

Dalenius, T. and S.P. Reiss. 1982. "Data Swapping: A Technique for Disclosure Control." *Journal of Statistical Planning and Inference* 7: 73–85.

Doyle, P., J.I. Lane, J.M.M. Theeuwes, and L. Zayatz. 2001. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: Elsevier Science B.V.

Duncan, G., M.J. Elliot, and J.J. Salazar. 2011. *Statistical Confidentiality: Principles and Practice*. New York: Springer.

Fienberg, S.E. and J. McIntyre. 2005. "Data Swapping: Variations on a Theme by Dalenius and Reiss." *Journal of Official Statistics* 9: 383–406.

Fraser, B. and J. Wooton. 2005. "A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing." Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, November 9–11. Available at: www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.35.e.pdf (accessed April 2015).

Gouweleeuw, J., P. Kooiman, L.C.R.J. Willenborg, and P.P. De Wolf. 1998. "Post Randomisation for Statistical Disclosure Control: Theory and Implementation." *Journal of Official Statistics* 14: 463–478.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer, and P.P. de Wolf. 2012. *Statistical Disclosure Control*. Chichester: John Wiley & Sons.

Salazar-Gonzalez, J.J., C. Bycroft, and A.T. Staggemeier. 2005. "Controlled Rounding Implementation." Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, November 9–11. Available at: www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.36.pdf (accessed April 2015).

Shlomo, N. 2007. "Statistical Disclosure Control Methods for Census Frequency Tables." *International Statistical Review* 75: 199–217. Doi: http://dx.doi.org/10.1111/j.1751-5823.2007.00010.x.

Shlomo, N. and C. Young. 2008. "Invariant Post-tabular Protection of Census Frequency Counts." In *In PSD'2008 Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and Y. Saygin, 77–89. Berlin: Springer.

Willenborg, L.C.R.J. and T. de Waal. 2001. *Elements of Statistical Disclosure Control*. New York: Springer.

Received July 2013
Revised October 2014
Accepted November 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 325–347, http://dx.doi.org/10.1515/JOS-2015-0020

Statistical Metadata: a Unified Approach to Management and Dissemination

Marina Signore[1], *Mauro Scanu*[1], *and Giovanna Brancato*[1]

This article illustrates a unified conceptual approach to metadata, whereby metadata describing the information content and structure of data and those describing the statistical process are managed jointly with metadata arising from administrative and support activities. Many different actors may benefit from this approach: internal users who are given the option to reuse information; internal management that is supported in the decision-making process, process industrialisation and standardisation as well as performance assessment; external users who are provided with data and process-related metadata as well as quality measures to retrieve data and use them properly. In the article, a general model useful for metadata representation is illustrated and its application presented. Relationships to existing frameworks and standards are also discussed and enhancements proposed.

Key words: Generic Statistical Information Model (GSIM); Generic Statistical Business Process (GSBPM); organisational and support processes; SDMX; paradata; quality indicators.

1. Introduction

Metadata systems are extremely important management tools for all public and private institutions managing statistical data, and especially for today's National Statistical Institutes (NSIs). These systems are vital for the dissemination of statistics; they are the pillars of statistical data warehouses, and are gaining increasing importance in statistical data exchange among different organisations, especially at supranational level, where they are used in addition to facilitate communication and a common understanding of the data being exchanged. Important metadata system examples are described in the 18 statistical metadata case studies available in the United Nations – Economic Commission for Europe (UNECE) statistics wiki METIS (www1.unece.org). Many NSIs rely on sets of specific metadata systems, such as the eight metadata systems declared by Statistics Sweden, including one for classifications (KDB), one for products (the product data base) and one for governance (FMOD, applied in the management of processes related to the statistical production process). Others are mainly composed of a unique centralised system, such as the Statistics Canada Integrated Metadata Base (IMDB), obtained through the consolidation of already existing separate metadata systems, whose aim is to display data sources and methods for each statistical programme and survey, collecting reference metadata including survey methodology and data accuracy, definitions of concepts, variables and classifications for all subject-matter areas (domain-specific metadata

[1] ISTAT, Via Cesare Balbo 16, Rome 00184, Italy. Emails: signore@istat.it, scanu@istat.it, and brancato@istat.it

systems for business or social surveys exist, but they are highly integrated with IMDB, see, Greenhough et al. 2014). All these metadata systems strive to rely on conceptual models that act as current international standards: for example, European NSIs should be compliant with the European Statistics Code of Practice (CoP) that sets the standard for developing, producing and disseminating European statistics (Eurostat 2011). More generally, international institutions and NSIs worldwide agree on the Generic Statistical Business Process Model – GSBPM (UNECE 2013a) and the Generic Statistical Information Model – GSIM (UNECE 2013b). However, these standards and the related metadata systems can still be improved with regard to the relationships between different metadata typologies, such as those specifying the meaning of data and those describing the underlying processes (Signore et al. 2013a). The Generic Activity Model for Statistical Organizations – GAMSO (UNECE 2015), while only at version 0.2, seems to be a promising endeavour in connecting activities such as Strategy, Capability and Corporate Support to statistical production, thus supporting the views expressed in the present article.

In this article, we propose a unified conceptualisation of metadata that allows a thorough description of data and processes (both of statistical and support nature). Well-known metadata classes are reorganised and the role of business-related metadata, that is, metadata supporting the management of statistical organisations, is emphasised.

More specifically, three main metadata typologies (i.e., metadata related to data structure and content, process-related and business-related metadata) are discussed in this article. They are introduced in Section 2 and compared with the GSIM conceptual framework. We propose a unified metadata management system that comprises activities both at process/product level as well as at the institutional level. In this respect, this approach could be useful for assessing, for instance, the compliance of ESS Members to the European Statistics CoP that involves assessment at the abovementioned levels. One example of the applicability of this model is furnished by the unified metadata system developed by the Italian National Institute of Statistics (Istat) described by Signore et al. (2013b).

In Section 3, we introduce a metadata conceptualisation useful for connecting the three metadata typologies to the statistical process phases. After this, we present a model for describing metadata related to data structure and content as well as data transformations along the statistical process (Section 4). Existing limitations in standards such as GSIM are discussed and possible enhancements are proposed.

Process-related metadata, including quality indicators, and business-related metadata are described in more detail in Sections 5 and 6, respectively. As a result of the proposed approach, the ties between metadata and quality are given greater importance. Even though there are increasing examples of a joint use of process-related metadata and quality indicators (see, for instance Götzfried et al. 2011), a common conceptualisation is useful for both disciplines, which to a large extent developed independently of one another. For this reason, we propose some enhancements to GSBPM. One consequence of such an integrated approach is a broader perspective: quantitative information (i.e., quality indicators and paradata, Couper 1998) should be considered alongside the above-mentioned typologies of metadata (qualitative information).

In Section 7, we comment on the joint use of process- and business-related metadata for improving statistical processes and products. We conclude with some final remarks in Section 8.

