JOS

Journal of Official Statistics

An International Quarterly Published by SCB, Statistics Sweden www.jos.nu

September 2013

Journal of Official Statistics vol. 31, i. 4 (2015)

Letter to the editor.. p. 537
Jakub Bijak, Isabel Alberts, Juha Alho, John Bryant, Thomas Buettner, Jane Falkingham, Jonathan J. Forster, Patrick Gerland, Thomas King, Luca Onorante, Nico Keilman, Anthony O'Hagan, Darragh Owens, Adrian Raftery, Hana Ševčíková, Peter W.F. Smith

Using auxiliary sample frame information for optimum sampling of rare populations.. p. 545
Martin Barron, Michael Davern, Robert Montgomery, Xian Tao, Kirk M. Wolter, Wei Zeng, Christina Dorell, Carla Black

Response burden in official business surveys: measurement and reduction practices of national statistical institutes.. p. 559
Mojca Bavdaž, Deirdre Giesen, Simona Korenjak Černe, Tora Löfgren, Virginie Raymond-Blaess

Statistical estimators using jointly administrative and survey data to produce french structural business statistics.. p. 589
Philippe Brion, Emmanuel Gros

First impressions of telephone survey interviewers... p. 611
Jessica Broome

Quarterly regional gdp flash estimates by means of benchmarking and chain linking..p. 627
Ángel Cuevas, Enrique M. Quilis, Antoni Espasa

Coordination of conditional poisson samples... p. 649
Anton Grafström, Alina Matei

Cultural variations in the effect of interview privacy and the need for social conformity on reporting sensitive information.. p. 673
Zeina M. Mneimneh, Roger Tourangeau, Beth-Ellen Pennell, Steven G. Heeringa, Michael R. Elliott

Frameworks for guiding the development and improvement of population statistics in the united kingdom... p. 699
James Raymer, Phil Rees, Ann Blake

B-graph sampling to estimate the size of a hidden population................................ p. 723
Marinus Spreen, Stefan Bogaerts

Quality indicators for statistical disclosure methods: a case study on the structure of earnings survey... p. 737
Matthias Templ

Effects of cluster sizes on variance components in two-stage sampling.................. p. 763
Richard Valliant, Jill A. Dever, Frauke Kreute

On proxy variables and categorical data fusion.. p. 783
Li-Chun Zhang

Book review: online panel research: a data quality perspective................................ p. 809
Carina Cornesse, Annelies G. Blom,

Book review: practical tools for designing and weighting survey samples................ p. 813
Mariano Ruiz Espejo

Book review: managing and sharing research data: a guide to good practice........... p. 817
Timothy Michael Mulcahy

Editorial collaborators.. p. 821

Index to volume 31, 2015... p. 827

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 537–544, http://dx.doi.org/10.1515/JOS-2015-0033

Letter to the Editor

Probabilistic Population Forecasts for Informed Decision Making

Demographic forecasts are inherently uncertain. Nevertheless, an appropriate description of this uncertainty is a key underpinning of informed decision making. In recent decades, various methods have been developed to describe the uncertainty of future populations and their structures, but the uptake of such tools amongst the practitioners of official population statistics has been lagging behind. In this letter we revisit the arguments for the practical uses of uncertainty assessments in official population forecasts, and address their implications for decision making. We discuss essential challenges, both for the forecasters and forecast users, and make recommendations for the official statistics community.

Probabilistic Population Forecasts Revisited

Demographic forecasts are concerned with the future population size and structure by sex, age and possibly also some other attributes of interest, such as region of residence, marital status, household type, or other.

As stated by Jan M. Hoem (1973, 9), "the chief purpose of making a population forecast . . . is to contribute to improved planning and better decisions". However, the history of error in population forecasts is as old as the history of these forecasts themselves (Hajnal 1955). Hence, an appropriate description of the forecasting uncertainty is a key aspect of informed decision making. Recognising this, in the early 1970s a small, yet influential group of statistical demographers, becoming increasingly uneasy with the continuing use of deterministic variant 'projections', already suggested that probability distributions should be used to describe the forecast uncertainty (e.g., Keyfitz 1972). At that time, however, it was noted that the available technical resources would not stand up to the task in a general case (Hoem 1973).

The times have changed. Over the past four decades, the methods of statistical demography have been developing very rapidly, especially in the area of stochastic population forecasting at the national level. Increasingly, more arguments and suggestions have been put forward for applying these methods in practice. To mention a few examples: Alho and Spencer (1997) argued that probability distributions would allow the users to

Acknowledgments: This letter stems from the workshop on "The use of probabilistic forecasts with focus on population applications", held in London on June 19, 2014, and organised by the ESRC Centre for Population Change and the EPSRC Care Life Cycle project, University of Southampton, and the Probabilistic Population Projection Group, University of Washington, Seattle, WA. The financial support of the Economic and Social Research Council (ES/K007394/1) and the Engineering and Physical Sciences Research Council (EP/H021698/1) is gratefully acknowledged. We also thank Giampaolo Lanzieri, Richard Pereira, Philip Rees and other workshop participants for stimulating discussions, as well as an anonymous reviewer for helpful suggestions. The views expressed in this letter are exclusively those of the authors, and should not be attributed to any institution with which they are affiliated.

prepare appropriate contingency plans. Tuljapurkar (1992), de Beer (2000) and Bijak (2010) have recommended taking advantage of decision theory, allowing for different – possibly asymmetric – objective or loss functions of the forecast users. Lee (1998) added the possibility of making derived forecasts, where population predictions could be integrated with economic ones, as well as the analysis of conditional forecasts, with some sources of uncertainty removed.

Despite these methodological developments and recommendations, probabilistic population forecasting methods have been incorporated into official statistical practice only in a handful of countries – chiefly in the Netherlands and New Zealand. Ambitious plans laid out at the US Census Bureau a decade ago (Long and Hollmann 2004) have since been mothballed. Progress was additionally hampered by the lack of established methodology for forecasting subnational populations or disaggregating the forecasts by various groupings of interest (household position, labour market status, etc.). To our knowledge, there have been hardly any policy applications of formal decision analysis or similar techniques, with the notable exception of Alho et al. (2008).

However, a major step forward was taken on July 11, 2014 (World Population Day), when the UN Population Division for the first time issued official probabilistic population projections for all countries, using the methodology of Raftery et al. (2012). These were the basis for the article of Gerland et al. (2014), which argued that the world population is unlikely to stop growing this century – a probabilistic statement. This attracted considerable media coverage, much of which showed an understanding of the probabilities reported (e.g., Carrington 2014; Schiermeier 2014). We expect this to spur a revival of interest in official probabilistic forecasting of populations. Anticipating this revival, we want to reopen the discussion on the potential advantages and obstacles of producing and using the probabilistic population forecasts.

Challenges and Open Questions

Current practice in official population forecasting is not sufficient. Deterministic forecasts based on single numbers are bound to fail, and to surprise their end users time and again. Probabilistic forecasts, with probability distributions describing possible outcomes, can prepare the user for such outcomes. However, a very important aspect of the single-number forecasts is that they are easy to grasp in cognitive terms. Hence, to aid decisions, probability distributions need to be summarised in an appropriate way that will be useful for the users and correspond with their requirements.

Our basic premises are as follows. First, there is a need for an analytical framework for supporting policy and planning decisions under uncertainty, especially where there are some real concerns which can be expressed as losses – economic losses, or other, such as reputational. Second, deterministic scenarios can be misleading, have a zero probability under any continuous probability measure (or very close to zero in other cases), and are problematic to aggregate or compare with each other. They also attempt to answer a tautological question – what *would* happen under certain assumptions – when the real policy-relevant question is: what *will* happen (Keyfitz 1972; Hand 1994). Of course, a precise answer to this question is impossible, and probabilistic forecasts – similarly to deterministic scenarios – also depend on a number of assumptions, but they explicitly state the forecaster's belief as to how probable those conditions are. Third, probabilistic forecasts not only attend to the relevant question about the future, but also contain precise warnings about the uncertainty. We consider this to be an ethical virtue.

Various reasons have been put forward for a meagre uptake of probabilistic methods in official uses. Lutz and Goldstein (2004, 3–4) cite four arguments: a "misleading sense of precision" regarding probability ranges; the "mechanistic" nature of many forecasts, chiefly based on time series; technical and conceptual complexities and difficulties involved in making such forecasts; and a lack of skilled workforce at the statistical offices. Ten years later, however, while the official statistical agencies may still face technical, statistical, and computational challenges related to probabilistic forecasting, the goalposts have been moved. In our view, the above reservations can now be largely addressed, thanks to advances in methodology and statistical training, and the key contemporary challenges can be found elsewhere. Four of them are discussed in more detail below.

The first challenge is the user **attitude** towards forecasting uncertainty and towards risk in general. Uncertainty can be either perceived as a "curse" – lack of knowledge about the future; or as a "blessing" – if dealt with properly, this is additional information that can help us make better decisions. In particular, there is still a lack of clarity surrounding what can be gained – or lost – by using probabilistic forecasts in practice. Besides, the way uncertainty is dealt with also depends on the risk attitude of the users (Kahneman 2011), with options ranging from downplaying uncertainty for the sake of efficiency or potential gains, to preparing for the 'worst-case' scenarios under high risk aversion. As Kahneman (2011, 263) has put it, "an unbiased appreciation of uncertainty is a cornerstone of rationality, but it is not what people and organizations want."

The second challenge results from the **specificity** of various user needs and circumstances. The horizons for forecasts, projections, and decisions differ; so do the potential consequences of these decisions, as well as the level of risk aversion of the decision makers. The choice between a few predefined variants is not sufficient, as they are unlikely to correspond to user needs, especially if only offered at national level. On the other hand, offering decision support via probabilistic forecasts requires striking a delicate balance between what is needed by the users and what can be realistically offered by the forecasters. Examples range from local investment decisions, in the case of subnational forecasts (NZIER 2014), to macroeconomic policy issues, such as the sustainability of pension and other social security systems (Alho et al. 2008). Such decisions usually have long term and potentially very costly consequences, so it is all the more important to base them on a comprehensive analysis of potential forecast errors.

The third challenge is how to deal with **information** – specifically, statistical data and inferences made on their basis – which may be either incomplete or superfluous, and possibly conflicting. Here, the role of prior beliefs and expert judgement comes to the fore, and an appropriate approach to elicitation becomes crucial (O'Hagan et al. 2006). The same applies to eliciting from the users their attitudes to risk and loss or utility functions, which approximate the decision setting – the relative losses of underpredicting or overpredicting the parameters of interest (see Bijak 2010). The key questions are: what are the practical implications of probabilistic forecasts, and, if the forecasts are wrong, what is at stake? Elicitation requires caution, especially as the perceptions of concepts such as probability, utility, or loss are not uniform. Besides, cognitive biases have to be considered here – especially overconfidence and illusion of certainty, which are a subconscious way of avoiding the cognitive effort of processing more information than just single-point predictions or guesses (Kahneman 2011; Raftery 2014).

Finally, the fourth challenge is related to **validation**, the calibration and testing of probabilistic forecasts, chiefly through comparing them with known outcomes (Alho and Spencer 1997). Even though this aspect is more technical, it is a crucial complement for some other challenges, in particular attitudes: to appreciate the role of uncertainty, the users need to trust that it is calculated correctly. Here, the main question concerns the aim of probabilistic forecasting: is it to *describe* the predictive uncertainty, or to *minimise* it, which can be misleading? Alternatively, as suggested by Gneiting et al. (2007), a compromise could be to minimise uncertainty for a well-calibrated model, where the expected (*ex ante*) and observed (*ex post*) empirical frequencies of events match each other. In such models, events with predicted 50% probability would happen half of the time on average, the events with 90% probability would occur nine out of ten times, and so on.

Where Next? Practical Recommendations

To address the challenges mentioned above, the starting point could be to change the discourse about uncertainty from just a lack of knowledge, to a more realistic and nuanced view. In that regard, the discussion about uncertainty could be reframed as being about confidence, or *additional* knowledge or information. Besides, being explicit and transparent about the forecasting uncertainty can be also associated with such virtues as honesty, humility, and trust.

This approach has already proved successful in the aviation industry, contributing to a substantial increase in safety levels in the recent decades. One of the underpinning cultural changes that the aviation community has witnessed was a shift from a reactive and punitive blame-for-error model to a "just culture". This concept can be defined as "a culture in which front line operators and others are not punished for actions, omissions or decisions taken by them that are commensurate with their experience and training, but where gross negligence, wilful violations and destructive acts are not tolerated" (EUROCONTROL 2014), and explicitly recognises the role of uncertainty as an inherent part of operations. Importantly, by allowing an honest discussion about errors, this model allows for learning from the mistakes, and helps prevent them in the future.

In order to convince the users and producers of population forecasts of the **added value** of an analysis of uncertainty, and to overcome some institutional inertia, the experience of other areas and disciplines could be looked at. Probabilistic forecasting has been successfully developed, for example, in some aspects of meteorology and climatology, aviation, and macroprudential economic regulation. In these areas, techniques of communicating uncertainty to the users and the general public are also being researched. This experience and expertise could be used in population forecasting. Similarly, population forecasts are a crucial input for many policy areas, for example with respect to such structural measures as pension reforms. Given that population is often used as an exogenous variable in the macroeconomic system, its forecasts will be helpful in supporting decisions regarding the endogenous policy variables, such as interest rates.

In particular, the meteorological community has been grappling with issues surrounding uncertainty in weather forecasts for over a century (WMO 2008). Unlike in the case of the aviation industry, with its high level of regulation and entry barriers, the users of weather forecasts are much more diverse. The recent *Guidelines on Communicating Forecasts Uncertainty* (WMO 2008) offer several arguments for communicating uncertainty to the users. Besides the clear applicability for decision making, increasing users' confidence that the forecasts are a result of an honest, objective, and scientific endeavour, and besides managing the users' expectations, it is also pointed out that uncertain weather forecasts simply reflect the state of the science (WMO 2008). This point is even more important in demography and other social domains, where, thanks to human agency and ingenuity, we do not know (and will be never able to know exactly) what drives the individual decisions on, for example, whether and when to have children or to migrate, or the reasons why some people die earlier than others, or why the different demographic processes change over time. In that sense, probabilistic forecasts provide an important epistemological statement about the limited state of knowledge in population sciences – and about the limits of forecasting more generally.

Addressing the second challenge requires **bespoke approaches**, with forecasts tailored to the specific needs of different types of users and different audiences (Raftery 2014). There are vast differences between high-level, longer-term, strategic decision making, and practical, more immediate, operational-level planning, which requires quantitative input for decisions (Bijak 2010). In that respect, full probabilistic forecasts offer a general solution, from which the specific options can be derived. Some users (and uses) may require no point forecasts or estimates at all. And if scenarios are needed, they can be obtained from trajectories based on quantiles from predictive distributions. Finally, conditional probabilistic forecasts, assuming that some variables are known, can help answer policy-relevant "what-if" questions. Interactive, versatile online tools might help the users here. In any case, the user appreciation of the benefits of probabilistic forecasts can help the official statistical agencies justify the resources needed for their development.

Tailoring the predictions, and eliciting the relevant information, such as prior beliefs, expert judgement, or loss functions, requires **interaction** with users. The prerequisites here involve an open, two-way dialogue, with frequent exchange of information between forecasters and users. This exchange can become routine if the forecasts are periodically updated, as is often the case with official population forecasts. Some of the related challenges can be overcome by appropriate methods of communication, such as the use of visualisations (Spiegelhalter et al. 2011). This aspect would benefit from wider insights from cognitive science on such issues as statistical literacy, education, and training, not only related to the end users of forecasts, but also the general public (see also Kahneman 2011). Similarly to the case of weather forecasting, this is especially important for nonspecialist users, who may benefit particularly from appropriate visualisations, interactive online tools, and similar materials.

Not surprisingly, more **methodological research** on a number of technical issues is required. In particular, there is need to design an appropriate framework for calibrating whole time series of observations. Besides, for rare events, there may not be enough observations to properly calibrate the extremes (tails) of the distributions (see e.g., Taleb 2007). In such cases, exploration of methods and techniques of risk management can be promising, whereby future events are classified according to a combination of their probability and impact. As mentioned above, there is also a need to develop a wider range of methods for the types of forecasts that play the greatest role in actual policy and expenditure decisions, for example at the subnational level.

However, in order to achieve a paradigm shift in practical applications of probabilistic population forecasts, the focus should not be on methods, but rather on possible impacts and consequences of decisions. In such a way, the ongoing change of methodological perspective in demographic forecasting, from deterministic point forecasts through variant scenarios to probabilistic predictions, would continue incrementally towards interactive decision support at a variety of levels of policymaking – from national to subnational, in parallel with the methodological developments for the latter. Of course, as a prerequisite, various sources of uncertainty need to be acknowledged and combined in the forecasts, ideally within a joint and coherent framework, such as the one offered by Bayesian statistics.

The challenges of the practical uses of probabilistic forecasts are important, but they are now well recognised and are not insurmountable. The methodology is ripe, and insights from other areas of application are encouraging. In many other areas, the concepts of uncertainty and risk have already entered the language and practice of the decision makers and other forecast users. As for population forecasts, several pioneer countries, as well as the United Nations Population Division, have also taken up to the challenge. We hope this trend continues – where there's a will, there's a way.

References

Alho, J.M. and B.D. Spencer. 1997. "The Practical Specification of the Expected Error of Population Forecasts." *Journal of Official Statistics* 13: 203–226.

Alho, J.M., S.E. Hougaard Jensen, and J. Lassila. 2008. *Uncertain Demographics and Fiscal Sustainability*. Cambridge: Cambridge University Press.

Bijak, J. 2010. *Forecasting International Migration in Europe: A Bayesian View*. Dordrecht: Springer.

Carrington, D. 2014. "World Population to hit 11bn in 2100 - With 70% Chance of Continuous Rise." *The Guardian*, September 18, 2014. Available at: http://www.theguardian.com/environment/2014/sep/18/world-population-new-study-11bn-2100 (accessed on 20 September 2014).

De Beer, J. 2000. *Dealing with Uncertainty in Population Forecasting*. Voorburg: Statistics Netherlands.

EUROCONTROL. 2014. "Just Culture." Online material. Brussels: European Organisation for the Safety of Air Navigation. Available at: https://www.eurocontrol.int/articles/just-culture (accessed on 10 April 2014).

Gerland, P., A.E. Raftery, H. Ševčíková, N. Li, D. Gu, T. Spoorenberg, L. Alkema, B.K. Fosdick, J. Chunn, N. Lalic, G. Bay, T. Buettner, G.K. Heilig, and J.R. Wilmoth. 2014.

"World Population Stabilization Unlikely This Century." *Science* 346: 234–237. Doi: http://dx.doi.org/10.1126/science.1257469 (accessed on 20 September 2014).

Gneiting, T., F. Balabdaoui, and A.E. Raftery. 2007. "Probabilistic Forecasts, Calibration and Sharpness." *Journal of the Royal Statistical Society, Series B* 69: 243–268. Doi: http://dx.doi.org/10.1111/j.1467-9868.2007.00587.x.

Hajnal, J. 1955. "The Prospects for Population Forecasts." *Journal of the American Statistical Association* 50: 309–322. Doi: http://dx.doi.org/10.1080/01621459.1955.10501267.

Hand, D.J. 1994. "Deconstructing Statistical Questions." *Journal of the Royal Statistical Society, Series A* 157: 317–356. Doi: http://dx.doi.org/10.2307/2983526.

Hoem, J.M. 1973. *Levels of Error in Population Forecasts*. Article No. 61. Oslo: Statistisk Sentralbyrå.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Keyfitz, N. 1972. "On Future Population." *Journal of the American Statistical Association* 67: 347–363. Doi: http://dx.doi.org/10.1080/01621459.1972.10482386.

Lee, R.D. 1998. "Probabilistic Approaches to Population Forecasting." *Population and Development Review* 24: 156–190. Doi: http://dx.doi.org/10.2307/2808055.

Long, J.F. and F.W. Hollmann. 2004. "Developing Official Stochastic Population Forecasts at the US Census Bureau." *International Statistical Review* 72: 201–208. Doi: http://dx.doi.org/10.1111/j.1751-5823.2004.tb00233.x.

Lutz, W. and J.R. Goldstein. 2004. "Introduction: How to Deal with Uncertainty in Population Forecasting?" *International Statistical Review* 72: 1–4.

NZIER. 2014. "Costly Investment Decisions Require Improved Population Forecasts." *NZIER Insight* 47. Wellington: New Zealand Institute of Economic Research.

O'Hagan, A., C.E. Buck, A. Daneshkhah, J.E. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley, and T. Rakow. 2006. *Uncertain Judgements: Eliciting Expert Probabilities*. Chichester: Wiley.

Raftery, A.E. 2014. Use and Communication of Probabilistic Forecasts. Mimeo; University of Washington. Available at: http://arxiv.org/abs/1408.4812 (accessed on 22 August 2014).

Raftery, A.E., N. Li, H. Ševčíková, P. Gerland, and G.K. Heilig. 2012. "Bayesian Probabilistic Population Projections for All Countries." *Proceedings of the National Academy of Sciences* 109: 13915–13921.