2. A Unified Metadata Management System

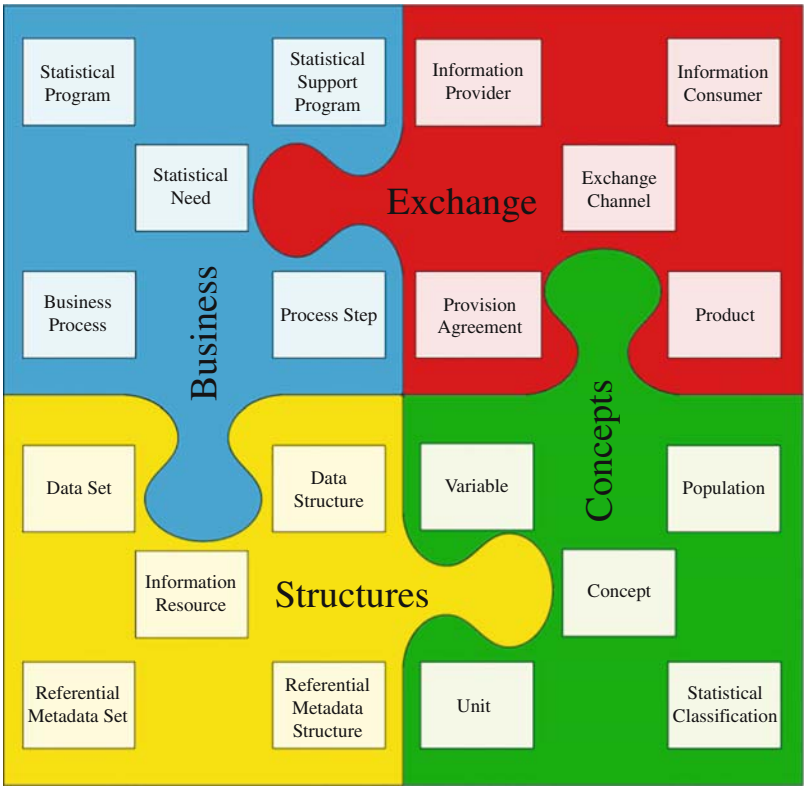
In our approach, metadata are viewed as part of a wider system where different metadata typologies are related to one another and connected to quality.

Hence the need to explicitly consider three metadata categories is advocated:

- *Metadata related to data structure and content* include all that is necessary to give a definition and a meaning to statistical data;
- *Process-related metadata* describe the statistical business process in terms of methods (e.g., sampling, collection methods, editing processes) and quality (e.g., timeliness, accuracy).
- *Business-related metadata* are useful for the management of an NSI in planning, executing and assessing both statistical and support activities. These metadata allow stakeholders to *connect* data and processes with NSI's strategic objectives (e.g., long-term goals) and with the different management plans of each Institute (e.g., methodological investments); to *assign* responsibilities and resources to the objectives; to *plan* schedules and timetables of different actions; to *evaluate* the achievement of objectives and to *assess* performance and efficiency.

This approach matches the definitions provided by Androvitsaneas et al. (2006) and SDMX (2009), as well as the information objects identified in GSIM (UNECE, 2013b), but is more suitable for a unified vision. In particular, metadata related to data structure and content correspond to respectively structural and conceptual reference metadata in Androvitsaneas et al. (2006) and SDMX (2009) terminology. Process-related metadata correspond to methodological and quality-reference metadata. Business-related metadata are not explicitly considered in Androvitsaneas et al. (2006) and SDMX (2009); instead, they are identifiable within the business top-level group of GSIM, although they are not fully defined. Figure 1 maps the metadata categories as defined in this article and the main information objects included in the four top-level GSIM groups.

This classification calls for a unified metadata management system as outlined in Figure 2, in which metadata typologies are linked to the institutional macro phases and to generic statistical process phases. Connections with the areas of the European Statistics Code of Practice (Eurostat 2011), namely institutional environment, statistical processes and statistical outputs, are also represented in Figure 2. In particular, *planning* and *assessment* are performed at an institutional level, thus reflecting strategic and overall activities, while *execution* pertains to the generic statistical process (and related outputs). The latter also includes a specific planning phase, identified by the GSBPM phases (UNECE 2013a) "Specify needs", "Design" and "Build", as well as a proper assessment phase, "Evaluate", which are all performed at a process level. The need to go beyond the generic statistical process to model metadata was already recognised by Androvitsaneas et al. (2006), who introduced the concept of the statistical system as a whole and of quality dimensions such as metadata on objectivity and credibility (pertaining to the institutional level). According to Sundgren (2004), the metadata managed in a system should be captured as early as possible in the process of developing, implementing and operating a production system; moreover, the same metadata should not be captured more than once.



GSIM top-level groups and information objects		Proposed classification
BUSINESS	Statistical Program	Business-related metadata
	Statistical Support Program	
	Statistical Need	
	Business Process	Process-related metadata
	Process Step	
EXCHANGE	Information Provider	
	Information Consumer	
	Exchange Channel	
	Provision Agreement	
	Product	
STRUCTURES	Referential Metadata Set	
	Referential Metadata Structure	
	Information Resource	
	Data Set	Metadata related to data structure and content
	Data Structure	
CONCEPTS	Variable	
	Population	
	Concept	
	Unit	
	Statistical Classification	

Fig. 1. Mapping between the GSIM information objects and the proposed classification

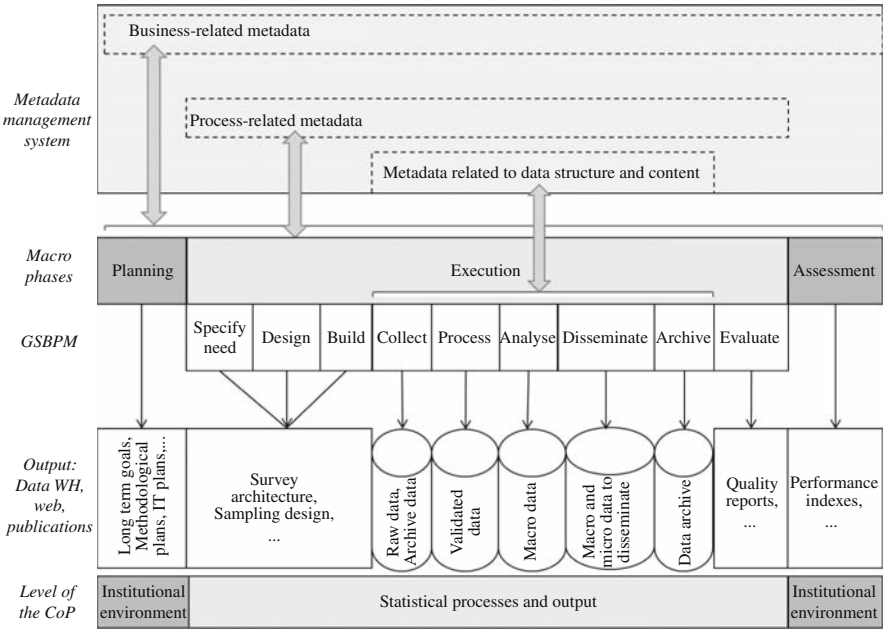


Fig. 2. The Istat metadata management system

One possible application of the proposed model is the Istat Unified Metadata System (Sistema Unitario di Metadati, SUM; described by Signore et al. 2013b). SUM is based on the following key concepts and objectives: data retrieval and usability (by associating proper meaning to data as well as methodological and quality information); metadata reuse (in order to harmonise concepts and reduce documentation burden); traceability (with the objective of statistical process transparency and process automation); integration (with the aim of making the different sectors within a statistical institute speak with one voice and support standardisation).

3. Modelling Metadata in the Statistical Business Process

The model for metadata identification described in the present section highlights the relationships between metadata related to data structure and content and the process- and business-related metadata that will be described in the next sections.

The proposed approach, properly adapted to building a model for statistical business process description and relative metadata, traces back to Saint Thomas Aquinas's eight circumstances determining the ethic of an action, and is commonly used in the fields of journalism and police investigation as questions for gathering basic information. This paradigm is not new and has also been used in similar contexts, for example, in metadata classification (Statistics Sweden 2008) and in other applications, for example, in the definition of a framework for Enterprise Architecture (Zachman 1987).

According to GSIM (UNECE 2013c), a business process is "[t]he set of Process Steps to perform one or more Business Functions to deliver a Statistical Program Cycle or Statistical Support Program". In other terms, the process steps can be viewed as a sequence

of logically connected elementary operations, or subprocesses in the GSBPM terminology (UNECE 2013a), aimed at producing statistical information, that is, statistical product or output. Besides these, other support processes exist that can be directly oriented towards the production goal or that are more cross-sectional, such as staff training. GSIM partly recognises the role of such processes, introducing the Statistical Support Program, whereas they are not explicitly included in GSBPM.

In order to categorise metadata within the statistical business process or relate them to it, the following fundamental questions should be answered: *cur, quid, quomodo, quibus auxiliis, quis, ubi, quando,* and *quantum* (respectively: why, what, how, by which means, who, where, when and how much). This set of questions allows us to exhaustively represent a circumstance, fact, situation and, in the field of official statistics production, a statistical business process. However, additional specifications have to be added concerning the quality dimension, that is, the level of performance attained and the fitness for use of the produced statistical data.

***Cur* (why) –** In the philosophical interpretation, this element represents the cause of the action. Translating this concept into the metadata model, it represents the main production objective of the statistical process, responding to both users' needs and national or international regulations.

***Quid* (what) –** The *what* element identifies data at any stage of the production process. It can refer to final products as well as to intermediate outputs. Data are formally defined by metadata related to data structure and content (Section 4) and described for users by process-related metadata (Section 5).

***Quomodo* (how) –** The *how* component is the pivotal element, representing the way activities are carried out to meet the *why* objective. These activities can vary in nature: *i)* statistical operations; *ii)* management or administrative activities (included in GSIM as Statistical Support Program); *iii)* quality control actions.

Statistical activities are operationalised in elementary operations describing a statistical process, usually arranged in hierarchical structures, thus facilitating their organisation (UNECE 2013a; Brancato and Simeoni 2012). They represent the key elements for the description of statistical business processes; together with other components (*who*, *by which means*) they constitute the core of process-related metadata. Depending on the objective, elementary operations can have a more generic nature, as in GSBPM, or can be more detailed to provide hints about the methodology, as in the Istat documentation system (Brancato and Simeoni 2012). A thorough description of the methodology supports the construction of metadata related to data structure and content on the transformations the data undergo (see, Section 4). For example, in the phase "Process", process-related metadata describe that a given editing and imputation procedure has been applied, for instance "Deterministic error and outlier detection and imputation based on deterministic rules (IF-THEN)"; metadata related to data structures and content describe the variables involved and the exact parameters of the rules, thus supporting a full understanding of the data.

However, there are a number of support activities of an administrative and management nature concurrent with the statistical production. One example is the procedure needed to select a private company to carry out data collection in outsourcing. These operations are among the key elements of business-related metadata (see, Section 6).

Finally, during survey execution, an extensive number of activities is devoted to quality controls, constituting the third type of the *how* elements. Similarly to statistical operations, quality control activities can be structured hierarchically (Brancato and Simeoni 2012; D'Angiolini et al. 1998).

Each *how* element can be ordered. Although, as already pointed out, operations or subprocesses follow a logical sequence, they may be performed in different orders. In addition, some operations can be performed more than once, generating cycles (UNECE 2013a).

The *how* elements are strictly connected to the *what* elements, the latter representing inputs and outputs (UNECE 2013a). The output of each operation or activity is the result of a transformation process and represents the input for the next one.

***Quibus auxiliis* (by which means) –** In the proposed model, this element is represented by the IT tools used to perform the elementary operations, that is, a software application or a system, implementing one or more methods. For example, for the CAPI technique (*how*), the software used to implement the electronic questionnaire (*by which means*) can be specified. IT tools' characteristics and rules for their use within the administration are usually documented and stored in software and application catalogues.

***Quis* (who) *and ubi* (where) –** The *who* element represents what GSIM identifies as the Agent. The following *who* elements can be defined: *i)* the person/s responsible for the operation; *ii)* the person/s executing the operation; *iii)* other actors involved in the operation.

The *where* element specifies the organisational unit where the activity is carried out and is strictly related to the *who* element.

For instance, persons in charge of and performing the activities may belong to the same entity, but not necessarily so. In addition, depending on the organisational structure, an agent such as the interviewer may belong to different units within the statistical office, or may depend on an outsourcing company in charge of the interviewing activity.

The *who* and the *where* elements may assume increasing importance given the tendency to abandon the classical stovepipe production process model in favour of adopting more integrated production models, the former being characterised by a concentration of responsibilities, the latter by more distributed tasks. When external bodies are involved in process operations, the identification of responsibilities may be attributed in administrative documentation, such as contracts setting out the terms and conditions of the service (e.g., when resorting to outsourcing). Finally, some elementary operations may involve actors external to the Institution, such as the reporting unit in the data collection activity.

***Quando* (when) –** The *when* element is vital metadata to properly document validity of elementary operations, as well as changes in procedures and methodologies. Each elementary operation and its characteristics therefore need to be associated to a time reference or a validity period, as also proposed by Banca D'Italia (2007).

Periodicity is another *when* element, that is, the frequency of an operation (release, estimate, etc.). In the majority of statistical processes, all operations (e.g., data collection, processing and dissemination) are performed according to the same periodicity (monthly, yearly, etc.), although there are cases in which this is not true. In addition, the concept of estimate periodicity can be defined, that is, the time frequency of results or estimates released to users. It corresponds to the Frequency of data compilation in the Metadata Common Vocabulary terminology (SDMX Content-oriented guidelines 2009).

***Quanto* (how much) –** This element is proposed to document resources (human, IT) and costs (financial, time) required to carry out statistical, administrative and quality control activities. It represents the typical business-related metadata useful for measuring efficiency and performing cost analysis.

4. Metadata Related to Data Structure and Content

Metadata related to data structure and content define statistical data. These data are usually structured in a table of macro data or a data set of micro data. Hence, their primary content consists of metadata defining *i*) the structure and *ii*) the structure components of the data structure by means of the appropriate concepts. These metadata are defined along the lines of well-established models, such as UNECE (1995) and GSIM (UNECE 2013c). The approach outlined in this article pinpoints, and wherever necessary reorganises, some aspects of these models in order to foster industrialisation. The main aim is to ensure data traceability along a data production process, and metadata reuse in different phases of a process and among processes. In order to be able to trace metadata along the production process, we support the GSIM idea of describing data as the input or output of a process, so that:

- any output is a function of one or more inputs;
- the function transforming the inputs into an output is an operator, that is, a method where all the parameters characterising a transformation should be carefully described, so the same result can be reproduced whenever necessary.