Schiermeier, Q. 2014. "World Population Unlikely to Stop Growing This Century." *Nature News*. September 18, 2014. Available at: http://www.nature.com/news/world-population-unlikely-to-stop-growing-this-century-1.15956 (accessed on 20 September 2014).

Spiegelhalter, D., M. Pearson, and I. Short. 2011. "Visualising Uncertainty about the Future." *Science* 333: 1393–1400. Doi: http://dx.doi.org/10.1126/science.1191181.

Taleb, N.N. 2007. *The Black Swan. The Impact of the Highly Improbable*. New York: Random House.

Tuljapurkar, S. 1992. "Stochastic Population Forecasts and Their Uses." *International Journal of Forecasting* 8: 385–391.

WMO. 2008. "Guidelines on Communicating Forecast Uncertainty." Technical document WMO/TD No. 1422. Geneva: World Meteorological Organization. Available at:

https://www.wmo.int/pages/prog/amp/pwsp/documents/GuidelinesonCommunicating-Uncertainty_TD-4122.pdf (accessed on 10 April 2015).

Jakub Bijak
(corresponding author)
University of Southampton
Department of Social Statistics and Demography
Southampton SO17 1BJ, UK
Email: j.bijak@soton.ac.uk

Isabel Alberts
German Weather Service

Juha Alho
University of Helsinki

John Bryant
Statistics New Zealand

Thomas Buettner
Formerly the UN Population Division

Jane Falkingham
University of Southampton

Jonathan J. Forster
University of Southampton

Patrick Gerland
UN Population Division

Thomas King
University of Newcastle

Luca Onorante
Central Bank of Ireland

Nico Keilman
University of Oslo

Anthony O'Hagan
University of Sheffield

Darragh Owens
Aviation Training Consultant

Adrian Raftery
University of Washington

Hana Ševčíková
University of Washington

Peter W.F. Smith
University of Southampton

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 545–557, http://dx.doi.org/10.1515/JOS-2015-0034

Using Auxiliary Sample Frame Information for Optimum Sampling of Rare Populations

Martin Barron[1], *Michael Davern*[1], *Robert Montgomery*[1], *Xian Tao*[1], *Kirk M. Wolter*[1], *Wei Zeng*[1], *Christina Dorell*[2], *and Carla Black*[2]

We investigate disproportionate stratified sampling as a possibly efficient method of surveying members of a rare domain in circumstances in which there is no acceptable list of members. In this work, we assume that information is available at the sampling stage to stratify the general-population sampling frame into high- and low-density strata. Under a fixed constraint on the variance of the estimator of the domain mean, we make the optimum allocation of sample size to the several strata and show that, in comparison to proportional allocation, the optimum allocation requires (a) a smaller total sample size but (b) a larger number of interviews of members of the rare domain. We illustrate the methods using information about American consumers maintained by market-research companies. Such companies are able to develop lists of households that are thought to have a defined attribute of interest, such as at least one resident in a user-specified age range. The lists are imperfect, with false positives and negatives. We apply an age-targeted list to the National Immunization Survey (NIS), conducted by the Centers for Disease Control and Prevention, which targets the relatively rare population of children age 19–35 months. The age-targeted list comprises the high-density stratum and the rest of the survey's sampling frame comprises the low-density stratum. Given the optimum allocation, we demonstrate potential cost savings for the NIS in excess of ten percent.

Key words: Optimum allocation; cost model; variance; disproportionate stratification; rare population; age-targeted list; telephone surveys; National Immunization Survey.

1. Introduction

Surveys of rare populations are common in a variety of scientific fields. For example, health surveys often target low-prevalence domains, such as people with a specific disease, a specific chronic condition, a special healthcare need, or people who have received specific healthcare services. While in general there is no universally accepted demarcation between rare and nonrare, we have in mind possible rare domains that comprise less than ten percent of the general population.

[1] NORC – University of Chicago, 55 East Monroe Street Suite 3000, IL 60603-5805, Chicago, Illinois, U.S.A. Emails: martin-barron@norc.org, davern-michael@norc.org, montgomery-robert@norc.org, tao-xian @norc.org, wolter-kirk@norc.org, and zeng-wei@norc.org.
[2] National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, 1600 Clifton Rd, Atlanta, GA 30333, U.S.A. Emails: eqw1@cdc.gov and cblack2@cdc.gov.

Acknowledgments: The findings and conclusions in this article are those of the author(s) and do not necessarily represent the views of Centers for Disease Control and Prevention. Kirk Wolter presented this article in October 2012 at the International Conference on Survey Methods for Hard to Reach Populations in New Orleans, Louisiana. The authors thank the associate editor and referees for helpful comments on earlier drafts of the article.

We consider the problem of sampling when two circumstances are true: (1) no acceptable sampling frame exists for the rare domain of interest, henceforth denoted by D, and (2) an acceptable sampling frame does exist for the general population and auxiliary information is available at the time of sampling that enables the survey statistician to partition this frame into high- and low-density strata. The former are presumed to have higher prevalence rates (also called the *eligibility rate*) of the rare population than the latter. A sample is selected from each stratum; a brief screening interview is administered to persons in the sample to ascertain membership in D; and then members receive the main survey interview and nonmembers are not interviewed. Practical applications of this problem may encounter a range of eligibility rates in the various strata. Throughout this article, we use the labels *high density* and *low density* simply to indicate that one or more strata have higher eligibility rates, perhaps much higher, than the other strata, not to imply any absolute level of eligibility.

One example of this sampling problem occurs when a list (possibly quite imperfect, reflecting false positives and false negatives) of members of D exists and is available at the time of sampling. The list itself may be considered the high-density stratum and all persons represented on the general sampling frame and not on the list may be considered the low-density stratum. A second example occurs when the sampling frame is stratified by census variables that are thought to be associated with membership in D. Such examples may become increasingly important in the future as cost pressures on surveys mount.

Our main aim in this article is to develop a method of *disproportionate stratification* in which the high-density strata are sampled at higher rates than the low-density strata. We examine whether the use of different sampling rates can result in lower data-collection costs than when the same sampling rate is used across the entire sampling frame. Aspects of this sampling problem have been treated previously by Sudman (1972), Waksberg (1973), and Kalton and Anderson (1986). Kalton (2009) arrived at the general conclusion that disproportionate stratification can reduce cost only when three conditions are true: (a) the prevalence rates in the high-density strata are much higher than those in the low-density strata, (b) the high-density strata contain a substantial portion of the overall rare domain D, and (c) the per-unit cost of the main data collection must be high relative to the cost of screening. Valliant et al. (2014) study the use of stratification of address-based samples of households in which the strata are defined by auxiliary information from commercial sources.

Our specific aims are to give a precise definition of the method of disproportionate stratification and demonstrate the optimum design and its sample sizes within this class (Section 2), to describe certain information available from market-research companies that can be used for implementation of such stratification (Section 3), and to illustrate the optimum design and select market-research information using an age-targeted list applied to the National Immunization Survey (NIS), a project conducted on an ongoing basis by the Centers for Disease Control and Prevention to measure the vaccination status of young children (Section 4). The article closes with a brief summary and recommendations (Section 5).

2. Methods for an Optimum Allocation

Two notions of optimality are standard in survey sampling: first, one can fix the variance of a key survey statistic of interest and design the sample to minimize the cost of data collection, or second, one can fix the cost of data collection and design the sample to minimize the variance of the key statistic. Both notions of optimality lead to a similar relative allocation of the sample size across the several strata (Cochran 1977). We will focus on the first notion of optimality, and comment briefly on the second notion at the end of this section.

We consider a sampling design in which there are L strata indexed by h, and, without loss of generality, take the eligibility rate of the rare domain D to be decreasing from $h = 1$ to $h = L$. Simple random samples are taken from each of the strata, resulting in the selection of some members of the rare domain and some nonmembers.

A brief screening interview is conducted to determine the members of D, followed by the main interview of such members. In this section, we consider the ideal circumstance of complete response, while in Section 4 we give an illustration in which nonresponse does occur. Furthermore, throughout the article, we assume that domain membership can be ascertained without error in the screening interview. This setting is in contrast with some survey applications in which reporting, coding, or definitional problems can result in erroneous classifications of sampling units as in D or not in D.

We let c_{scr} denote the cost (or hours) per screening interview and c_{inv} the cost (or hours) per main interview. We let n_h be the number of completed screening interviews and m_h the number of completed main interviews in stratum h. Moreover, we let $r_h = N_{Dh}/N_h$ denote the population eligibility rate (size of the rare domain D as a proportion of the size of the sampling frame) in stratum h and $r = \sum_{h=1}^{1} W_h r_h = N_D/N$ the overall eligibility rate across the entire sampling frame, where $W_h = N_h/N$ is the proportion of units on the sampling frame that are classified in stratum h.

Total expected survey costs can be expressed by

$$T = \sum_{h=1}^{1} \left(c_{scr} n_h + c_{inv} E\{m_h\} \right) = \sum_{h=1}^{1} t_h n_h \, , \tag{1}$$

where $t_h = c_{scr} + c_{inv} r_h$ is the average combined cost per unit in the sample. On average, each unit in the sample incurs its own cost of screening plus a fractional share of the cost of the main interview, where the fraction is the eligibility rate. For simplicity, we have omitted fixed costs from the model, because they have no bearing on the optimum allocation. Also for simplicity, we have assumed that the per-unit costs are identical in the two strata. The methods extend directly to the case where the cost components vary by stratum, such as when response rates vary by stratum.

We assume the main aim of the survey is to estimate the mean of the rare domain, say $R = Y/X$, where Y_{hi} is the variable of interest for members (h, i) of the rare domain and is zero for nonmembers, X_{hi} is 1.0 for members of the rare domain and is zero for nonmembers, and Y and X are the population totals of these variables. We let $\hat{R} = \hat{Y}/\hat{X}$ be the standard ratio estimator of R, where $\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} d_{hi} y_{hi}$ is the estimated domain total of the variable of interest, $\hat{X} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} d_{hi} x_{hi}$ is the estimated total number of members of the rare domain, and $d_{hi} = N_h/n_h$ is the design weight for all $i = 1, \ldots, n_h$ and $h = 1, \ldots, L$.

Assuming that finite population correction terms can be ignored and that the means and variance components are of similar value in the various strata, the Taylor series approximation to the variance of the estimator is given approximately by

$$Var\{\hat{R}\} \doteq \frac{S^2}{r^2}\sum_{h=1}^{L}\frac{W_h^2 r_h}{n_h}\,, \tag{2}$$

where S^2 is the variance component among members of the rare domain. Kalton and Anderson (1986) give a similar expression for this variance. An alternative exact expression for the variance can be given in lieu of (2) in the event that the variance components differ from stratum to stratum.

Given the foregoing, the classical optimum allocation (Cochran 1977) of the sample to the two strata, which minimizes cost subject to a constraint on the variance, is given by

$$n_h^o = a_h n^o\,, \tag{3}$$

where

$$a_h = \frac{W_h\sqrt{r_h}/\sqrt{t_h}}{\sum_{h'=1}^{L} W_{h'}\sqrt{r_{h'}}/\sqrt{t_{h'}}}\,, \tag{4}$$

$$n^o = \frac{S^2}{V^o r^2}\sum_{h=1}^{L} W_h\sqrt{r_h}\sqrt{t_h}\sum_{h=1}^{1} W_h\sqrt{r_h}/\sqrt{t_h}\,, \tag{5}$$

and V^o is the specified fixed constraint on the variance. The sample size within a stratum is proportional to the size of the stratum and to the root of the eligibility rate in the stratum, and inversely proportional to the root of the per-unit cost of data collection in the stratum. The expected number of interviews of members of the rare domain is

$$m^o = \sum_{h=1}^{1} n_h^o r_h = \frac{S^2}{V^o r^2}\left(\sum_{h=1}^{1} W_h r_h \frac{\sqrt{t_h}}{\sqrt{r_h}}\right)\left(\sum_{h=1}^{L} W_h r_h \frac{\sqrt{r_h}}{\sqrt{t_h}}\right), \tag{6a}$$

and the minimum total cost under the optimum allocation is

$$T^o = \frac{S^2}{V^o r^2}\left(\sum_{h=1}^{1} W_h\sqrt{r_h}\sqrt{t_h}\right)^2. \tag{6b}$$

An alternative sampling design that is used in many surveys involves the selection of the sample without regard to the high- and low-density strata, or effectively the selection of the sample from the sampling strata using proportional allocation. The sample sizes required to achieve the variance constraint are

$$n_h^p = W_h n^p \tag{7}$$

$$n^p = \frac{S^2}{V^o r^2}\sum_{h=1}^{1} W_h r_h\,, \tag{8}$$

the expected number of interviews of members of the rare domain is

$$m^p = \frac{S^2}{V^o r^2} \left(\sum_{h=1}^{1} W_h r_h \right)^2 , \tag{9a}$$

and the total cost given this allocation is

$$T^p = \frac{S^2}{V^o r^2} \sum_{h=1}^{1} W_h r_h \sum_{h=1}^{1} W_h t_h \, . \tag{9b}$$

A measure of the cost savings associated with the optimum allocation is the ratio of total costs T^o/T^p, where the superscripts "o" and "p" signify optimum and proportional allocation, respectively. This ratio is guaranteed to be less than or equal to 1 by construction. If the eligibility rates are homogeneous, that is, $r_h = r$, for all h, then the ratio is equal to 1. Cost can be reduced relative to proportional allocation when the eligibility rates are variable and there are high-density strata of non-negligible size.

In comparing optimum and proportional allocations when variance is fixed, two inequalities are true: (i) $n^o/n^p \le 1$ and (ii) $m^o/m^p \ge 1$. Because $T^o/T^p \le 1$, the ratio of sample sizes is $n^o/n^p \le c_{scr} + c_{inv} r^p / c_{scr} + c_{inv} r^o$, where $r^p = \sum_{h=1}^{1} W_h r_h = r$ and $r^o = \sum_{h=1}^{1} a_h r_h$. Inequality (i) follows from the fact that $r^o \ge r^p$. Applying the Cauchy-Schwarz inequality to (6) and (9) gives inequality (ii).

Summarizing the results for fixed variance, the optimum allocation results in cost savings relative to proportional allocation; it requires a smaller total sample size but a larger number of interviews of members of the rare domain than does proportional allocation. The optimum allocation involves disproportionate sampling, it creates a weighting effect, and it therefore requires more interviews to achieve the fixed variance.

Briefly, for fixed cost, the variance-minimizing optimum allocation is given by (3) and (4), where $n^o = T^o \left(\sum_{h=1}^{1} W_h \sqrt{r_h} / \sqrt{t_h} \right) / \left(\sum_{h=1}^{1} W_h \sqrt{r_h} \sqrt{t_h} \right)$. Consider the special case $c_{scr} = 0$, $c_{inv} = 1$, and $T^o = \sum_{h=1}^{1} n_h r_h$, which corresponds to fixing the expected sample size in the rare domain D. For this case, the optimum allocation is proportional allocation with $n^o = T^o/r$.

3. Market-Research Lists for Stratification

Market-research companies have developed proprietary databases containing demographic, behavioral, and consumer information on people and households throughout the world. These data can be used as the basis for the stratification used in Section 2. Even though the specific details of their construction are proprietary, it is known that the databases are compiled from product registrations, store loyalty programs, credit-card purchases, cable-television viewing, internet searching, smartphone applications, coupon redemptions, mobile health devices, voter registration databases, publicly available real-estate transactions, as well as many other sources. And while the data from market-research companies are not always accurate at the individual case level (Pasek et al. 2014), they may still be useful for stratifying a survey sampling frame of the general population into high- and low-density strata for households or people who have the rare characteristic of interest. Using the lists provided by market-research companies containing names, telephone numbers or addresses (depending on the sampling frame used), the sampling statistician can divide the sampling frame into two or more strata based on whether the market-research company has associated the name, telephone number or address with a specific rare trait or characteristic of interest (domain D).

The general approach of stratifying the sampling frame into high- and low-density strata is not limited to lists provided by market-research companies. For example, if a team of researchers was interested in studying asthma among children using an address-based sample frame, they might be able to obtain a high-density list of addresses from administrative data of children on Medicaid (Medicaid is a government health-insurance program for needy people in the U.S.) with asthma-related prescriptions. The low-density stratum would be comprised of all remaining addresses. And there could be combinations with one high-density list coming from a state Medicaid agency of addresses of child beneficiaries with asthma-related prescriptions, a second list coming from a market-research company that identifies households likely to have children, and a third low-density frame of all remaining addresses not on either of the two high-density lists. Other applications of this method could entail using voter registration lists as the high-density frame for an address-based sample of likely voters for a local election, and the low-density frame could be all the remaining addresses. Market-research companies and administrative data sources offer ample opportunities to take advantage of this kind of methodology, as many lists are available to stratify the sampling frames into high-density and low-density strata that presumably have differing eligibility rates for members of the rare domain D. Lists used for stratification could target information on age, race, ethnicity, people who purchased and registered specific products (e.g., insulin pumps or asthma prescriptions), disease registries, voter registration lists, and lists of households who redeem specific coupons.

The methods presented in the foregoing section for sampling and interviewing members of a rare domain therefore have application to at least two related problems:

1. A comprehensive sampling frame exists, which contains information that permits the population to be partitioned into two or more sampling strata that vary in their density of the rare domain, D.
2. There are initially two (or more) sampling frames: one containing a complete list of the overall population, and one (or more) containing only a subset of the first list that is rich in members of the rare domain, D. By matching the second list(s) to the first, a revised sampling frame can be obtained that identifies two or more sampling strata: cases on the second list (the high-density stratum) and cases not on the second list (the low-density stratum).

The lists used to stratify the sampling frame (e.g., an age-targeted list from a market-research company or Medicaid enrollment data on likely asthma patients) are subject to error, including the telephone numbers or addresses of households that do not actually have the rare attribute (false positives), and excluding the telephone numbers or addresses of households that do have the attribute (false negatives). Due to their origin in the market-research field, some lists may be skewed towards more affluent households that have landline telephone numbers, register automobiles, and buy things on credit. As long as the entire population of D is covered by at least one of the lists or sampling strata, there is no bias in estimators of population parameters of interest.

4. Application: The National Immunization Survey

As an illustration of the method of disproportional sampling, we apply the concept of age targeting to the design of the National Immunization Survey (NIS). The NIS uses two phases of data collection to obtain information for a large national probability sample of young children: a random-digit-dialing (RDD) telephone survey designed to identify households with children between 19 and 35 months, followed by a mail survey of the vaccination providers of the children identified in the household survey (called the Provider Record Check), which obtains provider-reported vaccination histories for the children. At the close of the telephone interview the interviewer asks the respondent, the child(ren)'s parent or guardian, for consent to contact providers and for their names and addresses, and the Provider Record Check is conducted only for children for whom oral consent is given. Data from the Provider Record Check yield each child's number of doses for each of eleven vaccines. These counts are compared to the recommended number of doses for each vaccine (CDC 2010) to determine whether the child is up to date (UTD).

The NIS is designed to produce direct, sample-based estimates of *vaccination coverage rates* (UTD children as a proportion of all age-eligible children) within each of 56 estimation areas, consisting of 46 whole states, six large cities, and four rest-of-state areas (CDC 2012b). The estimation areas are the primary sampling strata in the NIS sampling design. A dual-frame RDD sampling design is used within each estimation area. The landline RDD sample has been conducted since 1994, while the cell-phone RDD sample was introduced in the fourth quarter of 2010.

The NIS deploys a new and independent RDD sample every calendar quarter. Vaccination coverage rates, R, are estimated using the combined sample from an annual time period. The estimator within a given estimation area is a ratio of the form $\hat{R} = \hat{Y}/\hat{X}$, where $\hat{Y} = \sum_{i \in s_c} W_i Y_i$ is an estimator of the total number of children who are UTD with respect to a given vaccine, $\hat{X} = \sum_{i \in s_c} W_i X_i$ is an estimator of the total number of age-eligible children, s_c is the set of children for whom the NIS interview (including PRC) is complete within the annual time period, Y_i is an indicator variable signifying whether the i th child is UTD, $X_i = 1$ for age-eligible children and $= 0$ for all other units in the population, and W_i is the survey weight taking into account the probability of selection, adjustments for both household and provider nonresponse, and calibration to known population counts. See the NIS Data User's Guide (CDC 2012b) for a description of the methods of weighting.

The population domain studied in the NIS is considered to be rare. In 2011 only about 18 percent of the resolved telephone numbers in the landline sample were working residential numbers and two percent of the completed screening interviews resulted in finding eligible children age 19–35 months. Given the rarity of the domain, it is reasonable to examine whether it would be possible to gain cost efficiency by using a disproportionate sampling design within high- and low-density sampling strata within each estimation area.