These operators include: population subsetting, variable transformations (variable sums, new categorisations including coding, etc.), microdata integration, imputation and editing of micro- or macrodata, transformations useful to avoid data disclosure, transformations from micro- to macrodata (averages, medians, variable totals, concentration indexes, variances, etc.), transformations from macro- to macrodata (balances, ratios, index numbers, percentage variations, etc.).

Following this approach, any output is connected to other data according to a backward-oriented "plug in" approach (from dissemination to collection). This is consistent with UNECE (1995), making explicit the way data are transformed through the operator (using notation in UNECE, 1995, a macrodata output is defined by the formula $< O(t), v(t), f >$ where $O(t)$ are microdata at time t, $v(t)$ are variables observed on the units of $O(t)$ and f is an operator of aggregation from micro- to macrodata). The use of a syntax based on statistical concepts is appropriate for describing the input and output content of an operator, enhances the possibility of reusing the same metadata throughout the whole data production process and allows metadata harmonisation between different data production processes. This syntax is mostly available in GSIM and includes the whole set of Concept metadata, some GSIM Structure Metadata (i.e., those with a specific focus on data structures both for micro- and macrodata), and specifications of other concepts related to data transformation along the process phases, such as the GSIM Process Step and Rule.

As a matter of fact, GSIM does not directly address transformation tools, instead leaving them to the GSBPM description, which is usually quite generic and not operative. These aspects are deemed to be crucial to enable a statistical system to trace data and the corresponding metadata transformations along the statistical process. For this reason, the

main concepts for describing the content of a data structure are proposed in the following, pinpointing the main differences with the corresponding terms in GSIM.

Reference population – Data (either micro- or macrodata) may have a reference population, that is, the set of individuals, entities or objects on which a set of phenomena are observed (for microdata) or to which a statistical operation is applied in order to obtain macrodata. The concept of reference population includes the GSIM concepts Unit, Unit Type and Population. The distinction between Population and Unit becomes apparent once a population term is associated to either a microdatum or a macrodatum, respectively. As far as Unit Type is concerned, this concept is managed through hierarchical (subgroup) relationships as well as relationships between entities.

Variable – Our concept is in line with GSIM. The variable Value Domain specifies the statistical variable type (categorical or numerical/quantitative).

Classifications – Classifications are managed mostly, though not completely, as in GSIM. Classification versions are organised in levels containing mutually exclusive and collectively exhaustive categories. The most relevant exception compared to GSIM is the possibility of defining a classification version not only in a strictly hierarchical way (i.e., lower-level categories are subcategories of a category the next level up). In this context it is mandatory for a classification to have a lowest level of mutually exclusive and collectively exhaustive categories that are not split into subcategories (elementary events, as defined in probability theory). Any other category in use for the definition of at least one figure can be obtained by the set operations of complementation, union of and intersection of finitely many sets of events (i.e., categories in the algebra of the events as defined in probability theory). These categories should be included either in the classification (if a level of higher order than the lowest with such a category can be defined) or in a classification variant. Higher-order levels actually in use are not necessarily organised in a hierarchical way: complex graphs can describe the relationship between levels. Classification variants should not necessarily fulfil the rule of being composed of level(s) consisting of mutually exclusive and collectively exhaustive categories, as is the case for the list of classification categories actually in use in a disseminated contingency table.

Operator – This concept is useful for defining data on the basis of the way it was transformed from the previous phase in the statistical process, with all its parameters.

Data Content (Measure) – Data Content is a key aspect to be included among metadata related to data structure and content: it is defined as the combination of input data and method of transformation (statistical operator), and feeds the GSIM concept Measure in a data structure. Based on the description of output as a function of input, this approach identifies different forms of the Data Content. In the case of macrodata, the Data Content (macrodata output) has a simple and a composite form. Simple macrodata result from the combination of reference population, a numerical variable of interest (if not present, the counting variable) and the statistical operator used to pass from the micro- to the corresponding macrodata: this is the case, for instance, for average household income (households = population; income = statistical variable; average = statistical operator). Composite macrodata are instead described by the operation between already existing macrodata, as is the case for index numbers, ratios, balances, etc.: for instance, the net occupancy rate of beds in hotels and similar accommodation is defined as the ratio (statistical operator) between the nights spent

in hotels and similar accommodation in a month (first macrodatum) and the number of beds on offer in the given month (second macrodatum). Each component of a composite macrodatum inherits its detailed description as a simple (as in the previous example) or composite macrodatum, depending on their nature. The Data Content for microdata results from the combination of reference population and the set of numerical variables associated to the population units, as well as the statistical operator used for their production. The importance of the interaction between these elements is clearly stated in metadata literature (e.g., in Section 1.2.2 in UNECE 1995, and references therein), and traces back to the foundation of statistical inference: for example, Fisher (1925) states that "Statistics may be regarded *i)* as the study of populations, *ii)* as the study of variation, *iii)* as the study of methods of the reduction of data", and gives definitions of and justifications for this in the introduction to the monograph. While GSIM declares that "measures correspond to Represented Variables with uncoded Value Domains (Described Value Domains)" (UNECE 2013b, item 101), the proposed approach is to code each output, relate it to its statistical description (in terms of reference population, statistical variables, statistical operator and so on) and allow its use as a tool for easing metadata traceability and reuse.

Data Structure – Data structure is the detailed description of the content of a data set of either macro- or microdata produced alongside the statistical process. There are essentially four distinct groups of data produced in a statistical process: collected data (Collect phase); validated data (Process phase); data analysed by means of a statistical procedure (Analysis phase); disseminated data (Dissemination phase). Data sets in the previous four phases can be also archived; however, this is an overarching activity (see, GSBPM) and not relevant in this framework. In line with GSIM, two mandatory components are to be included in a data structure: measure and dimensions. The measure defines the actual content of the table, and is consequently described by the Data Content (with the indicator or the list of indicators of a macrodata set, or the variables observed on a microdata set). The dimensions are aimed at clearly defining how the measure is further disaggregated according to a set of categorical statistical variables or other coded concepts. Macrodata dimensions consist mainly (but not only) of a set of categorical variables crosscutting the measure (for instance average household income per region, number of household components and household typology). Microdata also include a unit identifier among dimensions, in order to appropriately list the set of microdata records in the data set and possibly link them with those in other data sets. A fundamental dimension of both micro- and macrodata sets is time. In order to better specify the content of a data set, or of each data-set cell, a data structure can also have attributes such as the observation status (if a number is provisional or definitive) or the confidentiality status of a cell.

Other concepts – In addition to the previous list of metadata, our approach also considers all concepts useful for a data structure description, such as those related to time (from the most general as frequency and time period, to specific ones such as school year or the edition of national accounts data), to visualisation issues (number of decimals, unit multiplier, etc.) and comprehension issues (observation status, confidentiality status, etc.) with their corresponding code lists.

This type of metadata organisation as well as the strict definition of Data Content represents a useful tool for tackling a number of issues. First of all, the statistical content of a data set is maintained in a unique list (Data Content) and it is therefore easy to see actual

outputs (or, in other words, the set of products obtained in each phase), while there is no need to consider intersections between different metadata in order to reach this goal. Secondly, users find the content of each number in a data set in the description of a single item of the Data Content dimension, simplifying the understanding of data: in other words, our approach does not allow for more than one GSIM measure in a data structure, and this measure is organised as a Data Content. Furthermore, the organisation of Data Content helps data producers to always be exhaustive in the data definition, without taking anything for granted. Finally, the organisation of the data structure as explained above addresses one of the aspects highlighted in Gelsema (2012): describing the connection between data and metadata content in the flow of a statistical process in a functional way. This last aspect plays a fundamental role in the management of metadata related to data structure and content: these should not be used just for the static description of a data set, but should show how data and the associated metadata evolve and change along the production process steps (see, Bergamasco et al. 2013 for a description of the metadata relationship throughout different data production phases). As a real-life example, the structural metadata organisation introduced in this section has been used to model metadata in SUM and in the Istat corporate dissemination system I.Stat (http:\\dati.istat.it). This approach has been really useful in creating a metadata system with specialised search tools.

One key issue is using the organisation of metadata related to data structure and content depicted in this Section in accordance with the available metadata standards, such as SDMX and DDI. The Appendix shows how SUM relates metadata related to data structure and content and the SDMX artefacts.

5. Process-Related Metadata

In the last years, activity on reference metadata (Androvitsaneas et al. 2006) has been focused on conceptual harmonisation across the European Statistical System, as requested by 2009 Commission Recommendation on reference metadata (European Commission 2009). This recommendation invited NSIs to adopt the set of statistical concepts and subconcepts attached to the recommendation based on the Euro SDMX Metadata Structure, ESMS (Eurostat 2009). More recently, within the framework of quality reporting, an ultimate reporting strategy is being fostered, integrating these concepts with additional ones, as well as with quality indicators such as sampling error, response rate, timeliness and punctuality among others (Götzfried et al. 2011). With respect to process-related metadata, the current standards are thus provided on the one hand by Eurostat reference metadata templates and on the other hand by the GSBPM business process documentation.

The framework for modelling metadata described in Section 3 corresponds to a great extent to Eurostat as well as to the GSBPM standards. The area concerning quality management represents one exception. In GSBPM, this is defined as an overarching process, providing little coverage of quality standards (Eltinge et al. 2013). We believe that integrating the business process model with quality activity and quality indicators is necessary in order to correctly interpret and evaluate standard quality indicators in the light of the relevant metadata, and set improvement goals based on objective information.

As an example of this integration, Istat has developed a system named SIDI-SIQual, which currently collects, stores and disseminates most of the process-related metadata and

quality indicators included in the abovementioned standard structures (Brancato et al. 2004; D'Angiolini et al. 1998). The system is developed according to the model presented in Section 3 and represents a major component of SUM.

Following the approach presented in this article, this system documents quality not only through quality indicators but also in terms of quality control actions, that is, activities aimed at: *i)* preventing errors; *ii)* monitoring process operations and *iii)* evaluating quality. In order to describe all the statistical activities, a hierarchical thesaurus of standard terms, from planning to dissemination, is available (operations thesaurus) whose mapping to the GSBPM items is ensured; for each process operation, a set of possible quality control activities is defined and organised in another hierarchical thesaurus (quality control actions thesaurus). These operations and quality control actions represent the *how* elements of the proposed conceptual model, and will be complemented with the documentation of the support activities in the future when realising the business-related metadata area. For each *how* element, also the *by which means* elements, that is, the software used to perform the activity, as well as the *when* elements, that is, the time reference of the operation/quality control action and various periodicities (periodicity of data collection, processing, dissemination and main releases), are documented in SIDI-SIQual.

The *who* element, that is, who is in charge of each operation, that was initially available within the Istat system was deleted because it was deemed irrelevant. The shift from a stovepipe production model to a more decentralised one might require the reintroduction of this item.

Together with the relevant process-related metadata, standard quality indicators are also stored in the Istat system: coverage, nonresponse, coding, editing and imputation, revision policy, timeliness and punctuality, comparability, coherence and resources (Brancato et al. 2004).

As the repository of the majority of the metadata items and quality indicators required by Eurostat for quality reporting, SIDI-SIQual is currently being enhanced in order to produce SDMX-compliant reference metadata files, (Simeoni 2013) following ESMS and ESQRS structures (Eurostat 2010).

6. Business-Related Metadata

As previously stated, the proposed approach also enables the modelling of business-related metadata, in particular those more strictly linked to the statistical process, in order to provide pieces of information useful for the more efficient management of statistical activities.

Business-related metadata have been defined as metadata arising from administrative and management activities performed to support statistical production and dissemination (i.e., support activities). As pointed out above (see, Figure 1), in GSIM they correspond to the metadata pertaining to the Business group which "is used to capture the designs and plans of Statistical Programs, and the processes undertaken to deliver those programs" (UNECE 2013b) or, in other terms, for high-level management. The GSIM conceptual model can be applied together with GSBPM at different levels of specification. At the most aggregated level, GSBPM corresponds to the statistical process as a whole, while GSIM specifies the corresponding information objects, namely Statistical Need for the input and Assessment for the output.

Business-related metadata are needed particularly by managers to assess efficiency, user satisfaction and acceptance by respondents and funders as well as to balance the needs of statistical information against other needs (Sundgren 2004). Thygesen and Nielsen (2012) further exploit GSBPM, suggesting a model for managing user needs and feedback.

GSIM provides a framework for the conceptualisation of business-related metadata: given their usefulness for the management of an NSI, business-related metadata can be described according to the macro phases *planning*, *execution* and *assessment*, already introduced (see, Figure 2). The proposed approach encompasses the GSIM input and output objects, as well as supports activities connected to statistical production (i.e., Statistical Program in the GSIM terminology) that are only partly represented in GSBPM as overarching processes (e.g., quality and metadata management).

Below, a first identification of business-related metadata for each macro phase is proposed, without claiming to be exhaustive or asserting that all business-related metadata should be managed in a unique repository. The intention is to contribute to fostering a more structured and detailed description of support activities in reference models, such as GSIM and GSBPM, particularly in connection with statistical metadata and quality.

With regard to the *planning* phase, the classes of business-related metadata generated during the different stages relate any statistical activity (data and processes) to the NSI strategic objectives (long- and mid-term goals) and to the different sectorial plans (e.g., methodological, IT, dissemination, etc.), thus setting out resources, schedules and timetables. The final aim is to ensure high quality and an efficient statistical production (regulated by Annual Statistical Planning). Annual Statistical Planning is the repository (in the form of either a document or information system) containing process-related metadata as well as metadata related to data structure and content.

In the *execution* phase, documents containing business-related metadata can be the service charters (for internal and external users), institutional policies, service-level agreements and contracts concerning, for instance, survey operations carried out in outsourcing. They give rise to classes of business-related metadata such as service standards, expected outputs, procedure targets and quality targets, time schedules of subprocesses and associated responsibilities and costs. Links to quality indicators can also be established when administrative procedures and contracts concern the execution of survey operations in outsourcing (e.g., interviewer selection, data collection, data capture, etc.). With regard to the provision of administrative data sources, additional information might concern the input quality of the administrative source.