In what follows, we work with age-targeted lists of landline telephone numbers compiled by Marketing Systems Group (MSG) from consumer databases maintained by the marketing-research companies InfoUSA, Experian, Acxiom, and Targus. MSG and other vendors have the capability to produce lists that target various age ranges. We have conducted research for the NIS using lists targeted at ages 0–5 and 0–17 and find that both lists yield similar results. We report the results of our investigation of the list that targets households with someone age 0–17. Because the large NIS screening sample is also used for a companion survey of American adolescents aged 13–17 years, called the NIS-Teen, we report the results of our investigation of the list that targets households with someone age 0–17. This list should support the needs of both the NIS and the NIS-Teen. However, we continue this brief illustration only for the NIS sample. Because age-targeted lists are not available for cell phones, we work only with the landline sample in this illustration.

In some applications of consumer databases in sampling rare populations, it may be possible to classify the units in the overall population into three strata: (i) in the targeted domain, (ii) not in the targeted domain, and (iii) domain status indeterminate. For the current application, however, we were only able to classify telephone numbers into two categories: on or not on the age-targeted list.

Because the NIS is an important national healthcare survey that must represent the entire population of age-eligible children to the greatest extent feasible, we use the age-targeted list for stratification purposes rather than for purposes of restricting the sampling frame. The set of all telephone numbers on the landline sampling frame that are also on the list shall be deemed the high-density stratum ($h = 1$), and the set of all other numbers on the landline sampling frame that are not on the list shall be deemed the low-density stratum ($h = 2$), with $L = 2$. We observe that some market-research surveys that target consumers in a specific age range may choose to restrict the sampling frame by selecting the sample solely from an age-targeted list. This practice saves screening costs while incurring potentially large errors of undercoverage (failing to represent persons actually in the age range but not on the list). Our approach aims to achieve both complete representation of the population and some efficiency in data collection through the use of disproportionate sampling.

We illustrate the optimum allocation in terms of the annual sample size for a single, typical estimation area. A strategy of oversampling (undersampling) the high-density (low-density) stratum will tend to result in both (i) a higher observed eligibility rate in the sample and more productive data-collection operations, and (ii) a weighting effect (due to disproportionate sampling) in the estimation of population parameters of interest and, therefore, a larger sample size to maintain variance at a fixed level. A key question before us is to what extent total data collection cost can be reduced as the net effect of these two factors, one of which tends to decrease cost while the other tends to increase it.

We determine the optimum allocation under the following ideal assumptions: (a) that there is no nonresponse in the household or provider surveys, (b) that each household in the landline population of households is connected to one and only one landline, and (c) there is at most one child aged 19–35 months in the household. If the methods cited here were used in actual practice, the sample sizes would have to be adjusted for these various factors.

The model for data collection costs is (1), where $L = 2$ and n_h is the sample size of households in stratum h. The per-unit cost components, t_h, reflect numerous features of the NIS design, including the cost per telephone number for obtaining the age-targeted flag, the cost per telephone number for sample preparation and sending advance letters; the cost per telephone number for the screening interview (including both resolution of residential telephone number status and age screening); the cost per incentive given; and the cost per age-eligible household for the main interview and the PRC. The per-unit cost components must be loaded with both the costs directly expended on completed cases and a pro-rata share of the costs of all efforts expended on unproductive cases, for example, households and providers that break off or otherwise fail to complete the survey. We have analyzed recent NIS cost data and determined that the ratio of the per-unit cost components is $t_1/t_2 = 5.1$. Thus the per-unit cost of data collection in the high-density stratum is about 5 times the per-unit cost in the low-density stratum. This result is to be expected, because, as we will show, the overall eligibility rate is much higher in the high-density stratum, and therefore this stratum requires more interviewing effort than does the low-density stratum.

The vaccination coverage rates in the high- and low-density strata are quite similar, usually differing by only one or two percentage points. Thus, given the foregoing assumptions, the variance of the estimated vaccination coverage rate, $\hat{R}$, is given approximately by (2), where r_h is the overall eligibility rate within stratum h (encompassing both the age-eligibility rate and the rate of working residential numbers among the resolved telephone numbers in the selected sample), r is the overall eligibility rate across both of the sampling strata within the estimation area, $W_h = N_h/N$ is the proportion of landlines on the area-specific sampling frame that are classified in stratum h, $S^2 = R(1-R)$ is the variance component in the domain of age-eligible children.

With the cost and variance models in hand, the optimal allocation of the total sample size to the two sampling strata within an estimation area is given by (3) and (4) and the total sample size by (5).

We estimate the overall eligibility rates and population proportions using NIS data from the third and fourth quarters of 2010 (henceforth referred to as Q3–Q4 2010). Since we actually conducted the NIS in these two quarters, we know which of the selected landline telephone numbers were associated with a household with a resident child in the eligible age range, and we have since been able to determine retrospectively which of the selected landline numbers were on the age-targeted lists in those quarters. The overall eligibility rates and population proportions are given in Table 1.

While the overall eligibility rate is not high in absolute terms in either stratum, the rate in the high-density stratum is relatively much higher than the rate in the low-density stratum. The rate in the high-density stratum is almost 14 times greater than that in the low-density stratum, and about 58 percent $= r_1W_1/r_1W_1 + r_2W_2$ of the population of age-eligible children is classified in the high-density stratum. While the statistics presented in Table 1 are at the national level, we will take them to be appropriate for calculating the optimum allocation for a single, typical estimation area.

The Centers for Disease Control and Prevention have specified that the NIS sample size in an estimation area shall be large enough so that the coefficient of variation of the estimated vaccination coverage rate is 7.5 percent when the true rate is 50 percent. Thus, we can take $V^o = 0.001406$ as the value of the fixed variance. When the true vaccination coverage rate is 0.50 (or 50 percent), the variance component for eligible children is $S^2 = R(1-R) = 0.25$.

Plugging the foregoing parameter values into (3), (4), and (5) gives the optimum allocation to the high-density stratum, $n_1^o = 3,824$, the low-density stratum, $n_2^o = 22,875$, and the total sample size $n^o = 26,699$, which are cited in Table 2. The optimum allocation is expected to result in 320 completed interviews in the estimation area, with 223 in the high-density stratum and 97 in the low-density stratum.

Table 1. Overall eligibility rates and population proportions at the national level: NIS Q3–Q4 2010

Parameter	Low-density stratum, $h = 2$	High-density stratum, $h = 1$	Overall landline RDD sampling frame
Eligibility Rate, r_h	0.30%	4.10%	0.65%
Proportion of the Landline RDD Sampling Frame, W_h	0.9075	0.0925	1.0000

The ratios of the optimum sample sizes and the optimum sampling fractions are displayed in Table 3. Optimality calls for the high-density stratum to be sampled at a rate 1.64 times the rate of sampling in the low-density stratum. While the ratio of the population sizes is about 0.10, the ratio of the sample sizes is about 0.17.

By comparison, if we were to use the sampling design that actually was used for the NIS, which is essentially a proportional-allocation design, the corresponding total sample sizes given our assumptions would be those that appear in Table 4. The same sampling precision can be achieved in two different ways: (a) use of the current design, or (b) use of the optimum-allocation design. The latter design requires about 11,266 fewer telephone numbers in the released sample, because we have oversampled the high-density stratum that has the higher eligibility rates. However, the optimum-allocation design introduces a disproportionate allocation of the completed interviews and a corresponding weighting effect, and thus it requires about 15 more completed interviews to achieve the specified level of precision.

Our methods may be contrasted to those of Srinath et al. (2004), who previously tested the use of the Experian list for improving the efficiency of NIS sampling. They determined a method of sample allocation to minimize the variance of the estimated vaccination coverage rate subject to fixed sample size, and concluded that the estimator suffers from a loss of precision due to the weighting effect. From our work in Sections 2 and 3, it is clear that the optimum allocation, which involves disproportionate sampling, requires more interviews to maintain a constant level of precision. It is also clear that optimum allocation can maintain precision while reducing data-collection costs, at least for the age-targeted lists studied here.

Plugging the expected sample sizes in Table 4 into the cost model, we find that the ratio of data-collection costs, T^o/T^p, is about 0.87. The optimum allocation is expected to save about 13 percent in data-collection costs relative to the current NIS design for the landline

Table 2. Optimum allocation and expected sample sizes in a typical estimation area to minimize total cost subject to the specified variance constraint (7.5 Percent coefficient of variation for the estimated vaccination coverage rate)[a]

Landline RDD sample components	Low-density stratum, $h = 2$	High-density stratum, $h = 1$	Total landline sample size
Sample size, n_h^o	22,875	3,824	26,699
Eligible households with complete NIS interview	97	223	320

[a] The sampling sizes are computed using the national rates in Q3–Q4 2010.

Table 3. Ratios of population sizes, optimum aample sizes, and sampling fractions

W_1/W_2 = high-density population size/low-density population size	0.1019
n_1^o/n_2^o = high-density sample size/low-density sample size	0.1672
f_1^o/f_2^o = high-density sampling fraction/low-density sampling fraction[a]	1.6406

[a] The sampling fraction is $f_h^o = n_h^o/N_h$.

RDD sample. This percentage translates into considerable potential cost savings across 56 estimation areas per year. Most telephone surveys do not have, and thus do not bear the costs of, a second phase of data collection like the PRC. To test our methods in this more common setting, we repeated all of the calculations in this section assuming no PRC costs, and found that the resulting cost savings relative to proportional allocation amount to about 15 percent.

5. Summary

In this study of the use of disproportionate stratification for sampling a rare domain *D*, we made a number of assumptions, including that (a) the sampling frame covers a general population that contains both members and nonmembers of the rare domain; (b) domain membership is not known at the time of sampling; (c) the sampling design involves simple random sampling within two or more strata that vary in the density of the rare domain; (d) the parameter of interest is the mean of the rare domain; (e) the estimator of the domain mean is the standard ratio estimator; (f) classification of sampling units in or out of the rare domain based on the screening interview is conducted without error; (g) the cost of data collection arises as in (1); and (h) the variance of the ratio estimator can be represented by (2). We focused on the optimum allocation of the sample size to the several strata when one's object is to minimize the cost of data collection subject to a constraint on the variance of the ratio estimator (we also briefly treated the optimization problem when the object is to fix cost or to fix the number of interviews achieved for members of the rare domain). We find the optimum allocation to a stratum is proportional to the size of the stratum and to the root of the eligibility rate in the stratum, and is inversely proportional to the per-unit cost of data collection in the stratum. Given our assumptions, the optimum-allocation design, which oversamples the high-density stratum, introduces no bias into the

Table 4. Expected sample sizes within a typical estimation area to achieve the specified variance constraint (7.5 percent coefficient of variation for the estimated vaccination coverage rate) for two allocation regimes

Landline RDD sample components	Expected sample size given current NIS design	Expected sample size given optimum-allocation design	Difference in expected sample size (current design minus optimum-allocation design)
Sample size, n_h^o	37,965	26,699	11,266
Eligible households with complete NIS interview	305	320	−15

ratio estimator of the domain mean. Because the optimum-allocation design, by definition, minimizes the cost of data collection, it must result in non-negative cost savings relative to a proportional-allocation design. The cost savings could be small unless (a) the eligibility rates in the high-density strata are much higher than those in the low-density strata; (b) a substantial portion of the rare domain is classified in the high-density strata; and (c) the per-unit cost of the main interview is high relative to the screening cost. While the optimum-allocation design potentially saves cost, it does so through disproportionate sampling of the strata, which creates a weighting effect. Thus it actually requires more completed interviews than does the less efficient proportional-allocation design.

We illustrated the optimum-allocation design using the NIS, in which the rare domain is children 19–35 months and the parameters of interest are vaccination coverage rates for this domain. Results for the NIS are limited to the age-targeted lists obtained from the MSG vendor for the period Q3–Q4 2010.

Other surveys operating in future time periods and targeting different domains of interest should test the lists available to them. The method of disproportional stratification is broadly applicable to lists available from market-research companies as well as those derived from administrative data sources. Examples include targeted lists of people or households defined by age, race, ethnicity, income, disease registry, health insurance claims data, and voter registration status.

In deciding whether to use the optimum-allocation design, the survey statistician should be mindful of any secondary objectives for the rare-population survey, other than those embodied in the optimized objective function. For estimating other population parameters of interest, such as means for crosscutting domains, the optimum-allocation design could result in a decrease in sample size and an increase in the standard error of the estimator. These issues should be tested before the decision to implement the optimum design is taken.

6. References

Centers for Disease Control and Prevention (CDC). 2010. “Recommended Immunization Schedules for Persons Aged 0–18 Years – United States, 2010.” *Morbidity and Mortality Weekly Report* 58 (51 & 52); 1–4. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5851a6.htm (accessed 10/13/2015).

Centers for Disease Control and Prevention (CDC). 2012a. “National, State, and Local Area Vaccination Coverage Among Children Aged 19-35 Months – United States, 2011.” *Morbidity and Mortality Weekly Report* 61: 689–696. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6135a1.htm (accessed 10/13/2015).

Centers for Disease Control and Prevention (CDC). 2012b. National Immunization Survey: A User’s Guide for the 2011 Public-Use Data File. Available at: http://www.cdc.gov/nchs/nis/data_files.htm

Cochran, W.G. 1977. *Sampling Techniques*, (3rd ed.). New York: John Wiley & Sons.

Kalton, G. 2009. “Methods for Oversampling Rare Subpopulations in Social Surveys.” *Survey Methodology Journal* 35: 125–141.

Kalton, G., and D.W. Anderson. 1986. “Sampling Rare Populations.” *Journal of the Royal Statistical Society, Series A* 149: 65–82. Doi: http://dx.doi.org/10.2307/2981886.

Pasek, J., S.M. Jang, C.L. Cobb, J.M. Dennis, and C. DiSorga. 2014. "Can Marketing Data Aid Survey Research? Examining Accuracy and Completeness in Consumer File Data." *Public Opinion Quarterly* 78: 889–916. Doi: http://dx.doi.org/10.1093/poq/nfu043.

Srinath, K.P., M.P. Battaglia, and M. Khare. 2004. *A Dual Frame Sampling Design for an RDD Survey that Screens for a Rare Population*, In Proceedings of the Survey Research Methods Section, American Statistical Association, Toronto, August 8–12, 2004. (pp. 4424–4429). Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000462.pdf (accessed 10/13/2015).

Sudman, S. 1972. "On Sampling of Very Rare Human Populations." *Journal of the American Statistical Association* 67: 335–339. Doi: http://dx.doi.org/10.1080/01621459.1972.10482383.

Valliant, R., F. Hubbard, S. Lee, and C. Chang. 2014. "Efficient Use of Commercial Lists in U.S. Household Sampling." *Journal of Survey Statistics and Methodology* 2: 182–209. Doi: http://dx.doi.org/10.1093/jssam/smu006.

Waksberg, J. 1973. *The Effect of Stratification with Differential Sampling Rates on Attributes of Subsets of the Population*. In Proceedings of the Social Statistics Section, American Statistical Association, New York City, December 27–30, 1973. (pp. 429–434). Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/sections/srms/Proceedings/ (accessed 10/13/2015).

Received May 2013
Revised February 2015
Accepted March 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 559–588, http://dx.doi.org/10.1515/JOS-2015-0035

Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes

Mojca Bavdaž[1], *Deirdre Giesen*[2], *Simona Korenjak Černe*[3], *Tora Löfgren*[4], *and Virginie Raymond-Blaess*[5]

Response burden in business surveys has long been a concern for National Statistical Institutes (NSIs) for three types of reasons: political reasons, because response burden is part of the total administrative burden governments impose on businesses; methodological reasons, because an excessive response burden may reduce data quality and increase data-collection costs; and strategic reasons, because it affects relations between the NSIs and the business community. This article investigates NSI practices concerning business response burden measurement and reduction actions based on a survey of 41 NSIs from 39 countries. Most NSIs monitor at least some burden aspects and have implemented some actions to reduce burden, but large differences exist between NSIs' methodologies for burden measurement and actions taken to reduce burden. Future research should find ways to deal with methodological differences in burden conceptualization, operationalization, and measurement, and provide insights into the effectiveness and efficiency of burden-reduction actions.

Key words: Administrative burden; data collection; establishment surveys.

1. Introduction

The Fifth Principle of the United Nations' Fundamental Principles of Official Statistics (United Nations 1994, 2014) explicitly requires the data source to be selected "with regard to quality, timeliness, costs and the burden on respondents." Response burden in official business surveys is thus not a new issue. It has long been a concern for National Statistical Institutes (NSIs) (e.g., Sunter 1977; Astin 1994; Willeboordse 1997; Hedlin et al. 2005) for three types of reasons: political, methodological and strategic. The political reasons stem

[1] University of Ljubljana, Faculty of Economics, Kardeljeva pl. 17, SI-1000 Ljubljana, Slovenia. Email: mojca.bavdaz@ef.uni-lj.si
[2] Statistics Netherlands, Postbus 4481, 6401 CZ Heerlen, The Netherlands. Email: d.giesen@cbs.nl
[3] University of Ljubljana, Faculty of Economics, Kardeljeva pl. 17, SI-1000 Ljubljana, Slovenia. Email: simona.cerne@ef.uni-lj.si
[4] Statistics Norway, Pb 8131 Dep, NO-0033 Oslo. Email: tora.lofgren@ssb.no
[5] Sogeti, Route de Longwy 36, L-8080 Bertrange, Luxembourg. Email: virginie.raymond-blaess@sogeti.lu

Acknowledgments: The research reported herein received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 244767. We acknowledge the valuable contributions of many colleagues, especially our BLUE-ETS partners Dag Gravem and Gustav Haraldsen (Statistics Norway); Rudi Seljak and Petra Mohorič Peternelj (Statistical Office of the Republic of Slovenia); Dan Hedlin, Boris Lorenc, and Andreas Persson (Statistics Sweden). We would also like to thank all respondents to our survey, the editors, and referees. The views expressed in this article are those of the authors and do not necessarily reflect policies of their employers or the European Commission.

from the fact that administrative burdens imposed on businesses by legislation, which include mandatory statistical reporting, decrease the competitiveness of businesses by unproductively engaging their resources. Many countries have therefore implemented programs focused on reducing administrative burdens (OECD 2009). Examples of such programs include the Paper Work Reduction Act of 1980 in the United States and the President's Executive Order 13610 of May 10, 2012 to all US government agencies; Canada's Red Tape Reduction Commission (Red Tape Reduction Commission 2012); the EU 2007–2012 Action Programme for Reducing Administrative Burdens (European Commission 2007); and the EU Regulatory Fitness Programme (European Commission 2012b). The methodological reasons for concern about response burden are based on the growing evidence that excessive burdens may lead to problematic survey response behavior with potential consequences for data quality, especially nonresponse, late response, or measurement errors (see, for instance, Hedlin et al. 2005; Bavdaž 2010; Giesen 2012; Jones 2012; Lorenc et al. 2013; and Berglund et al. 2013). Perception of a survey task may even be more relevant in this context than the objective burden (e.g., Willeboordse 1997; Hak et al. 2003; Jones et al. 2005). Closely related to these methodological reasons are the strategic reasons, because good relations between NSIs and the business community have spillover effects in the whole field of official statistics. Businesses are an important stakeholder for NSIs because of their double role as reporting units and users of official statistics (Lorenc et al. 2012).

1.1 Burden Concept and Measurement

Despite its long presence and broad relevance, response burden is a vague concept. A politician may have in mind the total costs imposed on the whole business community, a manager may think of the time people take away from business tasks, a methodologist may focus on the feeling that a respondent experiences when confronted with a mandatory survey, and so on. Willeboordse (1997) defines response burden along four bipolar dimensions. First, he distinguishes between objective (actual) and subjective (perceived) response burden with regard to the choice of measurement perspective. Actual response burden means the money and/or time it takes to comply with data requests, and perceived burden refers to the respondents' assessment of how burdensome they find it to comply with the data request. Second, the concept may only refer to the burden itself (i.e., gross burden) or be broadened to consider the advantages of responding that reduce the amount of burden (i.e., net burden) for the unit. Third, the concept of response burden may concern the mere completion of the questionnaire (i.e., minimalistic burden) or include accompanying activities such as studying the instructions, data retrieval, and follow-up calls (i.e., maximalistic burden). Fourth, the concept may relate to the burden initially placed upon and, in an ideal world, expected from businesses (i.e., imposed burden) or to the burden that businesses bear *de facto* considering their actual response behavior (i.e., accepted burden). In an ideal world, all units would respond in a timely and accurate way; in reality, some units discard survey requests, others provide inaccurate data, and so on.

Moreover, different units of observation and various levels of aggregation may be relevant for different purposes. To illustrate the methodological challenges of burden measurement, Figure 1 shows relations between business units (BU; arranged in size

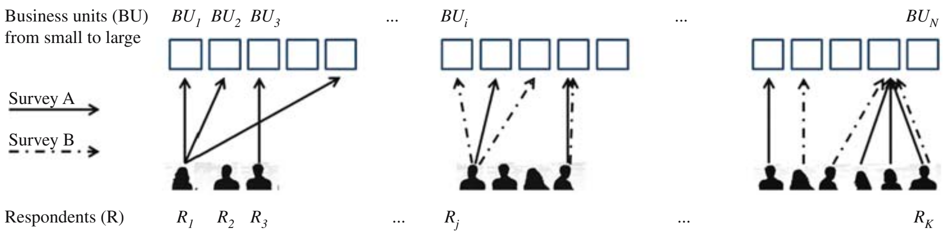


Fig. 1. Relations between business units (BU) and respondents (R) in two surveys

classes from small to large) and respondents (R) in two surveys (Survey A and Survey B). Small and also some medium-sized businesses typically hire accounting firms for all reporting (including statistical) matters. A single respondent may thus complete several questionnaires of the same survey for several businesses (see the left-most respondent involved in Survey A on behalf of several business units). By contrast, the same survey may involve several respondents at a large business (see the right-most group of respondents involved in Survey A for a large business unit). Two kinds of nesting are thus present that challenge the selection of the unit of observation: nesting business units within respondents and nesting respondents within business units. The scenario becomes loaded when surveys are added because the same respondent may be involved in more than one survey for either several businesses or a single business (see the middle group of respondents involved in both Survey A and Survey B).

The reality is more complex because surveys differ in their burden-relevant aspects (e.g., periodicity, questionnaire length, and data availability) and may involve other people in the response process in addition to respondents (e.g., data providers and authorities; see Bavdaž 2010). The same person can also have different roles in different surveys. Further complications relate to determining the relevant timeframe, delineation and dynamics of business units, changes in personnel involved in the response process, selection of the appropriate respondent for reporting burden data, and the timing and mode of collecting burden data.

The purpose of response burden monitoring ultimately determines what burden indicator (e.g., total or spread; actual or perceived) is relevant and at what level. The total actual burden at the national level may serve as a basic indicator of the total amount, progress, and outcome of national programs for administrative burden reduction. The total actual burden per survey may be considered when evaluating costs versus benefits of (new) statistical data. The spread of the total actual burden across business units, the total actual burden imposed on a business unit in a period of time, and the spread of a unit's burden in time may be useful when minimizing the impact that official surveys have at the business level. The burden that a respondent perceives in a specific survey task may contribute to better questionnaire design. Although perceptions are an inherent part of an individual, an indicator of perceived burden at a more aggregated level (a business unit, a survey, the national level) may provide greater insight because official surveys also "give rise to irritation and the perceived burden of statistics is often higher than the real burden" (European Commission 2012a, 33).

1.2 Research Problem and Research Questions

Although the Ninth Principle of the European Statistics Code of Practice specifies that "the statistical authority monitors the response burden and sets targets for its reduction over time" (European Commission 2011), the guidelines for burden conceptualization, measurement, and reduction are rather general (cf. Eurostat 2009; European Statistical System 2012) and do not offer solutions to methodological challenges. Therefore NSIs are relatively flexible and independent, but are also quite solitary in selecting conceptual definitions of response burden, tackling measurement issues, and prioritizing burden-reduction actions. With so many possible conceptual and operational differences, any comparison (e.g., across NSIs) becomes at least questionable, if not invalid, which obstructs insights into the matter and its improvement.

Our study thus aimed to provide a systematic review of the state of affairs at NSIs to help NSIs to better understand their position in comparison with other NSIs, learn from other NSIs and set priorities for actions. The study attempted to answer the following research questions:

1. How do NSIs measure response burden caused by business surveys?
2. What actions do NSIs use to reduce the response burden caused by business surveys?
3. What is known about the effectiveness of these burden-reduction actions?
4. Which, if any, are the differences between NSIs in their approaches to response burden?

Section 2 of this article describes the research method, Section 3 presents results according to the research questions, and Section 4 concludes with a discussion and summary of the findings. For the bibliography of all available documentation on response burden from our literature search including references of unpublished documents, see supplementary material on the JOS website (Supplemental_material_Bibliography_Bavdaz_et_al).

2. Research Method

A stepwise approach was used to answer the research questions. First, an extensive literature search was carried out for the period 2006–2010 (for more details, see Giesen and Raymond-Blaess 2011). This review did not find much (comparable) information about response-burden issues across NSIs and it was expected that many relevant reports would not be publicly available or updated to reflect the latest situation. A survey was thus conducted in the second step.

2.1 Questionnaire

Based on the literature review, a questionnaire was developed (see Appendix 2) that aimed to provide an overview of response-burden measurement (Part A) and reduction (Part B), and to identify any reports (additional to the ones found in our literature search) documenting response-burden measurement, response-burden reduction actions, and the effects of response-burden reduction actions.

The draft questionnaire was first reviewed by project and external experts, then revised, pretested at the NSIs of the Netherlands, Norway, Slovenia, and Sweden, and once again revised. The pretests showed two main challenges. First, respondents did not have an overview of all types of burden-reduction actions within their NSIs. For example, knowledge of questionnaire design and knowledge of sampling and estimation strategies were typically in separate departments. Therefore, attempts were made to establish presurvey contact with all NSIs to inform them about the survey and find the best respondent or response coordinator. Second, specifying the scope of a burden reduction action created conceptual problems (e.g., whether to refer to the number of surveys or businesses or respondents) and practical ones (e.g., how to treat surveys of different periodicities and lengths). Priority was given to overview rather than detail, and so the decision was made to focus on surveys and to ask for the proportion of surveys to which an action was applied using an ordinal scale:

- None: in none of our business surveys,
- Some: in some, but less than 50%,
- Most: in 50% or more, but not all,
- All: in all of our business surveys.

When developing questions about the response-burden measurement methods, four dimensions of response burden as defined by Willeboordse (1997) were taken into account:

- Objective (or actual) vs. subjective (or perceived) (Questions A1–A5 and A6–A8),
- Gross vs. net (Question A13),
- Imposed vs. accepted (Questions A5.3 and A5.2),
- Maximalistic vs. minimalistic (Question A4.a).

We asked whether the actual burden is calculated traditionally as time spent (Dale and Haraldsen 2007), in monetary costs as in the Standard Cost Model (European Commission 2009), or both. The perceived legitimacy of the survey request is probably an important aspect of how businesses perceive response burden (Dale and Haraldsen 2007), and so we also asked if NSIs had conducted any studies on how businesses perceive their organization. Furthermore, we asked about any registration of the NSI's response burden imposed on individual businesses and about any national registers of response burden caused by the government.

To assess which actions NSIs use for reducing response burden in business surveys, a list of possible reduction actions based on the literature review was created, but it only included those actions expected to be used by several NSIs and easy to capture with a single question; an open question was used to capture other actions (Question B5). Among two sets of questions, the first set referred to *the last five years* (2006–2010) and asked in what proportion of the NSIs' business surveys (none, some, most, or all surveys) the following actions had been implemented: reduction in sample sizes, reduction in the data-collection frequency, reduction in the number of requested items, and reduction in the number of recontacts with businesses (Question B1). The second set of questions referred to the current situation and asked for a list of thirteen statements to assess to which part of the NSIs business surveys (none, some, most, or all surveys) each statement applied. These statements were grouped by the use of alternatives to traditional data collection (Question B2), methods that make completing the questionnaire easier (Question B3), and actions that can improve communication and respondents' relationships to business surveys by attending to their needs (Question B4).

2.2 Survey Implementation

A letter with an invitation to participate in the web survey was sent to 45 NSIs in 43 countries covering all NSIs of the European Statistical System, (potential) candidate countries, and prominent NSIs in four non-European countries (see the list in Appendix 1). We included a request for relevant literature as an attachment to the invitation letter. We listed the literature we had already found related to that specific NSI (if any) and asked respondents to send us (references to) any other reports they could share with us. We specifically indicated that we were interested in any reports that describe the effects of burden-reduction actions on, for example, burden and data quality. This call for reports was also included as a question in the survey.

The web survey was online from November 2010 until February 2011. We saw that 41 of 45 NSIs from 39 of 43 countries responded. The achieved sample thus included 30 of the 31 NSIs in EU and EFTA countries, five of eight NSIs from (potential) candidate countries, and all six NSIs from non-European countries. Most of them responded electronically (a paper version was produced for others when requested) and after being sent reminders. For a few NSIs, we had to follow up contacts by telephone or email in order to clarify their answers or attempt to get substantive answers instead of "don't know." Our discussions with respondents revealed that it was sometimes challenging for them to answer our survey questions for all business surveys at their institute, especially because burden-measurement practices can vary over surveys and information about them does not seem to be located in a single place.

2.3 Analysis

The analysis consisted of various types of descriptive analysis. A cluster analysis aimed at identifying groups of NSIs with similar approaches to burden measurement and reduction. It was based on six binary variables describing the presence (or absence) of a specific practice:

- Actual response burden is measured in the five-year period studied (2006–2010),
- Perceived response burden is measured in the five-year period studied (2006–2010),
- Actual response burden is measured annually in the five-year period studied (2006–2010),
- Database on response burden for each business unit is kept by the NSI,
- Samples are coordinated and/or rotated (survey holidays) for all or most surveys,
- Electronic versions of self-completion questionnaires are available for all or most surveys.

The presence of a practice is considered positive: measurement of actual and perceived burden suggests NSIs' awareness of the problem; annual measurement of actual burden and a database at the business level indicate the possibility of monitoring and managing the burden; sample coordination and/or rotation for all or most surveys points to the use of more advanced statistical methods for burden reduction in a systematic way; and electronic questionnaires for all or most surveys suggest the adoption of modern technology.

After performing tests of several clustering methods, the clusters were identified using Ward's hierarchical clustering method (Ward 1963) based on the squared Euclidean distance for binary data. Analyses were done in R (R Core Team 2014) with the package *cluster* (Maechler et al. 2014). Comparisons of the clustering results were based on functions from the R-packages *fpc* (Hennig 2014) and *e1071* (Meyer et al. 2014).

The survey answers were treated as confidential unless the information was already in the public domain.

3. Results

3.1 Measurement of Response Burden

The majority of NSIs surveyed measure actual burden (i.e., the money and/or time it takes to comply); 34 out of 41 NSIs answered "yes" to the question: "In the last five years, 2006–2010, has the actual response burden incurred by businesses to comply with survey requests of your organization been calculated?" Nearly half of them (20) did this annually. Several NSIs that measure actual burden explained that this was only done for certain surveys; for example, some EU surveys or all mandatory surveys. Our follow-up contacts revealed that at least one respondent had interpreted our question as whether *total* response burden was calculated (for all survey requests). This lack of clarity in the question phrasing may have caused some other NSIs to answer "no" even though in fact they did carry out some kind of burden measurement. The reality might thus be slightly better than the results suggest.

NSIs measure perceived burden (i.e., respondents' assessments of how burdensome they find it to comply with the survey requests) less frequently: only twelve out of 41 NSIs measured perceived burden in the five-year period studied, most of those had also measured actual burden. Two-thirds of those measuring perceived burden did it every year. 17 NSIs reported that they had conducted studies on businesses' perception of the usefulness of statistics.

3.1.1 Measurement of Actual Response Burden

Out of 34 NSIs 16 calculated actual burden in time costs only and the same number of NSIs calculated both time and monetary costs, often by multiplying the time spent responding to surveys by an average wage rate. Some other NSIs also mentioned similar approaches, such as a monitoring system for the mean number of questionnaires filled in per business in a given time period.

NSIs reported using several types of data sources to calculate actual response burden. The most popular were data provided by survey respondents (29 NSIs) and expert estimates (25 NSIs). 13 NSIs used qualitative studies to assess the costs of complying. Other data sources were also reported: the frequency with which a business was drawn in samples (a practice also mentioned by other countries in some surveys); adjusted data from a previous survey; and interview time.

Table 1. Potential sources of burden explicitly included in the calculation of actual response burden (N = 33; one institute with an actual burden measurement is missing)

Sources of actual response burden	Yes	No	Don't know
Filling in the questionnaire.	31	1	1
Retrieving, collecting, and compiling the information requested.	28	4	1
Reading questions and instructions.	25	6	2
Administrative tasks (e.g., coordination) involved in survey completion.	18	10	5
Record formation specifically done for reporting obligations.	16	12	5
Recontacts with businesses about the data provided.	13	16	4
Other sources of response burden.	3	18	12

Of the 29 NSIs that used burden data provided by survey respondents, 14 collected it from subsamples and 21 collected it at the same time as the survey data they related to. Often, NSIs used several types of data sources to calculate response burden (the maximum reported by a single NSI was four different types of data sources).

Table 1 shows which potential sources of burden were explicitly included in the calculation of actual burden. For example, 18 NSIs included administrative tasks, 16 NSIs included record formation, and 13 NSIs included recontacts as part of the burden. These results suggest that response burden was operationally defined and measured in very different ways. Large discrepancies were further confirmed when comparing individual combinations of these sources. Only eight NSIs took into account all six sources of burden given in Table 1 and seven NSIs included all these sources except recontacts with businesses. Other NSIs reported using several different combinations of these sources. The most consistently used were the top three sources in Table 1 (filling in the questionnaire; retrieving, collecting and compiling requested information; and reading questions and instructions), which 25 NSIs reported they included in the calculation of actual burden. The other aspects of burden mentioned were "out-of-pocket costs/external costs" and "sixteen standard activities based on the standard cost model" (SCM Network 2005, 26–27).

An important difference in burden measurement is whether all questionnaires dispatched or only those returned are taken into account. 13 out of 34 NSIs measuring burden considered only the number of dispatched questionnaires and eleven NSIs only the number of returned questionnaires, whereas six NSIs considered both. A combination of both figures was used in some NSIs that indicated the use of different methods for different surveys. Ireland, on the other hand, does in fact publish two response-burden figures, one according to the Standard Cost Model (with the assumption of full compliance) and another one for the responding units only (Central Statistics Office 2012). The burden can also be estimated for nonrespondents (e.g., time taken to reach the decision not to respond).

3.1.2 Registers of Response Burden

Sixteen out of forty-one NSIs reported that they had a database (a register) of the burden imposed on each business unit. New Zealand used it to monitor burden ("respondent load") at the business level (Merrington et al. 2009). For each business they calculated the response burden and compared it to the relevant load thresholds for a business of that size. If businesses were unfairly burdened they were given some relief (e.g., participation in fewer surveys).

Moreover, in some countries registers were kept at the national level in order to monitor and/or reduce burden caused by all government surveys. These registers may be seen as a complement to NSIs' actual burden measurements. Such registers were reported by nine NSIs, such as the Statistical Clearing House (*www.sch.abs.gov.au*) in Australia, the Office of Management and Budget (Office of Information and Regulatory Affairs 2006) in the United States, and the Brønnøysund Register Centre (*www.brreg.no*) in Norway.

3.2 Burden-reduction Actions

Seventeen burden-reduction actions were assessed in the survey to ascertain the proportion of business surveys in which these actions had been applied. The arithmetic mean number of actions applied by the surveyed NSIs to at least some of their business surveys was twelve. One NSI had implemented none of the proposed actions and four NSIs had implemented 16 of the 17 proposed actions.

Figure 2 shows the extent to which the analyzed burden-reduction actions were present among the NSIs surveyed and how many of them applied these actions to at least half of their surveys. Burden-reduction actions that were more widely present across the NSIs tended to be more widely used within NSIs. Respondents could contact a help desk (*Help desk*) in nearly all NSIs and for a majority of surveys. Electronic versions of self-completion questionnaires (*E-qnr),* help for respondents on a website (*Website help*), and information on the concrete use of the statistical output based on the survey request (*Concrete use*) were also widely used, but around a third of NSIs surveyed still applied them to less than half of their surveys. Questionnaires were also widely tested with respondents (*Qnr testing*), but only around half of the NSIs surveyed used this testing in the majority of their surveys.

Some burden-reduction actions were present in at least 30 out of 41 NSIs surveyed, but they were not applied as often to the majority of surveys at these NSIs: sample coordination and/or rotation (*Sample coord*) was applied in the majority of surveys by only 13 NSIs, and register data replaced (part of) the data collection in the majority of surveys by nine NSIs (*Register data*). Despite their presence in more than 30 NSIs, only six NSIs applied the following three burden reduction actions to the majority of surveys: using smaller sample sizes (*Smaller samples*), requesting fewer survey items (*Fewer items*), and allowing nonautomatic fixed format files such as Excel (*Excel*).

Burden-reduction actions that were hardly ever or never used in the majority of surveys even when they were present in an NSI included: preprinting data from previous reporting periods in the questionnaire (*Preprinting*), fewer recontacts with businesses (*Fewer recontacts*), reduction of the data collection frequency (*Less frequently*) and the possibility of using automatic extraction from the businesses' administrative systems (*XBRL*). By contrast, of 19 NSIs that used a survey calendar to inform businesses of forthcoming survey requests (*Survey calendar*), as many as 14 used the calendar for the majority of surveys. Around half of the NSIs surveyed also used account managers for contacts with

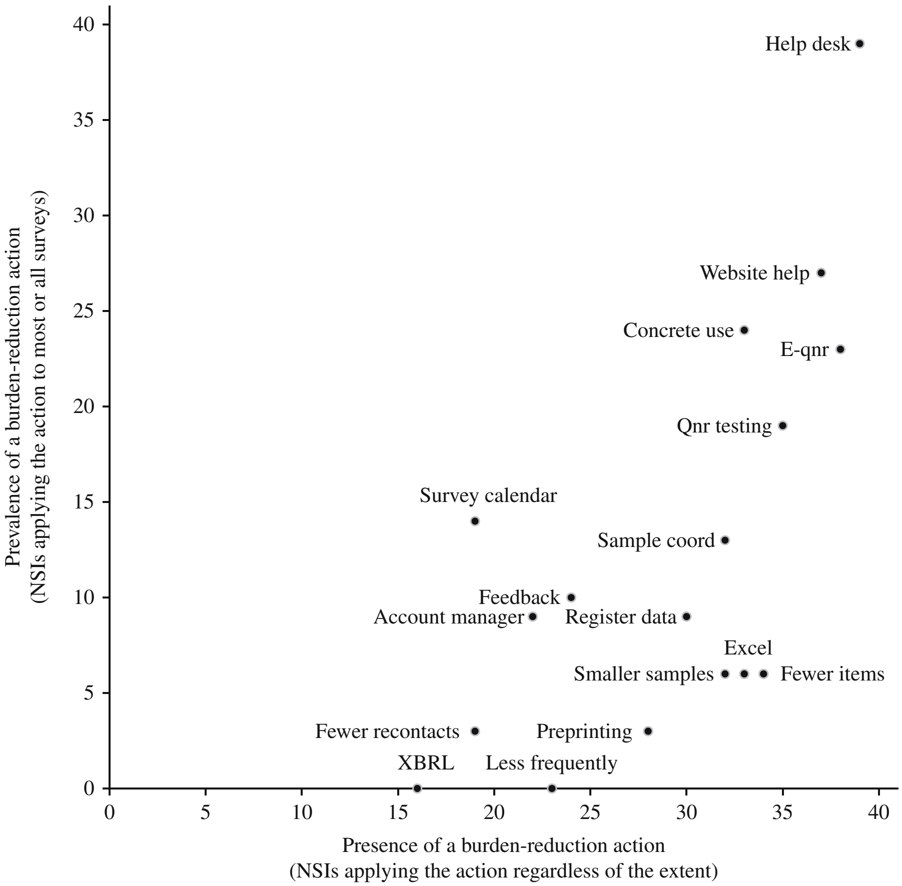


Fig. 2. Presence and prevalence of burden-reduction actions in NSIs surveyed (N = 41). Note: For complete descriptions of labels see Appendix 2, Questions B1–B4.

large businesses (*Account manager*) and personalized feedback for respondents (*Feedback*), but only about ten NSIs used those actions in the majority of surveys.

The open question containing a request to report any other unspecified reduction action yielded many responses. Some of them could be assigned to the themes of responses to the closed-ended questions. For example, some respondents interpreted the use of register data as something different to the use of administrative data. Appendix 3 gives an overview of the remaining other reduction actions and the number of times they were mentioned. It must be kept in mind that these actions are probably used at more NSIs, but these were not followed up in this study.

3.3 Effectiveness of Burden-reduction Actions

In response to our request for reports on the effects of burden-reduction actions, twelve NSIs sent us one or more reports about their efforts to reduce response burden. Some of these reports describe the development of response burden over time and, sometimes, separately for specific surveys. Examples of such publicly available reports are

Fröhlich et al. (2012) and Central Statistics Office (2012). However, very few publicly available studies investigate the effects of specific actions on response burden (Giesen and Raymond-Blaess 2011). Some exceptions are Ojo and Ponikowski (2010), who carried out a simulation study to explore the effects of dependent sampling, a method aiming at reducing response burden on the precision of estimates; a technical report by the Hungarian Statistical Office (2004) that describes a study on the expected effects of proposed burden-reduction measures on respondents and data users; and a study by Statistics Belgium (2010) that specifically states the effects of burden-reduction actions both in terms of response burden and in staff costs before and after implementation.

3.4 Approaches to Response Burden

Hierarchical clustering revealed three clusters (see Appendix 4 for details of cluster identification). The smallest cluster had approximately a quarter of the NSIs and the two other clusters had each about half of the remaining NSIs (see Table 2). The differences among clusters are particularly large with regard to the measurement of perceived response burden (in Cluster 2 all NSIs have already done it compared to only 29% of all NSIs) and much more moderate when it comes to the measurement of actual response burden (the proportion of NSIs in Cluster 1 that have already done it is 63%, compared to 83% overall).