Finally, in the *assessment* phase done at institutional level, business-related metadata can originate from users sending comments through the website, visiting contact points, calling a toll-free number, in addition to more formalised user satisfaction surveys (or image studies, etc.) as well as from internal staff satisfaction surveys, staff evaluation, and from the assessment of the effectiveness of human resource development initiatives (e.g., staff training courses). These classes of metadata can be useful for complementing process and product quality both from a qualitative and quantitative point of view.

Table 1 provides some examples of business-related metadata (classes of information) for each macro phase, also showing where such information is generated or documented (repositories/information systems). The last column provides a link to quantitative information that originates together with business-related metadata, as well as some

Table 1. Description of business-related metadata

Example of business-related metadata			
	Classes of information	Repositories/information systems	Link to other metadata
PLANNING	Users' needs Human Resources IT Resources Scheduling of activities Costs Critical events	Long-term goals	
		Mid-term goals	
		Multiyear Statistical Planning	
		Annual Statistical Planning	Process-related metadata Metadata related to data structure and content Regulatory requirements
		Sectorial plans (e.g., methodological investments, IT plan, events scheduling, dissemination plan, training plan)	
		Decisions by high-level Committees (e.g., Statistical Council, Quality Committee)	Process-related metadata Metadata related to data structure and content Regulatory requirements
		Documentation from the phases "Specify needs", "Design" and "Build" for a given statistical process	Process-related metadata Metadata related to data structure and content Regulatory requirements
		Risk management	
		Coordination of the National Statistical System	Process-related metadata Metadata related to data structure and content Regulatory requirements

Table 1. Continued

Example of business-related metadata			
	Classes of information	Repositories/information systems	Link to other metadata
EXECUTION	Service standards Expected service outputs Feedbacks from internal and external users Procedures' goals Target schedules Quality targets Costs (for respondents and producers) Survey operations responsibility and execution (e.g., internal/external bodies, people in charge)	Service charters (for internal and external users)	Process-related metadata Metadata related to data structure and content
		Institutional policies (e.g., dissemination policy, revision policy, confidentiality policy)	Process-related metadata
		Service level agreements for survey operations involving external bodies (e.g., provision of administrative data; selection of interviewers by municipalities)	Input quality of administrative data Paradata and quality indicators Process-related metadata Response and administrative burden
		Contracts for activities in outsourcing (e.g., data capture, data collection)	Paradata from monitoring systems

Table 1. Continued

Example of business-related metadata			
	Classes of information	Repositories/information systems	Link to other metadata
ASSESSMENT	Users' satisfaction Users' complaints Stakeholders' requirements Staff needs (e.g., training) Available expertise Lacking expertise Lack of resources Costs (for users) Training information (e.g., attendance, topics, duration)	Users satisfaction surveys	Process and product quality assessment Performance indicators (benchmarking results to planned ones) Users/Usages (usage rate)
		Web accesses to data and metadata	
		Feedbacks from stakeholders, journalists, public at large, etc.	
		Contact points	
		Staff satisfaction surveys	
		Staff evaluation	
		Human resources development initiatives (e.g., training courses)	

reference to process-related metadata and metadata related to data structure and content, which are deemed important.

As stated in Section 5, process-related metadata usually include quantitative measures used for quality analyses and reporting. More specifically, these measures are represented by paradata and quality indicators whose commonly accepted definitions are:

i) *paradata* are data about survey operations, for example, times of day on which interviews were conducted, interview duration, frequency of contacts with each interviewee or attempts to contact the interviewee (Couper 1998; Groves and Heeringa 2006);

ii) *quality indicators* are measures of the quality of statistical products or processes. Some quality indicators are obtained as a by-product of the production process, thus overlapping with paradata. In this case they are also called process variables (Ehling and Körner 2007).

However, support processes and management activities also produce quantitative information, that is, specific paradata (for example, number of available resources by professional expertise, punctuality and timeliness of administrative activities, evaluation score of training courses, compliance with quality targets/service goals, etc.) that can be exploited for a more comprehensive and detailed analysis of statistical activities (see, next section).

Indeed, business-related metadata and their associated paradata can complement information on quality components such as: *i)* accuracy, for instance whenever administrative procedures, service level agreements or contracts establish quality targets; *ii)* timeliness and punctuality, whenever support activities set time schedules that might affect statistical operations; *iii)* relevance and accessibility of data and metadata, by providing information collected through contact points, from the website, and by user satisfaction surveys; *iv)* completeness and coherence of the national statistical system as a whole, through coordination and planning activity.

In addition, business-related metadata provide information on costs for producers, respondents and users. Even though costs are not a quality dimension, they are undoubtedly a constraint for an NSI and need to be assessed to increase efficiency.

7. Complementing Process-Related Metadata With Business-Related Metadata

National Statistical Institutes are currently coping with challenging demands for new information needs and for more accurate and timely data while at the same time facing resource and financial restrictions. Thus it is vital to increase efficiency, reorganise output-oriented production, and rationalise, standardise and industrialise statistical processes. Business-related metadata can support such activities as they are useful for assessment performed at an institutional level (as referred to in Figure 2) if combined with process-related metadata, particularly with quality indicators. A similar assumption can be found in Eltinge et al. (2013), who propose a framework for improving statistical production systems by modelling, assessing and balancing multiple performance criteria including data quality, cost, risk and stakeholder utility. Furthermore, GSIM (UNECE 2013b; UNECE 2013c) introduces the concept of assessment for the Business group, defining it as

"the result of the analysis of the quality and effectiveness of any activity undertaken by a statistical organisation and recommendations on how these can be improved", even though the related metadata are not further specified among the conceptual objects.

Nevertheless, the role of business-related metadata in supporting the decision-making process and quality assessment needs to be more thoroughly investigated.

In fact, business-related metadata and paradata are important not only for efficiently managing statistical processes but also for ensuring process and product quality. For instance, the administrative procedure for selecting an external CATI company will affect the timeliness of survey data if not completed in a given period of time. Since support processes may affect data quality, measurements of such impact (i.e., business-related paradata) are required.

The range and number of potentially useful business-related paradata to be collected and monitored could be extremely wide; this speaks for a stepwise approach, as also proposed by Sundgren (2004), starting from those typologies more linked to quality indicators. Similarly, Thygesen and Nielsen (2012) encourage a more interdisciplinary approach in statistical quality and metadata definition and implementation, focusing on fulfilling user needs.

As possible examples of this, the efforts being made at Istat to try and establish stricter links between quality assessment and management activities can be mentioned. Above all, the improvement actions resulting from quality auditing and self-assessment are currently reported in the Annual Planning, thus allowing for monitoring and reporting on quality improvements, their associated costs, allocated resources and time schedule (Signore et al. 2012).

Following the needs expressed by Istat managers, Annual Planning is being exploited for business-related metadata to complement quality indicators already available in SIDI-SIQual. Information about the staff employed in each business process, as well as information about the costs of carrying out a survey (e.g., printing paper questionnaires and instructions, interviews, data capture) will be taken directly from Annual Planning. In future, the duration of each operation will also be taken from Annual Planning, thus providing a more meaningful understanding of timeliness, and allowing for the office-wide analysis of process operations to screen for bottlenecks.