NSIs in Cluster 1 ($N = 16$) manifested the most modest activities related to response burden issues. This cluster contains all NSIs that carried out none, one, or two of the six activities studied. About two-thirds of NSIs in this cluster measured the actual response burden in some way in the five-year period studied and less than half of them did it annually. The other four activities were present in a maximum of two NSIs. The defining

Table 2. Cluster sizes and proportions of NSIs within clusters with a specific practice considered in clustering

	Cluster 1 ($N = 16$) Modest response burden activity %	Cluster 2 ($N = 10$) Awareness of perceived response burden %	Cluster 3 ($N = 15$) Actual response burden in focus %	Total ($N = 41$) %
Actual response burden measured	63	90	100	83
Perceived response burden measured	13	100	0	29
Actual response burden measured annually	25	50	73	49
NSI database on response burden	13	20	80	39
Sample coordination and/or rotation for all or most surveys	0	60	47	32
Electronic questionnaires for all or most surveys	6	80	93	56

characteristic of this cluster is that none of its NSIs applied sample coordination and/or rotation to the majority of their surveys.

NSIs in both Cluster 2 ($N = 10$) and Cluster 3 ($N = 15$) showed much more activity with regard to response burden compared to Cluster 1 because they reported between three and five of the six activities. However, they had a different focus. NSIs in Cluster 3 concentrated on actual response burden. They all measured it in some way in the five-year period studied and the majority measured it annually, but none of them measured perceived response burden. By contrast, measurement of perceived response burden may be considered as the defining characteristic of Cluster 2, because all of its NSIs measured the perceived burden in some way in the observed five-year period. The other larger difference between Cluster 2 and Cluster 3 relates to response-burden databases. The majority of NSIs in Cluster 3 kept such a database, whereas only a minority of NSIs in Cluster 2 did. A closer look at the NSIs in Cluster 2 reveals that half of these NSIs applied sample coordination and/or rotation to the majority of their surveys and at the same time reported having no response-burden database. It is possible to claim that even these NSIs were practically ready for burden management, because burden registration is just a step away if a system infrastructure for sample coordination is already in place.

4. Discussion

Our study aimed to provide an overview of the situation regarding response-burden issues, focusing on Europe but with some extra-European countries included. It became clear during data collection that these issues cannot be covered in great detail because the data on burden-measurement and -reduction actions were either scattered around the NSIs or nonexistent. Most NSIs did not have a central person or department coordinating burden-measurement and burden-reduction actions. Notable exceptions were the Ombudsman for response burden at Statistics Canada (Sear 2011) and the Respondent Advocate at Statistics New Zealand (Statistics New Zealand 2008). Therefore it cannot be excluded that some actions were underreported and that the reality might be slightly better than the results suggest.

A closer inspection of burden measurement revealed that there were large differences in methodologies between NSIs and also within NSIs. These differences referred to the conceptual and operational definitions (e.g., monetary burden versus time burden, inclusion of recontacts and nonrespondents), type and number of data sources used, calculation procedures, and so on. Some differences might be negligible for the burden level (e.g., inclusion of nonrespondents when the response rate is high) but others quite substantial (e.g., recontacts in a complex survey with many questionable items). These methodological differences reflect differences in both the purpose and quality of burden measurement. In order to address the political reasons for burden measurement it might be sufficient for an NSI to consistently use current measurement through time, thus tracking only changes in the level of actual burden, which is quite low when presented in relative terms (e.g., compared to other administrative burdens). Comparisons for methodological or strategic reasons based on methodologically different indicators within an NSI and across NSIs are, however, much more problematic because they focus on burden levels (e.g., acceptable levels of actual burden per business, levels of perceived burden that affect a respondent's behavior in a survey, effects of a certain burden reduction action, etc.).

Our study results also indicate that most NSIs surveyed actively engage in activities related to response burden. The great majority of NSIs surveyed had – in accordance with the European Statistics Code of Practice – measured actual response burden in the five-year period studied, nearly half of them annually. These NSIs seemed prepared to respond to political pressures because they had a means of monitoring the actual burden imposed on businesses at the national level that is typically at the heart of political debates. Some NSIs also had policies guiding their burden-reduction activities. However, actions that directly reduce actual burdens imposed on businesses were not so widely applied within NSIs, regardless of whether they were common across NSIs (e.g., fewer survey items requested and smaller sample sizes) or less common (e.g., fewer recontacts with businesses and reduction of data-collection frequency). The highest prevalence was noted for register data replacing (part of) data collection, which nonetheless was still not common.

When analyzing burden-reduction actions, the first impression was that the NSIs really focused on strategic reasons and tried to establish and/or improve their relations with the business community by offering help and explaining how the collected data would be used. These actions were probably relatively easy to implement because they did not require much change in the work organization. Other actions likely required greater interventions because they demanded redesigned processes (e.g., electronic versions of self-completion questionnaires), a redesigned information system (e.g., sample coordination and/or rotation, survey calendar) or a broader knowledge (e.g., account managers for contacts with large businesses). Among these actions, the NSIs surveyed performed best on the electronic versions of self-completion questionnaires, probably because of other government initiatives for electronic reporting (e.g., on taxes), expected cost savings and business pressures. If the NSIs wanted to manage actual burden well – that is, to monitor its amount and spread over time – they first needed burden data per business over time. Such databases or registers were, however, set up in less than half of the NSIs surveyed, although a few NSIs might be close to having registers of this kind because they possessed the infrastructure for survey coordination.

Some actions mentioned above could also be understood as methodologically motivated burden-reduction actions, especially when considering the perceived burden, such as offering electronic versions of questionnaires, offering help, and explaining the reasons for survey requests (addressing the "irritation" burden, see High Level Group of Independent Stakeholders on Administrative Burdens 2009). Less than a third of the NSIs surveyed, however, measured the perceived response burden. Given that the survey questionnaire is the essential instrument of data collection and the main "source" of any kind of burden, and that the European Statistics Code of Practice explicitly prescribes systematic testing of questionnaires prior to the data collection, it was expected that testing questionnaires with respondents would be common across NSIs. A surprising finding was that half of the NSIs surveyed used testing with respondents in less than half of their surveys. However, some NSIs also had other initiatives that promised to make a questionnaire's completion easier, such as designing survey questions to be as close as possible to accounting categories, regularly reviewing questionnaires, testing web-questionnaire usability, and so on.

Given the burden-reduction actions implemented, the NSIs surveyed seemed to work simultaneously on political, strategic, and methodological reasons. Our study does not reveal how or why the NSIs decided on their combinations of implemented actions. Variations among the NSIs in these combinations may partly be caused by structural differences such as legal limitations (particularly with respect to getting access to administrative data) or other government initiatives, and by the human, technological, and financial resources available. They probably also reflect the fact that little is known about the effects of various response-burden reduction actions on response burden, data quality, and (net) costs for NSIs. Some burden-reduction actions have an obvious effect on actual response burden (such as substituting direct data collection with administrative sources), but even for these the effects on both actual burden and (net) costs are not often measured or publicized. Furthermore, there is little evidence for other actions, and there is even less evidence about effects on perceived response burden, data quality, and (net) costs for NSIs.

This lack of data is quite surprising for an information producer in the era of big data and omnipresent demands for improved efficiency. This overview of the situation might stir the NSIs to start collecting evidence in order to understand their own positions better. The overview discusses actions used at the time of data collection and indicates to what extent most of these actions were present and used in the NSIs surveyed, thus establishing a common reference or “norm”. Every NSI can now better compare itself to other NSIs. Such benchmarking then urges the NSIs to respond by at least reconsidering, if not improving, their own activities (see Triantafillou 2007). The best-performing NSIs may be encouraged to fill the remaining gaps and the underperforming NSIs to reach the “average”. Benchmarking can be supported by the cluster-analysis results, which suggested marked disparities in approaches to response burden among the NSIs surveyed. Some differences between the NSIs might be attributable partly to the diverse institutional environments in which they operate. These diverse institutional environments represent different levels of red tape, social responsibility, business friendliness, information disclosure, access to modern technology, and so on, but also different historical backgrounds. The situation seemed especially challenging for the NSIs of some smaller countries. These NSIs in particular may benefit from sharing knowledge on response burden among the NSIs in order to avoid reinventing the wheel.

4.1 Future Work and Research

It seems that successful management of response burden requires different disciplines within an NSI to work together; at least experts from statistical units, methodology, data collection, and communication should be involved. A central location for measuring and managing response burden seems an efficient way to facilitate and stimulate such cooperation within and across NSIs. It should also result in more data for benchmarking purposes, but these data can only be useful if they are comparable.

Methodological differences in burden conceptualization, operationalization, and measurement might be dealt with, to a certain extent, by estimating the effects of these differences on the results. Current knowledge on these issues is, however, limited. We therefore call for more research in order to better understand what concepts are relevant for what purpose, what sources of burden are (empirically) important in what context, what data sources are reliable, how often to measure the burden, how to measure the perceived burden of a single respondent and multiple respondents, and so on.

A longer-term objective, although one not easy to achieve, should be to attain some harmonization of burden definitions, measurement, and indicators with the purpose of allowing direct comparisons without corrections. Moreover, new indicators might be developed to quantify the burden per data point collected. The 2007 *Handbook for Monitoring and Evaluating Business Survey Response Burden* could be used as a starting point. A standardized framework, however, requires active dissemination and follow-up; the active involvement of Eurostat and other international organizations would certainly be helpful for such processes (see also Giesen et al. 2011).

We also recommend that NSIs first of all document and monitor their burden-reduction initiatives better, and share their knowledge both within and between NSIs. We also recommend more studies comparing burden-reduction action alternatives or at least describing the "before and after" situation to be able to make better decisions about priority actions. In order to make well-informed decisions, a step forward in the research into business survey data-collection methodology is indispensable. This research should take into account that it may not be easy to change the opinions and behavior of respondents to business surveys, who already have established routines and attitudes concerning NSI survey requests. The research into effects of burden-reduction actions should include both novice and experienced respondents and should monitor long-term effects. Furthermore, it seems advisable to design studies that can detect how business characteristics such as size class, type of industry, and past response behavior affect their reactions to burden-reduction actions. It may well be that NSI actions can be more effective and efficient if tailored to these characteristics.

Appendix 1: List of Targeted NSIs ($N = 45$) With An Indication of Nonresponse

NSIs of the EU and EFTA Countries (27+4 = 31 Units; 1 Nonrespondent)

1. Austria: Statistik Austria
2. Belgium: Statistics Belgium
3. Bulgaria: National Statistical Institute
4. Cyprus: Statistical Service of Cyprus
5. Czech Republic: Czech Statistical Office
6. Denmark: Statistics Denmark
7. Estonia: Statistics Estonia
8. Finland: Statistics Finland
9. France: National Institute of Statistics and Economic Studies (INSEE)
10. Germany: Federal Statistical Office
11. Greece: National Statistical Service of Greece
12. Hungary: Hungarian Central Statistical Office
13. Iceland: Statistics Iceland
14. Ireland: Central Statistics Office Ireland
15. Italy: Italian National Institute of Statistics (ISTAT)
16. Latvia: Central Statistical Bureau of Latvia
17. Liechtenstein: Office of Statistics
18. Lithuania: Statistics Lithuania
19. Luxemburg: National Institute of statistics and economic studies (STATEC)
20. Malta: National Statistics Office
21. Netherlands: Statistics Netherlands
22. Norway: Statistics Norway
23. Poland: Central Statistical Office
24. Portugal: Statistics Portugal
25. Romania: National Institute of Statistics
26. Slovakia: Statistical Office of the Slovak Republic
27. Slovenia: Statistical Office of the Republic of Slovenia
28. Spain: National Statistics Institute
29. Sweden: Statistics Sweden
30. Switzerland: Swiss Federal Statistical Office
31. United Kingdom: Office for National Statistics

NSIs of the (potential) Candidate Countries (8 Units; 3 Nonrespondents)

1. Albania: Institute of Statistics
2. Bosnia and Herzegovina: Agency for Statistics of Bosnia and Herzegovina
3. Croatia: Central Bureau of Statistics
4. FYROM: Statistical Office of Macedonia
5. Kosovo: Statistical Office of Kosovo
6. Montenegro: Statistical Office of Montenegro (MONSTAT)

7. Serbia: Statistical Office of the Republic of Serbia
8. Turkey: Turkish Statistical Institute

Non-European NSIs (6 Units; no Nonrespondents)

1. Australia: Australian Bureau of Statistics
2. Canada: Statistics Canada
3. New Zealand: Statistics New Zealand
4. USA: Bureau of Labor Statistics
5. USA: Census Bureau
6. USA: National Agricultural Statistics Service

Appendix 2: Web Survey Questionnaire

Note: Labels for burden-reduction actions given in bold in brackets in questions B1-B4 were added for easier interpretation of Figure 2 and did not appear in the questionnaire.

Part A: Measurement of Response Burden

A1. ** *A1 Help text: Businesses = organizations that produce goods and services for profit. Actual response burden = the money and/or time it takes to comply with survey requests.* **

This question is about *actual* response burden. We define actual response burden as the money and/or time it takes to comply with survey requests.

In the last five years, 2006–2010, has the actual response burden incurred by businesses to comply with survey requests of your organization been calculated?
** *1 choice only, no empty* **

1. Yes → A2
2. No → A5
3. Don't know → A5

A2. Has the actual response burden been calculated in time spent, monetary costs or both?
** *1 choice only, no empty* **

1. In time costs only
2. In monetary costs only
3. Both in time spent and monetary costs
4. Don't know

A3.a In the last five years (2006–2010), have the following kinds of data have been used to calculate the actual response burden of businesses?

** *1 choice only, no empty* **

Estimates from staff/experts.	Yes/No/Don’t know
Qualitative studies assessing the costs of complying (for example observation of respondents completing the questionnaire).	Yes/No/Don’t know
Information provided by respondents in surveys (for example through an additional survey question on time taken to complete questionnaire).	Yes/No/Don’t know
Other data** *if Other data = yes then A3.b* **	Yes/ No/Don’t know

A3.b Please briefly describe the other data used to calculate the actual response burden of businesses.
*** Large memo field ***

A4.a In the last five years (2006–2010), which potential sources of response burden have been explicitly included in the calculation of businesses’ actual response burden*?*
*** 1 choice only, no empty ***

Record formation specifically done for reporting obligations.	Yes/No/Don’t know
Administrative tasks (e.g., coordination) involved in survey completion.	Yes/No/Don’t know
Reading questions and instructions.	Yes/No/Don’t know
Retrieving, collecting and compiling requested information.	Yes/No/Don’t know
Filling in the questionnaire.	Yes/No/Don’t know
Recontacts with businesses about the data provided.	Yes/No/Don’t know
Other sources of response burden** *if Other sources = yes then A4.b* **	Yes/No/Don’t know

A4.b Please briefly describe the other sources of response burden used to calculate the actual response burden of businesses.
*** Large memo field ***

A5. For the last five years (2006–2010), which of the following statements are true for the methods used to calculate the actual business response burden due to survey requests of your organization?
*** 1 choice only, no empty ***

Actual response burden is calculated each year.	Yes/No/Don't know
Actual response burden is based on the number of businesses that *respond* to survey requests.	Yes/No/Don't know
Actual response burden is based on the total number of survey requests sent out (including nonresponse).	Yes/No/Don't know
Data used for actual response burden calculation are based on information provided by *samples* of business-survey respondents.	Yes/No/Don't know
Data on actual response burden are collected at the same time as the survey data they relate to (integrated or attached to survey request).	Yes/No/Don't know

A6. ** *A6 Help text: Perceived response burden = the respondents assessment/qualification of how burdensome the survey request is.* **

This question is about *perceived* response burden. We define perceived response burden as the respondents' assessments of how burdensome they find it to comply to the survey requests. This could be measured by questions on how time consuming and/or burdensome they think the survey questionnaire is.

In the last five years, 2006–2010, has the *perceived* response burden of business respondents caused by your survey requests been measured in some way?

1. Yes → A6
2. No → A9

A7. Has the perceived response burden of business respondents caused by your data requests been measured at least once a year in the last five years (2006–2010)?

1. Yes
2. No
3. Don't know

A8. Are the data on perceived response burden of businesses collected at the same time as the survey data they relate to?

1. Yes
2. No
3. Don't know

A9.a Do you have any additional information about the calculation and measurement of business response burden that would help us understand your practices? Further on in this questionnaire you can give references to any documents you might be able to share on response-burden measurement.

1. Yes
2. No → A10

A9.b ** If A9.a = yes then A9.b
Please put any additional information on the calculation and measurement of business response burden below.
Large memo field

A10. Does your organization keep a database on response burden for each business unit? By this we mean a register-like database that contains information on the total response burden for each business.

1. Yes
2. No
3. Don't know

A11. In your country, is there an authority or register that records survey requests posed on businesses by your organization as well as by other governmental organizations?
** *1 choice only, no empty* **

1. Yes
2. No ⇨ A13
3. Don't know ⇨ A13

A12. What is the name of the authority or register that records data requests by government organizations?
** *Medium sized memo field* **

A13. In the last five years, 2006–2010, has any study been done on how businesses perceive your organization – either in their capacity of data providers, data users or both? Please include any studies on businesses' perceived usefulness of statistics.
** *1 choice only, no empty* **

1. Yes
2. No → A15
3. Don't know

A14. Please describe how the data on businesses' appreciation of your organization have been collected. Any related documents about this you can share with us can be mentioned in question A16.
** *large memo field* **

A15. Can you help us find any recent (2005–2010) reports on how your institute measures business response burden and/or the businesses' appreciation for your institute. As an attachment to the invitation letter for this survey we included a list of papers we already found for your organization (if any).

Please enter any (other) references to reports below or send the reports to *rbsurvey@cbs.nl* or to Deirdre Giesen, Divison of Methodology and Quality, Room 1C33, PO Box 4481, 6401 CZ Heerlen, The Netherlands.

A16. Who has answered the questions above on the measurement of response burden?

Name: ––––*Noempty*
Function: ––––*Optional field*
Specific domain of expertise: ––––*Optional field*
Department: ––––*Optional field*
E-mail: ––––*Noempty, email check*
Telephone number: ––––*Noempty*

A17.a Who should we contact in your organization for additional information on the measurement of response burden?
*** 1 choice only, no empty ***

- Same person as mentioned in previous question yes/no
- Other person(s) yes/no

A17.b *If Other person(s) is yes than A17b*

Please mention name, telephone number, e-mail address and, if applicable, specific domain of expertise of the person(s) we can contact for additional information on the measurement of response burden in your organization.

Large memo field

Part B Reduction of Response Burden

The goal of the following questions is to assess which practices national statistical institutes use that can reduce response burden in business surveys.

B1 In the last five years (2006–2010), in which part of your business surveys have the following actions been implemented?
*** 1 choice only, no empty ***

- None: in none of our business surveys
- Some: in some, but less than 50%
- Most: in 50% or more, but not all
- All: in all of our business surveys

Reduction of sample size(s). **(Smaller samples)**	none/some/most/all/don't know
Reduction of the frequency of data collection. **(Less frequently)**	none/some/most/all/ don't know
Reduction of the number of requested items in survey requests. **(Fewer items)**	none/some/most/all/ don't know
Reduction of the number of recontacts with businesses. **(Fewer recontacts)**	none/some/most/all/ don't know

B2 Currently, to which part of your business surveys does each statement below apply?
*** 1 choice only, no empty ***

- None: to none of our business surveys
- Some: to some, but less than 50%
- Most: to 50% or more, but not all
- All: to all of our business surveys

Register information has replaced (part of) the data collected from businesses. **(Register data)**	none/some/most/all/don't know
(Part of) the data can be provided by automatic extracted files from the businesses' administrative systems, for example XBRL. **(XBRL)**	none/some/most/all/don't know
(Part of) the data can be provided by non-automatic fixed format files, for example excel files. **(Excel)**	none/some/most/all/don't know
Samples are coordinated and/or rotated (survey holidays). **(Sample coord)**	none/some/most/all/don't know

B3 Currently, to which part of your business-survey questionnaires does each statement below apply?
*** 1 choice only, no empty ***

- None: to none of our business-survey questionnaires
- Some: to some, but less than 50%
- Most: to 50% or more, but not all
- All: to all of our business-survey questionnaires

Data of previous reporting periods are preprinted in the questionnaires (e.g., dependent interviewing). **(Preprinting)**	none/some/most/all/ don't know
Questionnaires have been tested with respondents to assess how well they understand the questionnaire and are able to provide the data. **(Qnr testing)**	none/some/most/all/ don't know
Electronic versions of self-completion questionnaires are available. **(E-qnr)**	none/some/most/all/ don't know

B4 Currently, to which part of your business surveys does each statement below apply?
*** 1 choice only, no empty ***

- None: to none of our business surveys
- Some: to some, but less than 50%

- Most: to 50% or more, but not all
- All: to all of our business surveys

Survey requests are included in a survey calendar that gives businesses an overview of which surveys they can expect from your organization. **(Survey calendar)**	none/some/most/all/don't know
Respondents can contact a help desk if they have questions about a survey (e.g., a specific phone number and/or e-mail address). **(Help desk)**	none/some/most/all/don't know
Respondents can find help on a website (for example frequently asked questions). **(Website help)**	none/some/most/all/don't know
Information is provided on the concrete use of the statistical output based on the survey request. **(Concrete use)**	none/some/most/all/don't know
Respondents can receive personalized statistical feedback. **(Feedback)**	none/some/most/all/don't know
The contacts with large businesses are managed by a single account manager. **(Account manager)**	none/some/most/all/don't know

B5 Has your organization conducted any other activities to reduce response burden for business surveys? If so, please describe below.
Large memo field

B6 Can you help us find any recent (2005–2010) reports on how your institute aims to reduce businesses' response burden? We are particularly interested in any studies on the effects of these activities on response burden and data quality.

As an attachment to the invitation letter for this survey we included a list of papers we already found for your organization (if any).

Please enter any (other) references to reports below or send the reports to *rbsurvey@cbs.nl* or to Deirdre Giesen, Divison of Methodology and Quality, Room 1C33, PO Box 4481, 6401 CZ Heerlen, The Netherlands.

B7a Have the above questions on response-burden reduction been answered by the same person who answered the questions on response-burden measurement?

Yes/no

If No Then B7b

Who answered the questions on response-burden reduction?

Name: –––– *Noempty*
Function: –––– *optional field*
Specific domain of expertise: –––– *optional field*
Department: –––– *optional field*
E-mail: –––– *Noempty, email check*
Telephone number: –––– *Noempty*

B8a Who should we contact in your organization for additional information on response-burden reduction?

- same as person mentioned in previous question
- someone else

If someone else Then B8b

Please mention name, telephone number, e-mail address and, if applicable, specific domain of expertise of the person(s) we can contact for additional information on the reduction of response burden in your organization.

Large memo field
Closing message

Appendix 3: Other Burden-reduction Actions Reported by NSIs Surveyed

Policies outside and within the NSIs

- Better coordination across public agencies and authorities (3x)
- Seeking access to administrative data (3x)
- Policy not to collect data if information is available in administrative data (3x)
- Program of data collection split in two chapters, direct data collection and usages of administrative data from other government bodies (1x)
- Policy not to ask for the same information in different questionnaires (1x)
- 'No gold-plating' rule – implementing minimum requirements only (2x)
- 'One-in, one-out' rule (1x)
- Load Threshold Policy: proactive relief to businesses in accordance to size (1x)

Methods to make a questionnaire's completion easier

- Regular monitoring/reviewing of questionnaires to detect problems of respondents (3x)
- Testing usability of electronic web-based data collections (1x)
- Offering questionnaires in multiple modes (2x)
- Prefilling questionnaires with administrative data (1x)
- Redesign of questionnaires to align them as far as possible with the Profit & Loss and Balance Sheet account entries (1x)
- Establishment of Accounting Practices Unit that seeks to reconcile survey questions with business record keeping (1x)
- Establishment of response-improvement research staff to do research on questions (1x)
- All questionnaires can be downloaded and sent back electronically through a public website (1x)

Actions to improve communication and relationship with respondents

- Development of special shorter questionnaires for small businesses (2x)
- Reduction of the level of detail asked on a number of questionnaires (1x)
- Interaction between data collectors and respondents via ICT and Internet in order to complete questionnaires aiming at efficiency of the data-capture process (1x)
- Website developed specifically for businesses, both as respondents and users (1x)
- Accept a copy of the balance sheet of the annual account instead of filling in SBS questionnaire (1x)

Appendix 4: Clustering

After testing various clustering methods, Ward's hierarchical clustering method was selected because it offered the most meaningful interpretations of the results obtained. Several permutations of unit ordering were compared in order to observe the effect of unit ordering on clustering results. At lower levels (with many small clusters) all results mostly matched. At higher levels, three main groups with some differences were mostly detected.

The best clustering result was identified based on the criterion-function value (the within sum of squares), some other theoretical measurements for cluster validation—especially the silhouette plot, the average silhouette width, and the height of aggregation in the hierarchical tree (Kaufman and Rousseeuw 1990; Everitt et al. 2001)—and cluster interpretability. The criterion-function values (the within sum of squares) ranged from 64.02083 to 75.45125. The solution presented here, with three well-separated clusters, had the smallest obtained criterion-function value (64.02083) and offered a meaningful interpretation of the clusters. These clusters with 16, ten, and 15 NSIs can be seen clearly in the graphical presentation of the aggregation procedure (dendrogram) in Figure A1 (the plot was cropped at the bottom where NSI names appear for confidentiality reasons).

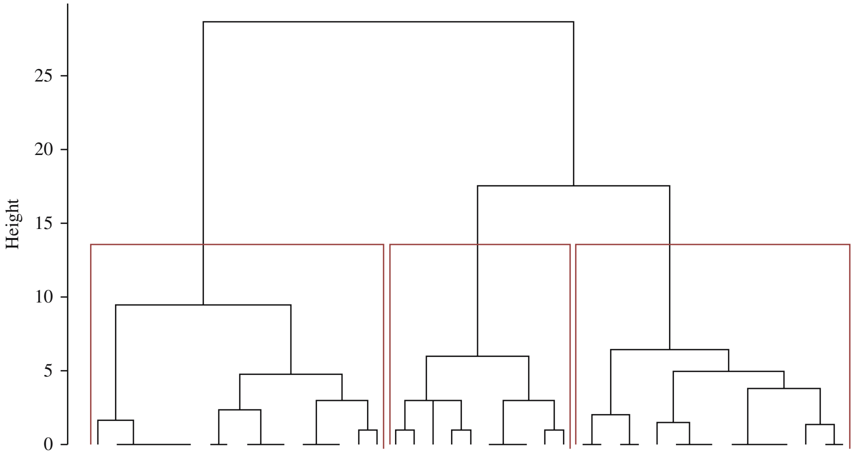


Fig. A1. Dendrogram of 41 NSIs, using six selected variables and obtained using Ward's hierarchical clustering method based on squared Euclidian distance for binary data.

The silhouette plot shows how well each individual unit fits into the cluster. The silhouette plot and the values of the average silhouette width of the clusters in Figure A2 suggest that units fit somewhat better in the larger two clusters compared to the smallest cluster.

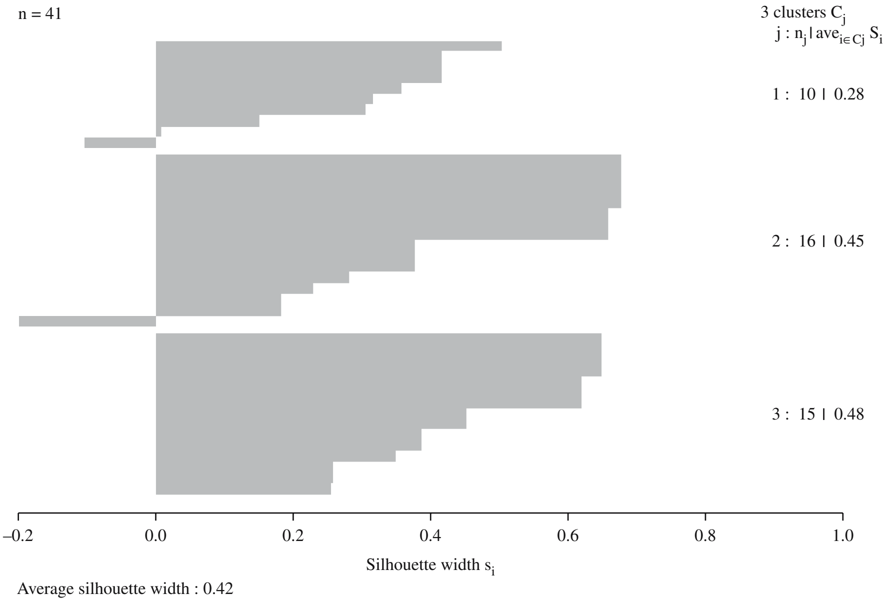


Fig. A2. Silhouette plot clustering 41 NSIs into three clusters.

5. References

Astin, J. 1994. "Statistical Form-Filling by Industry: Net Burden or Net Benefit?" *International Statistical Review* 62: 87–97. DOI: http://dx.doi.org/10.2307/1403547.

Bavdaž, M. 2010. "Sources of Measurement Errors in Business Surveys." *Journal of Official Statistics* 26: 24–42.

Berglund, F., G. Haraldsen, and O. Kleven. 2013. "Causes and Consequences of Actual and Perceived Response Burden Based on Norwegian Data." In *Deliverable 8.1 of the BLUE-ETS Project*, 29–35. Available at: http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable8.1.pdf (accessed October 2014).

Central Statistics Office. 2012. *Response Burden Barometer 2011. Measurement of Administrative Burden imposed on Irish Business by Central Statistics Offices Inquiries*. Available at: http://www.cso.ie/en/media/csoie/releasespublications/documents/multisectoral/2011/responseburden11.pdf (accessed April 2012).

Dale, T. and G. Haraldsen, eds. 2007. *Handbook for Monitoring and Evaluating Business Survey Response Burdens*. European Commission, Eurostat. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/HANDBOOK%20FOR%20MONITORING%20AND%20EVALUATING%20BUSINESS%20SURVEY%20R.pdf (accessed April 2013).

European Commission. 2007. *Action Programme for Reducing Administrative Burdens in the European Union*. Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions. Brussels, January 24, 2007. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2007:0023:FIN:EN:PDF (accessed April 2013).

European Commission. 2009. *Reducing Administrative Burdens in the European Union — Annex to the 3rd Strategic Review on Better Regulation*. Commission Working Document 16. Brussels, January 28, 2009. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0016:FIN:en:PDF (accessed April 2013).

European Commission. 2011. *European Statistics Code of Practice for the National and Community Statistical Authorities*. Adopted by the European Statistical System Committee, September 28, 2011. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF (accessed April 2013).

European Commission. 2012a. *Action Programme for Reducing Administrative Burdens in the EU Final Report*. Strasbourg, December 12, 2012. Available at: http://ec.europa.eu/smart-regulation/refit/admin_burden/docs/com2012_746_swd_ap_en.pdf (accessed October 2014).

European Commission. 2012b. *EU Regulatory Fitness*. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Strasbourg, December 12, 2012. Available at: http://ec.europa.eu/governance/better_regulation/documents/1_EN_ACT_part1_v8.pdf (accessed April 2013).

European Statistical System. 2012. *Quality Assurance Framework of the European Statistical System*. Version 1.1. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/QAF_2012/EN/QAF_2012-EN.PDF (accessed October 2014).

Eurostat. 2009. *ESS Standard for Quality Reports*. Eurostat Methodologies and Working Papers. Luxembourg: Office for Official Publications of the European Communities. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/ESQR_FINAL.pdf (accessed April 2013).

Executive Order 13610 of May 10. 2012. Identifying and Reducing Regulatory Burdens. Federal Register Vol. 77, No. 93. Available at: http://www.whitehouse.gov/sites/default/files/docs/microsites/omb/eo_13610_identifying_and_reducing_regulatory_burdens.pdf (accessed April 2013).

Everitt, B.S., S. Landau, and M. Leese. 2001. *Cluster Analysis*, 4th ed. London: Edward Arnold Publishers.

Fröhlich, M., U. Oschischnig, and N. Rainer. 2012. "Meldepflichten und Belastung der Wirtschaft durch Erhebungen der Statistik Austria 2001–2011." *Statistische Nachrichten* 9: 729–741. Available at: http://www.statistik.at/web_de/statistiken/unternehmen_arbeitsstaetten/respondentenbelastung/respondenten_belastungsbarometer/index.html (accessed April 2013).

Giesen, D., ed. 2011. *Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes*. Deliverable 2.2 of the BLUE-ETS Project. Available at: http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable2.2.pdf (accessed April 2013).

Giesen, D. 2012. “Exploring Causes and Effects of Perceived Response Burden.” In Proceedings of the Fourth International Conference on Establishment Surveys: American Statistical Association, June 11–14, 2012, Montréal, Canada. Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/meetings/ices/2012/papers/302171.pdf (accessed October 2014).

Giesen, D., G. Haraldsen, and M. Bavdaž. 2011. “Response Burden Measurement.” In Deliverable 2.2 of the BLUE-ETS Project, 15–23. Available at: http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable2.2.pdf (accessed April 2013).

Giesen, D. and V. Raymond-Blaess, eds. 2011. *Inventory of published research: Response burden measurement and reduction in official business statistics. A literature review of national statistical institutes’ practices and experiences*. Deliverable 2.1 of the BLUE-ETS Project. Available at: http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable2.1.pdf (accessed April 2013).

Hak, T., D. Willimack, and A. Anderson. 2003. “Response Process and Burden in Establishment Surveys”. In Proceedings of the Section on Government Statistics: American Statistical Association, August 3–7, 2003, San Francisco, 1724–1730. Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/sections/srms/proceedings/y2003/Files/JSM2003-000457.pdf (accessed April 2013).

Hedlin, D., T. Dale, G. Haraldsen, and J. Jones, eds. 2005. *Developing Methods for Assessing Perceived Response Burden. Research report*. Stockholm: Statistics Sweden, Oslo: Statistics Norway, and London: Office for National Statistics. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/DEVELOPING%20METHODS%20FOR%20ASSESSING%20PERCEIVED%20RESPONSE%20BURD.pdf (accessed April 2013).

Hennig, C. 2014. *fpc: Flexible procedures for clustering*, R package version 2.1–9.

High Level Group of Independent Stakeholders on Administrative Burdens. 2009. *Subject: Priority Area Statistics*. Available at: http://ec.europa.eu/smart-regulation/refit/admin_burden/docs/enterprise/files/hlg_opinion_070709_statistics.pdf (accessed October 2014).

Hungarian Statistical Office. 2004. *External Trade Statistics*, Technical Implementation Report. Transitional Facility, 2004. Project No 6. Agreement No 19100.2005.001-2005.532. Budapest: Department of External Trade Statistics. Unpublished document.

Jones, J. 2012. “Response Burden: Introductory Overview Lecture.” In Proceedings of the Fourth International Conference on Establishment Surveys: American Statistical Association, June 11–14, 2012, Montréal, Canada. Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/meetings/ices/2012/papers/302289.pdf (accessed April 2013).

Jones, J., J. Rushbrooke, G. Haraldsen, T. Dale, and D. Hedlin. 2005. “Conceptualising Total Business Survey Burden.” *The Survey Methodology Bulletin* 55: 1–10. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/survey-methodology-bulletin/smb-55/index.html (accessed April 2013).

Kaufman, L. and P.J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Lorenc, B., M. Bavdaž, D. Giesen, R. Seljak, and V. Torres van Grinsven. 2012. "Businesses as users of official statistics." In Proceedings of the Fourth International Conference on Establishment Surveys: American Statistical Association, June 11–14, 2012, Montréal, Canada. Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/meetings/ices/2012/papers/302173.pdf (accessed April 2013).

Lorenc, B, W. Kloek, L. Abrahamson, and S. Eckman. 2013. "An Analysis of Business Response Burden and Response Behaviour Using a Register of Data Provision." In Proceedings of the Conferences on New Techniques and Technologies for Statistics, March 5–7, 2013, Brussels. Available at: http://www.cros-portal.eu/content/ntts-2013-proceedings (accessed April 2013).

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik, M. Studer, and P. Roudier. 2014. *Cluster: Cluster Analysis Basics and Extensions*. R package version 1.15.3.

Merrington, R., B. Torrey, and L. van Heerden. 2009. "Measurement of Respondent Load at Statistics New Zealand." In Proceedings of the 57th session of the International Statistical Institute, August 16–22, 2009, Durban, South Africa. Available at: http://www.statssa.gov.za/isi2009/ScientificProgramme/IPMS/0246.pdf (accessed April 2013).

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, and C.-C. Lin. 2014. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6–4.

OECD. 2009. *Mandate of the Regulatory Policy Committee. A New Agenda for the Regulatory Policy Committee: Issues for the Next Three years, 2010-12*. Resolution of the Council concerning the Mandate of the Regulatory Policy Committee 11 December 2009. Available at: http://www.oecd.org/fr/gov/politique-reglementaire/44679685.pdf (accessed April 2013).

Office of Information and Regulatory Affairs. 2006. *Questions and Answers when Designing Surveys for Information Collections*. Washington: Office of Management and Budget. Available at: http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/pmc_survey_guidance_2006.pdf (accessed January 2013).

Ojo, O. E. and C. Ponikowski. 2010. "Evaluating the Effect of Dependent Sampling on National Compensation Survey Earnings Estimates." In Proceedings of the Section on Survey Research Methods: American Statistical Association. July 31–August 5, 2010, Vancouver, British Colombia, 2766–2775. Alexandria, VA: American Statistical Association. Available at: https://www.amstat.org/sections/SRMS/Proceedings (accessed April 2013).

Paper Work Reduction Act of 1980, United States Federal Law. Title 44, Section 35, United States Code. Available at: http://www.archives.gov/federal-register/laws/paperwork-reduction (accessed April 2013).

R Core Team. 2014. R: A Language and Environment for Statistical Computing (version 2.13.1). Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org (accessed July 2011).

Red Tape Reduction Commission. 2012. Recommendations Report. Cutting Red Tape. Freeing Businesses to Grow. President of the Treasury Board, 2012. Available at: http://www.reduceredtape.gc.ca/heard-entendu/rr/rr-eng.pdf (accessed April 2013).

Sear, J. 2011. "Response Burden Measurement and Motivation at Statistics Canada." In Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys, date of conference, 151–160. Heerlen: Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/23FD3DF5-6696-4A04-B8EF-1FAACEAD995C/0/2011proceedingsblueets.pdf (accessed April 2013).

SCM Network. 2005. *International Standard Cost Model Manual: Measuring and Reducing Administrative Burdens for Businesses*. Available at: http://www.administrative-burdens.com/filesystem/2005/11/international_scm_manual_final_178.doc (accessed October 2014).

Statistics Belgium. 2010. "Administrative Simplification Structural Business Statistics Survey." In Proceedings of the SIMPLY 2010 Conference, December 2–3, 2010, Ghent. Available at: http://www.simply2010.be/documents/papers/SESSION_2b_P5_BE.doc (accessed April 2013).

Statistics New Zealand. 2008. *Respondent Load Strategy for Statistics New Zealand: Strategies and Initiatives for Reducing Respondent Load*. Wellington: Statistics New Zealand. Available at: http://www.stats.govt.nz/~/media/Statistics/about-us/policies-protocols-guidelines/respondent-load-strategy/respondent_load_strategy.pdf (accessed April 2013).

Sunter, A. 1977. "Response Burden, Sample Rotation, and Classification Renewal in Economic Surveys." *International Statistical Review* 45: 209–222. DOI: http://dx.doi.org/10.2307/1402535.

Triantafillou, P. 2007. "Benchmarking in the Public Sector: A Critical Conceptual Framework." *Public Administration* 85: 829–846. DOI: http://dx.doi.org/10.1111/j.1467-9299.2007.00669.x.

United Nations. 1994. *Fundamental Principles of Official Statistics*. Available at: http://unstats.un.org/unsd/dnss/gp/fp-english.pdf (accessed October 2014).

United Nations. 2014. *Fundamental Principles of Official Statistics*. Available at: http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf (accessed October 2014).

Ward, J. H. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58: 236–244.

Willeboordse, A. 1997. "Minimizing Response Burden." In *Handbook on Design and Implementation of Business Surveys*, edited by A. Willeboordse. 111–118. Luxembourg: Eurostat. Available at: http://ec.europa.eu/eurostat/ramon/statmanuals/files/Handbook%20on%20surveys.pdf (accessed April 2013).

Received November 2013
Revised May 2015
Accepted May 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 589–609, http://dx.doi.org/10.1515/JOS-2015-0036

Statistical Estimators Using Jointly Administrative and Survey Data to Produce French Structural Business Statistics

Philippe Brion[1] *and Emmanuel Gros*[2]

Using as much administrative data as possible is a general trend among most national statistical institutes. Different kinds of administrative sources, from tax authorities or other administrative bodies, are very helpful material in the production of business statistics. However, these sources often have to be completed by information collected through statistical surveys. This article describes the way Insee has implemented such a strategy in order to produce French structural business statistics. The originality of the French procedure is that administrative and survey variables are used jointly for the same enterprises, unlike the majority of multisource systems, in which the two kinds of sources generally complement each other for different categories of units. The idea is to use, as much as possible, the richness of the administrative sources combined with the timeliness of a survey, even if the latter is conducted only on a sample of enterprises. One main issue is the classification of enterprises within the NACE nomenclature, which is a cornerstone variable in producing the breakdown of the results by industry. At a given date, two values of the corresponding code may coexist: the value of the register, not necessarily up to date, and the value resulting from the data collected via the survey, but only from a sample of enterprises. Using all this information together requires the implementation of specific statistical estimators combining some properties of the difference estimators with calibration techniques. This article presents these estimators, as well as their statistical properties, and compares them with those of other methods.

Key words: Structural business statistics; administrative data; multisources device.

1. Introduction

Using administrative data to produce official statistics is a big challenge for National Statistical Institutes (NSIs). Concerning business statistics, a lot of administrative sources are often available, and NSIs are using them more and more in an intensive way. A European ESSnet has been working on finding common ways for their use. However, an information collection carried out in 2009–2010 about existing practices among NSIs shows that various contexts do exist, especially concerning the legal basis underlying the use of administrative data, and the cooperation with administrative data holders (Costanzo 2011).