The exploitation of business repositories and the integration of some business-related metadata in the SUM system will support overall assessment by enabling middle and top management to accomplish a more comprehensive analysis of statistical production processes, widening the range of improvement actions beyond statistical aspects. This activity will also lead to a greater harmonisation of concepts used in different sectors of the Institute (administrative and support processes on the one side and statistical production processes on the other) to satisfy internal users' needs.

8. Final Remarks

This article proposes a unified approach for metadata conceptualisation and management. By jointly modelling metadata related to data structure and content and process- and business-related metadata, it is possible to fully describe data and the underlying statistical production processes. The proposed model also enables the exploitation of business-

related metadata for high-level planning and assessment, even though this is quite a challenging task that requires step-by-step implementation.

This approach also supports harmonisation. As a matter of fact, similar pieces of information are often collected by different sectors of the organisation according to a "silo" approach, resulting in multiple collections of the same items, in information that cannot be reused (by other sectors or in other process steps), or in less accurate information. The adoption of a unique metadata conceptual model facilitates overcoming these problems that many NSIs face by ensuring terminological coherence across different sectors within the organisation and consistency of targets, actions and results.

For this reason, the proposed approach is useful for NSIs in streamlining metadata and quality-related initiatives and in supporting ongoing standardisation and industrialisation processes. The implementation of a unified metadata management system can also lead to increased process efficiency and efficacy as well as to quality enhancement, as already highlighted. This approach is widely applicable in different contexts even though implementation requires adapting it to office-wide infrastructures and technical standards.

The present article also debates some current limitations and suggests possible enhancements of current standards. GSIM could be enhanced with a better representation of the data transformation over the course of the production process (e.g., the use of data content within a data structure). GSBPM could be improved by further detailing the quality management overarching process and proposing generic quality indicators. Both standards could better address business-related metadata and their relationships with other metadata.

In developing SUM, Istat has demonstrated the practical applicability of the conceptual model proposed in the present article, ensuring compliance with standards at the same time.

Of all the lessons learned in implementing SUM, it is worth mentioning the importance of addressing different experts in statistical organisations (e.g., statisticians, quality managers, survey managers, metadata-modelling and IT experts) who often work on specific aspects according to their expertise, thus lacking a view of the issue as a whole. In Istat's experience, a common perspective was found by assigning a prominent role to statistics (i.e., the statistician's point of view) in modelling and using metadata.

The main challenge ahead for SUM is to better conceptualise and integrate business-related metadata into the system, as well as to test their usefulness for institutional global assessment. Based on this experience, we deem it likely that other NSIs implementing unified metadata systems will encounter the same main challenges.

Appendix

The UNECE Frameworks and Standards for Statistical Modernisation group has already investigated the relationship between GSIM concepts with two standards available, SDMX (2009) and DDI (2014), which are usually used in NSIs. This mapping is extremely useful for metadata related to *data structure and content* (Section 4) because their aim is to define a standard to describe data content (DDI) or to allow data and metadata exchange (SDMX). Restricting attention to the "Concept" and "Structure" parts of GSIM used in Section 4, the corresponding concepts are more easily mapped in DDI (with some exceptions, such as classification) than in SDMX. As a matter of fact, SDMX lacks the

typical statistical organisation of metadata in terms of reference population, statistical variables, statistical operations, and so on. Given the mandatory application of SDMX in the European Statistical System, attention was focused on how to map SDMX concepts with those defined in Section 4. As a matter of fact, the use of the statistical role of metadata in conjunction with the SDMX machinery can be a tool to enhance integration within a National Statistical Institute and across different organisations through a unique representation of data structures. According to the SDMX standard, metadata related to data structure and content are mainly categorized in three large groups: *concepts*, that is, any element playing a role in a table description; *code lists*, specifying those concepts with a finite number of instances; *data structure definitions* (DSD), consisting of concepts, their associated (when necessary) code lists and the assignment of concept roles as dimensions, attributes and measures. This is the relationship between these SDMX concepts and the corresponding metadata related to data structure and content (Section 4).

Code list: define specific code lists for populations, numerical variables, statistical operators and Data Content. The Data Content code list can be defined in terms of its components as simple or composite macrodata by considering a number of code annotations: reference population, statistical operator, numerical variable, and so on.

Concept schemes: in order to allow reuse for micro data also, it is advisable to collect the names of the statistical categorical variables and the names of the other concepts used in a DSD (temporal, operational and explanatory concept) in two distinct concept schemes. In a given process, categorical variable names will be reused more than other concepts; among processes, the other concepts will be reused more than categorical variable names (most of the other concepts are available in the Cross Domain concept scheme, although some categorical variable names are also included as the Reference Area). Furthermore, it is appropriate to separate the two groups of concepts because there are operations (as marginalisation of contingency tables) that can be performed only on those dimensions consisting of statistical variables, and not on other kinds of dimension.

DSD: The DSD is organised in the following way. If the primary measure is associated to the concept Observed value (as it usually is in the SDMX context) one (and only one) DSD dimension should be devoted to the Data Content code list. All the categorical variables will be defined as dimensions, as well as time-related concepts (mainly time dimension and frequency). If useful, other concepts can be used as dimensions, as might happen for the time series adjustment concept. In any case, it would be better to organise the noncategorical variable concepts as attributes if possible and appropriate.

One feature that SDMX does not fully cover yet is the metadata relationship over the course of the statistical process. This aspect should be well modelled in order to ensure integration, traceability and metadata phase-based search. As a matter of fact, SDMX already has a module on "Expression and Calculations", although it still lacks an expression language. The issue of an expression language is under discussion in the SDMX Technical Working Group.

9. References

Androvitsaneas, C., B. Sundgren, and L. Thygesen. 2006. "Towards an SDMX User Guide: Exchange of Statistical Data and Metadata Between Different Systems, National

and International." OECD Expert Group on Statistical Data and Metadata Exchange, Geneva, 6–7 April 2006. Available at: https://sites.google.com/site/bosundgren/my-life (accessed November 2013).

Banca d'Italia. 2007. *The Matrix Model. Unified Model for Statistical Data Representation and Processing*. Available at: https://www.bancaditalia.it/statistiche/raccolta-dati/sistema-informativo-statistico/modellazione/matrixmod.pdf (accessed May 2015).

Bergamasco, S., A. Cardacino, F. Rizzo, M. Scanu, and L. Vignola. 2013. "A Strategy on Structural Metadata Management Based on SDMX and the GSIM Models." Work Session on Statistical Metadata (METIS), Geneva, 6–8 May 2013. Available at: http://www.unece.org/stats/documents/2013.05.metis.html#/ (accessed April 2015).

Brancato, G., C. Pellegrini, M. Signore, and G. Simeoni. 2004. "Standardising, Evaluating and Documenting Quality: the Implementation of Istat Information System for Survey Documentation – SIDI." In Proceedings of European Conference on Quality and Methodology in Official Statistics, Q2004, CD-ROM ISBN: 3-8246-0733-6. Mainz, 24–26 May 2004.