If we now consider the case of structural business statistics, the strategies of the different NSIs vary greatly, from the simple use of statistical surveys (without any use of

[1] INSEE Boulevard Adolphe Pinard F-75675 Paris Cedex 14 75675, France. Email: philippe.brion@insee.fr
[2] INSEE Department of Statistical Methodology, 18, Boulevard Adolphe Pinard, F-75675 Paris Cedex 14 75675, France. Email: emmanuel.gros@insee.fr

administrative sources) to the complete replacement of survey data with administrative sources. In between, a lot of NSIs use intermediate systems, combining administrative and survey data.

This article describes the French strategy adopted by Insee (National Institute of Statistics and Economic Studies) in order to build a new process of producing structural business statistics. As mentioned in Costanzo (2011), concerning the use of administrative data for business statistics, France is considered to have a specific model, highly centralized, due to a business register (SIRENE) that serves both administrative and statistical purposes. This model makes the use of administrative data concerning enterprises easier than in other countries, particularly due to the fact that each administration uses the same unit and the same ID number for the enterprises, which is the SIRENE ID number.

France has been using tax files to produce structural business statistics for a long time (Grandjean 1997). The richness of these files, composed of annual income statements sent by enterprises to the tax authorities, is very interesting, since the files provide detailed information about the accounting characteristics of all French businesses. However, for a long time, these files were available too late to answer certain needs, such as the supplying of preliminary results before the end of October of year ($n + 1$) for the European Structural Business Statistics (SBS) regulation. Furthermore, they did not provide information for all kinds of needs. Thus a statistical survey, limited to a sample of enterprises, was conducted at the same time: this statistical survey was the basis for the preliminary results sent to Eurostat, as the administrative data were used for the definitive results sent later.

This double system had a significant drawback, however: the two sources sometimes told different stories, even at a highly aggregated level. Using two different sources led, obviously, to the possibility of conflicting results. Here, one of the most important reasons identified related to the classification of enterprises within the NACE nomenclature. The two sources do not obtain the same quality of information for this variable (see Subsections 2.1 and 2.2 below), tax files being mainly based on the value of the code within the register, which cannot be updated in a continuous way for all enterprises. Since the results by industry are very important for structural business statistics, the divergences of the two systems were particularly problematic.

Hence a new system of production of French Structural Business Statistics, named ESANE (as *Elaboration des Statistiques ANnuelles d'Entreprises*), has been implemented to unite the two previous systems in just one, taking advantage of each of their characteristics (Brion 2011).

The originality of this device is that within it, variables obtained in the two sources (administrative files, statistical survey) are used jointly for the same enterprises, especially for classifying them within the NACE. By contrast, in many other systems, at least in European countries (ESSnet on administrative data 2011), the two sources generally complement one another for different categories of enterprises (for example, the statistical survey being limited to large enterprises, and the administrative data used for small and medium units).

This article is mainly dedicated to the questions of statistical estimators used in the device. The next section of the article provides a quick overview of the system. The following section is dedicated to the characteristics of the estimators that have been implemented. In Section 4, some other aspects of the system are mentioned briefly.

2. The French System of Structural Business Statistics

2.1. An Intensive Use of Administrative Data Combined With a Survey

The French system is mainly based on two administrative sources, completed by a survey. It is based upon a central administrative source: the annual statements of benefits sent by enterprises to the tax authorities (Chami 2010), containing accounting variables (between 500 and 1,000 according to the size of the enterprise). It should be noted that French statistical law makes Insee's access to these files possible. This material is very rich, since it concerns every unit of the three millions of enterprises under the scope of business statistics. Of course it cannot be used directly, mainly for two reasons:

- it has to be checked, because of missing data, or of multiple declarations: hence work is done by Insee to impute missing data (Deroyon 2013) and to deal with multiple declarations,
- not all information needed to produce the structural business statistics is available in these files, and additional information has to be obtained elsewhere.

A second interesting source is composed of the annual social security returns of the enterprises to the administration, giving information about employees and wages.

Using these two sources helps lessen the statistical burden on enterprises, but some additional information has to be collected to answer some of the users' needs. This is done through a statistical survey, because the required information is not available in administrative files. One cornerstone variable in particular is obtained thanks to the survey: the detailed breakdown of the enterprise's turnover according to its different activities. This information, among others, is needed at a very detailed level for the national accounts. Since only a "rough breakdown" – between production, sales and services – of the total turnover of the enterprise is available in the tax files, one main part of the statistical survey questionnaire is dedicated to this question: enterprises are asked to fill out a table giving the value of the turnover of each industry they are performing.

Other variables are collected through the survey, concerning restructuring of enterprises, data about nonsalaries, and other specific topics related to the economic sectors (relative to professional expenses, or to other specific aspects such as, for example, the number of trucks for road transportation). This survey is limited to a sample of enterprises (Haag 2010).

2.2. The Business Register and the Classifying of Each Enterprise Using the Nomenclature of Activities

As mentioned above, the French business register, SIRENE, serves both administrative and statistical purposes. The use of its ID number is mandatory for each French administration, and this makes the use of administrative files for statistical purposes very easy. In this way, there is no problem of undercoverage of the register.

Every French enterprise has, within SIRENE, a "principal activity code" named APE (in French *Activité Principale de l'Entreprise*), classifying it within the French NAF nomenclature of activities, which is derived from the European NACE. In this article, this value is named APE_{reg}. At the time of the creation of the enterprise, this value is coded by SIRENE clerks, according to the firm declaration.

However, this value is not necessarily updated in a continuous way for all French enterprises, especially for the numerous small ones. Some enterprises send information to modify the value of this code, but this is not the case for all of them. So directly using the value available in the register for producing statistics may raise quality-related questions: economic sectors are changing, for example, during the last years some enterprises have been moving from industry to the trade sector. The statistics that could be produced directly using the values of the code within the business register would not properly represent these changes.

Through the statistical survey, we obtain updated and rather objective information on the different activities conducted by the surveyed enterprises: each enterprise fills out a table giving a breakdown of its turnover according to the different activities it is performing, and an algorithm is then used to calculate an updated value of the APE code (using the breakdown of the turnover by activities as a proxy for the breakdown of value added of these activities, which should be, from a theoretical point of view, the basic information to classify the enterprise). This updated value, referred to in this article as $\text{APE}_{\text{survey}}$, may differ from the initial value of the register, and is only available for some of the enterprises, namely those that are surveyed. In the end, it is introduced into the business register, and may be used for the next drawing of samples; however, it cannot be fed back into the register and then used directly in the current survey as an auxiliary variable for statistical purposes (for example for calibration), since the partial updating of the register would lead to some bias.

2.3. *An Original Kind of Database*

Using administrative and survey data jointly leads, in a simplified presentation, to an incomplete rectangular data base (Figure 1). In this figure, rows represent enterprises and columns variables. The right part contains variables obtained through the administrative sources (mainly accounting variables), and the left part variables obtained through the statistical survey. This survey uses a sample stratified according to the activity and the size of the enterprises. The stratification variable used for the activity is obviously APE_{reg}, and the size is based on the number of employees. The sampling rates are different according to the size of the enterprise, and the take-all stratum (generally defined as more than 20 employees) contains the largest enterprises. The white area dominating the left part represents unobserved data (since in the sampled stratum only 85,000 enterprises are surveyed from the population of almost three million units).

This data base, where sampling weights exist for the left part only, is not easy to use, compared to an administrative data base (without sampling weights) or to a survey data base (with data limited to the sampled units, with sampling weights).

It should be noted concerning the classifying of the enterprise within the nomenclature of activities that two values may coexist in the database: the value of the register APE_{reg}, available for all three million enterprises, and the value of the updated $\text{APE}_{\text{survey}}$, which exists only for the units of the survey.

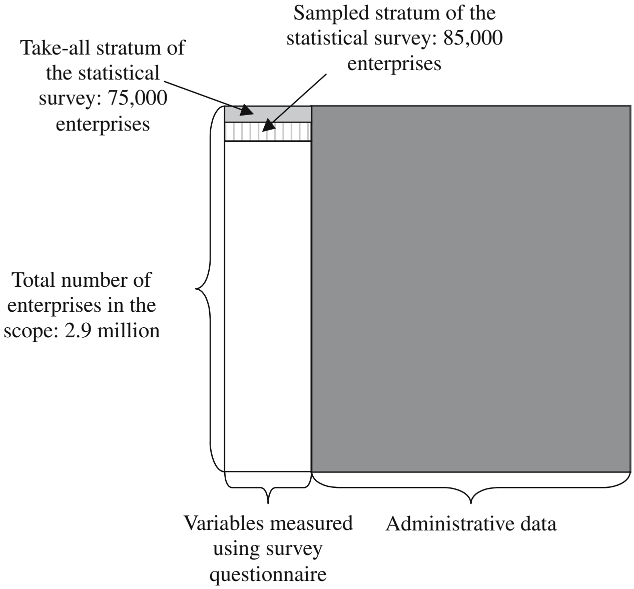

Sampled stratum of the statistical survey: 85,000 enterprises
Take-all stratum of the statistical survey: 75,000 enterprises
Total number of enterprises in the scope: 2.9 million
Variables measured using survey questionnaire
Administrative data

Fig. 1. ESANE, the multisources device for the French business statistics

3. The Statistical Estimators Used to Produce the Structural Business Statistics

3.1. What Kinds of Statistics Do We Want to Produce?

Structural business statistics have to give an appropriate picture of the population of enterprises, mainly concerning accounting variables (such as the turnover, the value added, the investments, etc.), but also characterizing enterprises by the industry to which they belong.

In this way, many of the produced statistics do not result from one variable only, but from a combination of two (or more) variables: a quantitative variable combined with a qualitative variable.

For example, if we consider the total turnover of an economic sector A, the quantity to estimate is:

$$\sum_{i \in U} \mathrm{Turnover}(i) 1\mathrm{I}_{\mathrm{APE=A}}(i),$$

where $1\mathrm{I}_{\mathrm{APE=A}}(i)$ is the indicator variable relative to the classifying of enterprise *i* in industry *A* (or sector: in this article we use sometimes the wording sector, understood as economic sector, not institutional sector referring to the system of national accounts), and *U* is the global population of enterprises.

The variable "turnover" is available in the tax files, while for the activity code the survey provides fresher and richer information than the register (even if a value does exist in the register).

Other kind of statistics are produced, for example statistics based only on survey variables but the following sections of the article focus mainly on the multisource statistics presented above, since sector-based statistics are one of the main results of the device.

3.2. Different Possible Methods

The objective is to rely, as much as possible, on the exhaustiveness of administrative sources, which concern hundreds of variables. This material has to be used jointly with the information available in the statistical survey, conducted on the sample of enterprises, particularly the up-to-date activity code.

Two "families" of methods may be considered:

- mass imputation (Kovar and Withbridge 1995), taking into account the observations of the sample to generate values for the white part of the rectangle of Figure 1; in particular, it is necessary to generate an updated APE code for each enterprise of the population (that means approximately three million enterprises),
- inference using specific statistical estimates.

Methodological studies have been conducted to compare the two kinds of methods (Brion 2007, performed on past data in NACErev1. It should be noted that Kroese and Renssen (2000) present some elements on the mass imputation method that are similar to those of (Brion 2007)). More precisely, the imputation method that has been evaluated consisted of imputing an updated value of the APE code for the nonsampled units by using probabilities of moving from the economic sector in which the enterprise is classified within the register to another sector, these probabilities being estimated on the sample for categories belonging to the same "four-digit" level within the register.

The first thing to note is that the mass imputation method leads to some potential bias in the way it is proposed here. The methodological studies have quantified the value of this bias as far from negligible: more precisely, for the trade sector, composed of 119 different values of the code APE, 15 have a potential bias with the proposed imputation method that is more, in absolute value, than ten percent of the total to estimate.

Then, in order to compare the mass imputation method with other estimates, the mean square error of every method needs to be computed: for the mass imputation method, its variance needs to be evaluated and to be added to the square of the bias that has been evaluated previously. The mass imputation method is compared to a difference estimator, which is unbiased, and close to the final estimators used in ESANE that are presented in next section (Brion 2007). Results show that, for the global trade sector, the root mean square error of the difference estimator is approximately half of the root mean square error of the mass imputation estimator. A comparison at a lower level of the nomenclature (four digits of the NACE) has been found that for 13 classes mass imputation was better, as for 100 classes the difference estimator was better. For this reason, it was decided to abandon the idea of using the mass imputation method. However, the question of the different kinds of methods to use to produce official statistics remains open (see for example Little 2012 and Brion 2012a).

Then, concerning "classic" statistical estimates, two usual strategies may be considered:

1. Only using the data coming from the units in the sample (and taking into account both survey and administrative variables for these units). This is a minimum approach, because it does not exploit the exhaustiveness of the administrative sources, which is consequently unsatisfactory.

2. Using calibration techniques (Deville and Särndal 1992) to improve the efficiency of the estimators. Here, the exhaustiveness of the administrative sources is used to modify the sampling weights according to calibration equations involving some of the administrative variables. This approach, which is an extension of the general regression estimator, will lead to a better precision of estimates for variables linked to the calibration variables.

In theory, this strategy allows us to take into account all information available in the administrative sources by computing a calibration estimator which takes into account all administrative variables. However, the huge number of fiscal variables (over 500) makes this approach totally impracticable: the calibration procedure (if it converges, which is clearly not guaranteed: indeed, with so many variables, even the regression estimator, which is a special case of calibration estimator, can be incomputable, due to colinearity problems for example) will lead to some negative and/or overly extreme weights, which will induce unrealistic sector-based estimates for some economic sectors, especially at a detailed level.

To avoid these problems, another strategy could be to take into account the information available in the administrative sources "variable by variable" and "sector by sector", by computing a simple regression estimator for each fiscal variable and each sector.

For a fiscal variable *Z* and a sector *A*, such a sector-based estimator would be:

$$\begin{aligned}\hat{\mathrm{Z}}^{\mathrm{A}}_{\mathrm{reg}} &= \sum_{\mathrm{s}} \frac{\mathrm{Z_i}\, 1\mathrm{I}_{\mathrm{APEsurvey=A}}(\mathrm{i})}{\pi_{\mathrm{i}}} + \hat{\beta}_{\mathrm{Z,A}} \left[\sum_{\mathrm{U}} \mathrm{Z_i}\, 1\mathrm{I}_{\mathrm{APEreg=A}}(\mathrm{i}) - \sum_{\mathrm{s}} \frac{\mathrm{Z_i}\, 1\mathrm{I}_{\mathrm{APEreg=A}}(\mathrm{i})}{\pi_{\mathrm{i}}} \right] \\ &= \hat{\mathrm{Y}}_{\pi} + \hat{\beta}_{\mathrm{Z,A}} \left[\mathrm{X} - \hat{\mathrm{X}}_{\pi} \right] \qquad (1)\end{aligned}$$

with $\mathrm{Y_i} = \mathrm{Z_i}\, 1\mathrm{I}_{\mathrm{APEsurvey=A}}(\mathrm{i})$ and $\mathrm{X_i} = \mathrm{Z_i}\, 1\mathrm{I}_{\mathrm{APEreg=A}}(\mathrm{i})$;

π_i is the inclusion probability of unit *i*;

$1\mathrm{I}_{\mathrm{APEreg=A}}(\mathrm{i})$ is the indicator variable using the value of the APE code within the register;

$1\mathrm{I}_{\mathrm{APEsurvey=A}}(\mathrm{i})$ is the indicator variable using the value of the APE code obtained through the statistical survey;

$\hat{\beta}_{\mathrm{Z,A}}$ is the coefficient of the simple regression of *Y* on *X* (the subscript *Z,A* reminds us that this coefficient depends at the same time on the fiscal variable Z and on the sector *A*).

It should also be noted that this regression estimator can also be formulated as a weighting estimator with weights $\mathrm{w_i^{Z,A}}$ that are different for each fiscal variable and each sector. This differs from the "classical" calibration estimator that leads to a single weight for each sampling unit regardless of the fiscal variable or sector.

Such an estimator allows us to produce sector-based estimates for all fiscal variables and all "sectoral" levels by systematically taking into account the exhaustiveness of the administrative sources. Unfortunately, such an approach is not appropriate to the context of ESANE. And consequently, these regression estimators are not used in the final system. Indeed, statistics produced in the ESANE device are subject to many consistency constraints, both “vertical” – consistency between estimations concerning different levels of hierarchically nested nomenclature – and “horizontal” – consistency between estimations relating to variables linked by accounting relationships – that the estimation method has to respect. However, the approach detailed above is not linear, because the $\hat{\beta}_{\mathrm{Z,A}}$ coefficients – or the weights $\mathrm{w}_{\mathrm{i}}^{\mathrm{Z,A}}$ if the estimator is formulated as a weighting estimator – change with the fiscal variable Z and the sector A, and consequently this approach does not ensure that the consistency constraints in the ESANE method would be respected.

Let us take, for example, three fiscal variables U, V and W linked by the accounting relationship $W = U + V$ and a group G of the NACE Rev.2 divided into two classes G1 and G2. We can compute the three sector-based estimators according to Formula (1):

$$\hat{\mathrm{U}}_{\mathrm{reg}}^{\mathrm{G}} = \sum_{\mathrm{s}} \frac{\mathrm{U_i}\, 1\mathrm{I}_{\mathrm{Group\_survey=G}}(\mathrm{i})}{\pi_{\mathrm{i}}} + \hat{\beta}_{\mathrm{U,G}} \left[\sum_{\mathrm{U}} \mathrm{U_i}\, 1\mathrm{I}_{\mathrm{group\_reg=G}}(\mathrm{i}) - \sum_{\mathrm{s}} \frac{\mathrm{U_i}\, 1\mathrm{I}_{\mathrm{group\_reg=G}}(\mathrm{i})}{\pi_{\mathrm{i}}} \right]$$

$$\hat{\mathrm{V}}_{\mathrm{reg}}^{\mathrm{G}} = \sum_{\mathrm{s}} \frac{\mathrm{V_i}\, 1\mathrm{I}_{\mathrm{Group\_survey=G}}(\mathrm{i})}{\pi_{\mathrm{i}}} + \hat{\beta}_{\mathrm{V,G}} \left[\sum_{\mathrm{U}} \mathrm{V_i}\, 1\mathrm{I}_{\mathrm{group\_reg=G}}(\mathrm{i}) - \sum_{\mathrm{s}} \frac{\mathrm{V_i}\, 1\mathrm{I}_{\mathrm{group\_reg=G}}(\mathrm{i})}{\pi_{\mathrm{i}}} \right]$$

$$\hat{\mathrm{W}}_{\mathrm{reg}}^{\mathrm{G}} = \sum_{\mathrm{s}} \frac{\mathrm{W_i}\, 1\mathrm{I}_{\mathrm{Group\_survey=G}}(\mathrm{i})}{\pi_{\mathrm{i}}} + \hat{\beta}_{\mathrm{W,G}} \left[\sum_{\mathrm{U}} \mathrm{W_i}\, 1\mathrm{I}_{\mathrm{group\_reg=G}}(\mathrm{i}) - \sum_{\mathrm{s}} \frac{\mathrm{W_i}\, 1\mathrm{I}_{\mathrm{group\_reg=G}}(\mathrm{i})}{\pi_{\mathrm{i}}} \right]$$

But as $\hat{\beta}_{\mathrm{U,G}} \neq \hat{\beta}_{\mathrm{V,G}} \neq \hat{\beta}_{\mathrm{W,G}}$, $\hat{\mathrm{W}}_{\mathrm{reg}}^{\mathrm{G}}$ is not equal to $\hat{\mathrm{U}}_{\mathrm{reg}}^{\mathrm{G}} + \hat{\mathrm{V}}_{\mathrm{reg}}^{\mathrm{G}}$, even if we have $W_i = U_{\mathrm{i}} + V_i$ for each unit i.