Brancato, G. and G. Simeoni. 2012. "Istat Statistical Process Modelling and the Generic Statistical Business Process Model: a Comparison." European Conference on Quality in Official Statistics, Q2012, Athens, 29 May – 1 June, 2012. Available at: http://www.q2012.gr (accessed November 2013).

Couper, M.P. 1998. "Measuring Survey Quality in a CASIC Environment." In Proceedings of the Survey Research Methods Section, 41–49. Available at: http://www.amstat.org/sections/srms/Proceedings/ (accessed March 2014).

D'Angiolini, G., M. Paolucci, and M. Signore. 1998. "Developing Tools for Managing, Exploiting and Disseminating Metainformation: the Istat's Experience." New Techniques and Technologies for Statistics (NTTS) Conference, 4–6 November 1998, Sorrento, vol.1, *Specialised Sessions Papers*, 119–206.

DDI Alliance. 2014. Getting Started with DDI. Available at: http://www.ddialliance.org (accessed October 2014).

Ehling, M. and T. Körner. 2007. *Handbook on Data Quality Assessment Methods and Tools*. Luxembourg: Eurostat.

Eltinge, J.L., P.P. Biemer, and A. Holmberg. 2013. "A Potential Framework for Integrating of Architecture and Methodology to Improve Statistical Production Systems." *Journal of Official Statistics* 29: 125–145. Doi: http://dx.doi.org/10.2478/jos-2013-0007.

European Commission. 2009. *Commission Recommendations of 23 June 2009 on Reference Metadata for the European Statistical System*. Official Journal of the European Union 2009/498/EC. Available at: http://eur-ex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32009H0498 (accessed May 2015).

Eurostat. 2009. *Euro SDMX Metadata Structure (ESMS)*. Available at: http://ec.europa.eu/eurostat/data/metadata/metadata-structure (accessed April 2015).

Eurostat. 2010. *ESS Standard for Quality Reporting Structure*, ESQRS release 1, October 2010. Available at: http://ec.europa.eu/eurostat/data/metadata/metadata_structure (Accessed May 2014).

Eurostat. 2011. *European Statistics Code of Practice*. Available at: http://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice (accessed May 2015).

Fisher, R.A. 1925. *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Gelsema, T. 2012. "The Organisation of Information in a Statistical Office." *Journal of Official Statistics* 28: 413–440.

Götzfried, A., H. Linden, and E. Clement. 2011. "Standards and Processes for Integrating Metadata in the European Statistical System." Workshop on Statistical Metadata (METIS), Geneva, 5–7 October 2011. Available at: http://www.unece.org/stats/documents/2011.10.metis.html (accessed March 2014).

Greenough, C., K. Mechanda, and F. Rizzolo. 2014. "Metadata in the Modernization of Statistical Production at Statistics Canada." European Conference on Quality in Official Statistics, Q2014, Vienna, 2–5 June 2014. Available at: http://www.q2014.at/papers-presentations.html (accessed May 2015).

Groves, R.M. and S.G. Heeringa. 2006. "Responsive Designs for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society, Series A* 169: 439–457. Available at: http://onlinelibrary.wiley.com/doi/10.1111/rssa.2006.169.issue-3/issuetoc (accessed March 2014).

SDMX 2009. *Content Oriented Guidelines*. Available at: http://www.sdmx.org/ (accessed November 2013).

Signore, M., R. Carbini, and M. D'Orazio. 2012. "Quality Assessment in Istat: the Combined Use of Standard Quality Indicator Analysis and Audit Procedures." European Conference on Quality in Official Statistics, Q2012, Athens, 29 May-1 June 2012. Available at: http://www.q2012.gr/articlefiles/sessions/6.2_Signore%20et%20al%20_Quality%20assessment%20at%20Istat%20paper.pdf (accessed November 2013).

Signore, M., M. Scanu, and G. Brancato. 2013a. "Statistical Metadata: a Unified Approach to Management and Dissemination." NTTS – Conferences on New Techniques and Technologies for Statistics, Brussels, 5–7 March 2013. Available at: http://www.cros-portal.eu/sites/default/files//NTTS2013%20Proceedings_0.pdf (accessed January 2015).

Signore, M., M. Scanu, A. De Santis, A. Ambrosetti, and V. Olivieri. 2013b. *La Governance del Sistema Unitario dei Metadati – Metadati Strutturali. Principi e struttura organizzativa per la gestione e armonizzazione dei metadati strutturali dell'Istituto*. Internal document, Istat, Rome Italy.

Simeoni, G. 2013. "Implementing ESS Standards for Reference Metadata and Quality Reporting at ISTAT." Joint UNECE/Eurostat/ OECD Work Session on Statistical Metadata. Geneva, 6–8 May 2013.

Statistics Sweden. 2008. "Classifications of Statistical Metadata." Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS). Luxembourg, 9–11 April 2008. Paper prepared by Bo Sundgren. Available at: http://www.unece.org/stats/documents/2008.04.metis.html# (accessed May 2015).

Sundgren, B. 2004. "Metadata Systems in Statistical Production Processes – for Which Purposes are They Needed, and How Can They Best Be Organized?" Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, 9–11 February 2004. Available at: https://sites.google.com/site/bosundgren/my-life (accessed November 2013).

Thygesen, L. and M.G. Nielsen. 2012. "How to Fulfill User Needs – Metadata, Administrative Data and Processes." European Conference on Quality in Official Statistics, Q2012, Athens, 29 May – 1 June 2012. Available at: http://www.q2012.gr/articlefiles/sessions/21.3_Thygesen-Nielsen_How%20to%20fulfill%20user%20needs%20-%20metadata,%20administrative%20data%20and%20processes%20version%20final.pdf (accessed March 2014).

United Nations Economic Commission for Europe (UNECE). 1995. *Guidelines for the Modeling of Statistical Data and Metadata*. Geneva: United Nations. Available at: http://www1.unece.org/stat/platform/display/metis/UNECE + Guidelines + for + the + Modelling + of + Statistical + Data + and + Metadata (accessed March 2014).

United Nations Economic Commission for Europe (UNECE). 2013a. GSBPM v5.0. Available at: http://www1.unece.org/stat/platform/display/GSBPM/GSBPM + v5.0 (accessed March 2014).

United Nations Economic Commission for Europe (UNECE). 2013b. Generic Statistical Information Model (GSIM): Communication Paper for a General Statistical Audience (Version 1.1, December 2013). Available at: http://www1.unece.org/stat/platform/display/gsim/GSIM + Communication + Paper (accessed April 2015).

United Nations Economic Commission for Europe (UNECE). 2013c. Generic Statistical Information Model (GSIM): Specification (Version 1.1, December 2013). Available at: http://www1.unece.org/stat/platform/display/gsim/GSIM + Specification

United Nations Economic Commission for Europe (UNECE). 2015. Generic Activity Model for Statistical Organizations (GAMSO) v0.2. Available at: http://www1.unece.org/stat/platform/display/GAMSO/GAMSO+Home (accessed April 2015).

Zachman, J.A. 1987. "A Framework for Information Systems Architecture." *IBM Systems Journal* 26, no. 3. Available at: http://www.zachmanframework.com/images/ZI_PIcs/ibmsj2603e.pdf (accessed March 2014).

Received November 2013
Revised February 2015
Accepted February 2015