And in the same way, we can compute for variable U the sector-based estimator for sectors G, G1 and G2:

$$\hat{\mathrm{U}}_{\mathrm{reg}}^{\mathrm{G}} = \sum_{\mathrm{s}} \frac{\mathrm{U_i}\, 1\mathrm{I}_{\mathrm{Group\_survey=G}}(\mathrm{i})}{\pi_{\mathrm{i}}} + \hat{\beta}_{\mathrm{U,G}} \left[\sum_{\mathrm{U}} \mathrm{U_i}\, 1\mathrm{I}_{\mathrm{group\_reg=G}}(\mathrm{i}) - \sum_{\mathrm{s}} \frac{\mathrm{U_i}\, 1\mathrm{I}_{\mathrm{group\_reg=G}}(\mathrm{i})}{\pi_{\mathrm{i}}} \right]$$

$$\hat{\mathrm{U}}_{\mathrm{reg}}^{\mathrm{G1}} = \sum_{\mathrm{s}} \frac{\mathrm{U_i}\, 1\mathrm{I}_{\mathrm{Class\_survey=G1}}(\mathrm{i})}{\pi_{\mathrm{i}}} + \hat{\beta}_{\mathrm{U,G1}} \left[\sum_{\mathrm{U}} \mathrm{U_i}\, 1\mathrm{I}_{\mathrm{Class\_reg=G1}}(\mathrm{i}) - \sum_{\mathrm{s}} \frac{\mathrm{U_i}\, 1\mathrm{I}_{\mathrm{Class\_reg1=G2}}(\mathrm{i})}{\pi_{\mathrm{i}}} \right]$$

$$\hat{\mathrm{U}}_{\mathrm{reg}}^{\mathrm{G2}} = \sum_{\mathrm{s}} \frac{\mathrm{U_i}\, 1\mathrm{I}_{\mathrm{Class\_survey=G2}}(\mathrm{i})}{\pi_{\mathrm{i}}} + \hat{\beta}_{\mathrm{U,G2}} \left[\sum_{\mathrm{U}} \mathrm{U_i}\, 1\mathrm{I}_{\mathrm{Class\_reg=G2}}(\mathrm{i}) - \sum_{\mathrm{s}} \frac{\mathrm{U_i}\, 1\mathrm{I}_{\mathrm{Class\_reg=G2}}(\mathrm{i})}{\pi_{\mathrm{i}}} \right]$$

The same causes produce the same effects, as $\hat{\beta}_{\mathrm{U,G}} \neq \hat{\beta}_{\mathrm{U,G1}} \neq \hat{\beta}_{\mathrm{U,G2}}$, $\hat{\mathrm{U}}_{\mathrm{reg}}^{\mathrm{G}}$ is not equal to $\hat{\mathrm{U}}_{\mathrm{reg}}^{\mathrm{G1}} + \hat{\mathrm{U}}_{\mathrm{reg}}^{\mathrm{G2}}$.

Another strategy is hence proposed: using combined statistical estimates mixing the principles of the difference estimators (Särndal et al. 1992) and the calibration techniques. This third option is detailed in the next section of the article.

3.3. The Statistical Estimators for Sector-Based Estimates at the Group (and Upper) Level

The idea is to start from the standard Horvitz-Thompson estimator and to use the exhaustiveness of the administrative sources to improve its efficiency as much as possible while keeping to all the consistency constraints of the ESANE device. In practice, as we have to deal with unit nonresponse, the "starting point" is in fact not the Horvitz-Thompson estimator but the reweighted-expansion estimator, with weights adjusted for unit nonresponse thanks to the response homogeneity groups method RHG.

First, as the turnover is a core variable – highly correlated with both turnover breakdown and the main accounting variables of the device such as value added –, we can use calibration techniques to modify the RHG-adjusted weights according to calibration equations involving turnover by sector. More precisely, the equations used here are:

$$\begin{cases} \sum_{i\in R} w_i T(i) 1I_{APEreg=A}(i) = \sum_{i\in U} T(i) 1I_{APEreg=A}(i) \\ \sum_{i\in R} w_i 1I_{APEreg=A}(i) = \sum_{i\in U} 1I_{APEreg=A}(i) \end{cases}$$

where:

- w_i is the calibrated weight of each enterprise i of the sample of respondents *R*,
- $1I_{APEreg=A}(i)$ is the indicator variable using the value of the APE code within the register,
- $T(i)$ is the value of the turnover of enterprise i in the tax files.

That is, we perform calibration on the total turnover and the number of enterprises by sector for each sector A of the ESANE device. In practice, this calibration is generally performed at the "3-digits" level of the sectoral classification, in order to limit the range of changes of the weights.

The calibration on the sectoral total of turnover permits us to improve the accuracy of sector-based estimates for all variables correlated with the turnover, while the calibration on the number of enterprises by sector aims to avoid too much distortion concerning the estimation of numbers of enterprises by sectors.

This calibration estimator thus incorporates all information available in the tax sources for the turnover variable, but, as previously stated, it does not allow the exhaustiveness of the administrative sources to be taken into account for other variables. In order to compensate for this drawback, we can use the principle of difference estimation and consider the following "combined estimator" for sector-based estimates relating to any administrative variable *Z*, such as turnover, value added, investments and so on:

$$\hat{Z}^{A}_{diff} = \sum_{i\in R} w_i Z_i 1I_{APEsurvey=A}(i) + \sum_{i\in U} Z_i 1I_{APEreg=A}(i) - \sum_{i\in R} w_i Z_i 1I_{APEreg=A}(i) \quad (2)$$

This estimator, based on the existence of two APE codes – the one of the register (APE_{reg}), available for all units, and the one derived from the survey (APE_{survey}), known only for the sample –, allows us to use all information available in the administrative sources for the variable *Z* while keeping to all the linear consistency constraints of the ESANE device because of its linearity.

Indeed, if we consider again the example of three fiscal variables *U,V* and *W* linked by the accounting relationship $W = U + V$ and a group G of the NACE Rev.2 divided into two classes G1 and G2, we can compute the three sector-based estimators according to Formula (2):

$$\hat{U}^{G}_{diff} = \sum_{i \in R} w_i U_i 1I_{Group\_survey=G}(i) + \sum_{i \in U} U_i 1I_{Group\_reg=G}(i) - \sum_{i \in R} w_i U_i 1I_{Group\_reg=G}(i)$$

$$\hat{V}^{G}_{diff} = \sum_{i \in R} w_i V_i 1I_{Group\_survey=G}(i) + \sum_{i \in U} V_i 1I_{Group\_reg=G}(i) - \sum_{i \in R} w_i V_i 1I_{Group\_reg=G}(i)$$

$$\hat{W}^{G}_{diff} = \sum_{i \in R} w_i W_i 1I_{Group\_survey=G}(i) + \sum_{i \in U} W_i 1I_{Group\_reg=G}(i) - \sum_{i \in R} w_i W_i 1I_{Group\_reg=G}(i)$$

and we have:

$$\begin{aligned}
\hat{U}^{G}_{diff} + \hat{V}^{G}_{diff} &= \sum_{i \in R} w_i U_i 1I_{Group\_survey=G}(i) + \sum_{i \in R} w_i V_i 1I_{Group\_survey=G}(i) \\
&\quad + \sum_{i \in U} U_i 1I_{Group\_reg=G}(i) + \sum_{i \in U} V_i 1I_{Group\_reg=G}(i) \\
&\quad - \left[\sum_{i \in R} w_i U_i 1I_{Group\_reg=G}(i) + \sum_{i \in R} w_i V_i 1I_{Group\_reg=G}(i) \right] \\
&= \sum_{i \in R} w_i \underbrace{(U_i + V_i)}_{W_i} 1I_{Group\_survey=G}(i) + \sum_{i \in U} \underbrace{(U_i + V_i)}_{W_i} 1I_{Group\_reg=G}(i) \\
&\quad - \sum_{i \in R} w_i \underbrace{(U_i + V_i)}_{W_i} 1I_{Group\_reg=G}(i) \\
&= \hat{W}^{G}_{diff}
\end{aligned}$$

In the same way, for variable *U* we can compute the sector-based estimator for Sector G, G1 and G2:

$$\hat{U}^{G}_{diff} = \sum_{i \in R} w_i U_i 1I_{Group\_survey=G}(i) + \sum_{i \in U} U_i 1I_{Group\_reg=G}(i) - \sum_{i \in R} w_i U_i 1I_{Group\_reg=G}(i)$$

$$\hat{U}^{G1}_{diff} = \sum_{i \in R} w_i U_i 1I_{Class\_survey=G1}(i) + \sum_{i \in U} U_i 1I_{Class\_reg=G1}(i) - \sum_{i \in R} w_i U_i 1I_{Class\_reg=G1}(i)$$

$$\hat{U}^{G2}_{diff} = \sum_{i \in R} w_i U_i 1I_{Class\_survey=G2}(i) + \sum_{i \in U} U_i 1I_{Class\_reg=G2}(i) - \sum_{i \in R} w_i U_i 1I_{Class\_reg=G2}(i)$$

and we have:

$$\begin{aligned}\hat{\mathrm{U}}_{\mathrm{diff}}^{\mathrm{G1}}+\hat{\mathrm{U}}_{\mathrm{diff}}^{\mathrm{G2}}=&\sum_{\mathrm{i}\in \mathrm{R}}\mathrm{w}_{\mathrm{i}}\mathrm{U}_{\mathrm{i}}\underbrace{\left(1\mathrm{I}_{\mathrm{Class\_survey=G1}}(\mathrm{i})+1\mathrm{I}_{\mathrm{Class\_survey=G2}}(\mathrm{i})\right)}_{1\mathrm{I}_{\mathrm{Group\_survey=G}}(\mathrm{i})}\\&+\sum_{\mathrm{i}\in \mathrm{U}}\mathrm{U}_{\mathrm{i}}\underbrace{\left(1\mathrm{I}_{\mathrm{Class\_reg=G1}}(\mathrm{i})+1\mathrm{I}_{\mathrm{Class\_reg=G2}}(\mathrm{i})\right)}_{1\mathrm{I}_{\mathrm{Group\_reg=G}}(\mathrm{i})}\\&-\sum_{\mathrm{i}\in \mathrm{R}}\mathrm{w}_{\mathrm{i}}\mathrm{U}_{\mathrm{i}}\underbrace{\left(1\mathrm{I}_{\mathrm{Class\_reg=G1}}(\mathrm{i})+1\mathrm{I}_{\mathrm{Class\_reg=G2}}(\mathrm{i})\right)}_{1\mathrm{I}_{\mathrm{Group\_reg=G}}(\mathrm{i})}=\hat{\mathrm{U}}_{\mathrm{diff}}^{\mathrm{G}}\end{aligned}$$

Moreover, as the variables $\mathrm{Z}_{\mathrm{i}}1\mathrm{I}_{\mathrm{APEsurvey=A}}(\mathrm{i})$ and $\mathrm{Z}_{\mathrm{i}}1\mathrm{I}_{\mathrm{APEreg=A}}(\mathrm{i})$ are usually well correlated and indeed often almost identical, this difference estimator is particularly appropriate to the ESANE device, and generally permits us to improve the quality of sector-based estimates.

It should be noted that the principle of difference estimation is used here in an unconventional way: indeed, in the conventional difference estimator (Särndal et al. 1992), the same set of auxiliary variables is used to perform estimation for all variables; conversely, in our combined estimator, the auxiliary variable $\mathrm{Z}_{\mathrm{i}}1\mathrm{I}_{\mathrm{APEreg=A}}(\mathrm{i})$ depends at the same time on the administrative variable *Z* and on the sector *A* and is consequently suited to the considered sector-based estimation.

Let us finally conclude with two comments on the relevance and the impact of calibration in our combined estimator. First, with calibrated weights, the combined estimators coincide with the calibrated estimators at the level of the nomenclature used for the calibration equations for the sector-based estimates relating to variables "turnover" and "number of enterprises". This gives coherence between statistics based on the administrative variables and estimates based on variables available only in the survey – obtained with the calibrated estimator. Finally, the use of calibrated weights in the combined estimator leads to improvements in the accuracy of sector-based estimates when $\mathrm{Z}_{\mathrm{i}}1\mathrm{I}_{\mathrm{APEsurvey=A}}(\mathrm{i})-\mathrm{Z}_{\mathrm{i}}1\mathrm{I}_{\mathrm{APEreg=A}}(\mathrm{i})$ is correlated with $\mathrm{T}_{\mathrm{i}}1\mathrm{I}_{\mathrm{APEreg=A}}(\mathrm{i})$ or $1\mathrm{I}_{\mathrm{APEreg=A}}(\mathrm{i})$.

3.4. A Quantitative Comparison of the Different Methods

In this section, we assess the impact of the methodological improvements implemented in the new system, namely the combined use of calibration techniques and difference estimators, for sector-based estimates at the "three-digit" (and above) level of the NACE Rev.2 classification. For this purpose, we consider the three following estimators:

- the reweighted-expansion estimator, with weights adjusted for unit nonresponse thanks to the response homogeneity groups method (named RHG),
- the calibrated estimator stemming from the calibration step performed in the ESANE device, which is equivalent to the GREG estimator using as auxiliary information the total turnover and the number of enterprises by sector, for each sector at the three-digit level of the sectoral classification (named GREG),
- and the combined estimator described in the previous section (named Esane).

Let us first note that, under the RHG model, these three estimators are unbiased – the reweighted-expansion estimator – or asymptotically unbiased – the GREG estimator and the Esane estimator. Consequently, we focus here on comparing the accuracy of these three estimators, measured by their coefficient of variation (CV).

To compute the coefficients of variation relating to the reweighted-expansion estimator, we use a self-made SAS macro which analytically computes variance, taking into account the stratified sampling design of the survey and the unit nonresponse adjustment using the RHG model.

The coefficients of variation relating to the GREG estimator are obtained by computing the variance of the reweighted-expansion estimator for the total of the residuals derived from the weighted least squares regression of the variable of interest $Y_i 1I_{APEsurvey=A}(i)$ on calibration variables.

Finally, the coefficients of variation relating to the Esane estimator are obtained by computing the variance of the reweighted-expansion estimator for the total of the residuals derived from the weighted least squares regression of the variable of interest $Y_i 1I_{APEsurvey=A}(i) - Y_i 1I_{APEreg=A}(i)$ on calibration variables.

We focus on a small group of core variables of the ESANE device: number of enterprises, turnover, salary, value added, gross operating profit, total assets, total liabilities and gross investments in tangible goods. Table 1 gives the result of this comparison for the six main production sectors covered by the ESANE device.

These results show that, at a global level, the Esane estimator gives better results than the two other estimators. At a more detailed level, the Esane estimator improvement performs better, as shown in Figures 2 and 3. These figures compare the different possible strategies for all variables and main sectors. They show that GREG performs better than RHG, and ESANE generally better than GREG.

But the improvement differs, obviously, depending on the relationship between the studied variable and the variables involved in the calibration procedure, especially with turnover, as Figures 4 and 5 show: for the variable "value added", the calibration step leads to an improvement of the estimators' accuracy, since the value added is positively correlated with the turnover; the "difference estimation" step leads to an another improvement of the estimators' accuracy, of the same order of magnitude as that of the calibration step.

Conversely, for the variable "gross investments in tangible goods", the improvement of the combined estimator is much more important. This is due to the richness of the tax file, which is used in the combined estimator, thanks to the principle of difference estimation, but not in the other methods – since, in the ESANE device, only the turnover and the number of enterprises by sector is used in the calibration equations. The link between the turnover and the investments is relatively weak, compared to the link between the value added and the turnover.

This first global assessment of an improvement of estimates' accuracy due to the combined use of calibration techniques and difference estimators to produce the sector-based estimates in the ESANE device is confirmed by the comparison of sector-based estimates' CV at the three-digit level of the French nomenclature, presented in Figure 6 (Table 2 in the Appendix gives the means and quintiles corresponding to these box plots). Indeed, the new statistical estimators generally lead to an average reduction of the CV, and

Table 1. Comparison between RHG, GREG and Esane estimators’ coefficients of variation (CV) (using Esane’s 2010 data)

	Estimators’s CVs relating to RHG estimators							
Sector	Number of enterprises	Turnover	Salary	Value added	Gross operating profit	Total assets	Total liabilities	Gross investments in tangible goods
Food-processing industry	3.9%	0.5%	1.1%	0.8%	0.7%	0.5%	0.6%	2.3%
Construction	0.9%	1.1%	0.9%	1.4%	5.5%	3.0%	3.2%	10.4%
Trade	1.1%	0.4%	0.5%	0.6%	1.8%	0.5%	0.5%	1.9%
Industry	1.3%	0.1%	0.2%	0.1%	0.4%	0.1%	0.1%	0.9%
Services	0.5%	0.4%	0.4%	0.5%	1.3%	0.9%	1.0%	4.0%
Transport	2.1%	0.4%	0.5%	0.5%	0.9%	1.2%	1.6%	8.6%
Total	0.40%	0.20%	0.21%	0.26%	0.77%	0.45%	0.53%	2.14%

	Estimators’s CVs relating to GREG estimators							
Sector	Number of enterprises	Turnover	Salary	Value added	Gross operating profit	Total assets	Total liabilities	Gross investments in tangible goods
Food-processing industry	3.2%	0.3%	0.6%	0.5%	0.6%	0.3%	0.4%	2.2%
Construction	0.3%	0.5%	0.7%	1.2%	5.6%	3.0%	3.2%	10.7%
Trade	0.5%	0.1%	0.3%	0.4%	1.7%	0.4%	0.4%	1.9%
Industry	1.3%	0.1%	0.1%	0.1%	0.3%	0.1%	0.1%	0.8%
Services	0.3%	0.2%	0.4%	0.4%	1.2%	0.8%	1.0%	4.1%
Transport	0.6%	0.2%	0.3%	0.3%	0.7%	0.5%	0.6%	5.0%
Total	0.09%	0.05%	0.16%	0.20%	0.72%	0.42%	0.49%	1.95%

	Estimators’s CVs in the ESANE device							
Sector	Number of enterprises	Turnover	Salary	Value added	Gross operating profit	Total assets	Total liabilities	Gross investments in tangible goods
Food-processing industry	3.2%	0.3%	0.4%	0.4%	0.4%	0.3%	0.4%	1.0%
Construction	0.3%	0.5%	0.2%	1.2%	6.2%	1.8%	1.8%	3.4%
Trade	0.5%	0.1%	0.2%	0.4%	1.6%	0.6%	0.5%	1.3%
Industry	1.3%	0.1%	0.1%	0.1%	0.3%	0.1%	0.1%	0.1%
Services	0.3%	0.2%	0.1%	0.3%	0.9%	0.4%	0.5%	0.5%
Transport	0.6%	0.2%	0.2%	0.1%	0.3%	0.2%	0.3%	0.3%
Total	0.09%	0.05%	0.01%	0.15%	0.56%	0.17%	0.22%	0.07%

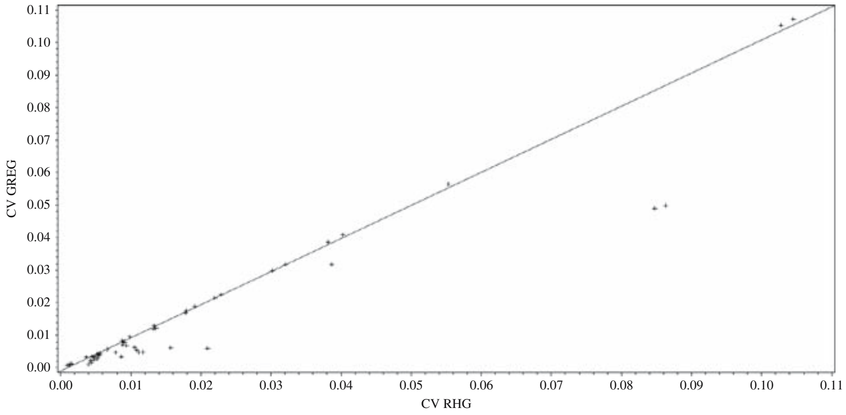


Fig. 2. Comparison of RHG and GREG coefficients of variation, estimations relating to the six main production sectors and the eight variables presented in Table 1

improve the accuracy of estimators in more than 80% of cases. Conversely, for the remaining 20%, the RHG estimator performs better than the ESANE one.

3.5. The Statistical Estimators for Sector-Based Estimates at Finer Levels

As indicated in the previous section, the implemented methods use the richness of the whole administrative data, and correct the problems of misclassifying some units within the registers.

However, these combined estimators have also some limits: particularly, they do not guarantee to always produce positive values, and can consequently lead to negative estimates even if all individual data for the variable of interest are positive. This proves

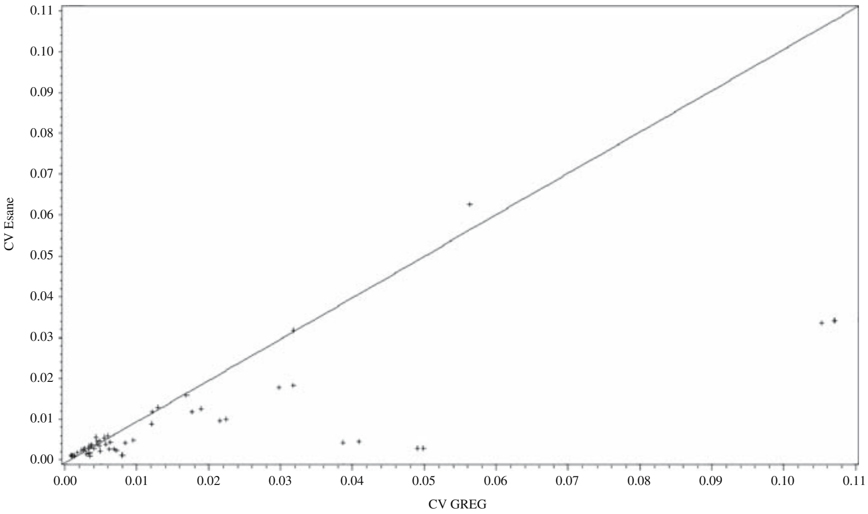


Fig. 3. Comparison of GREG and Esane coefficients of variation, estimations relating to the six main production sectors and the eight variables presented in Table 1

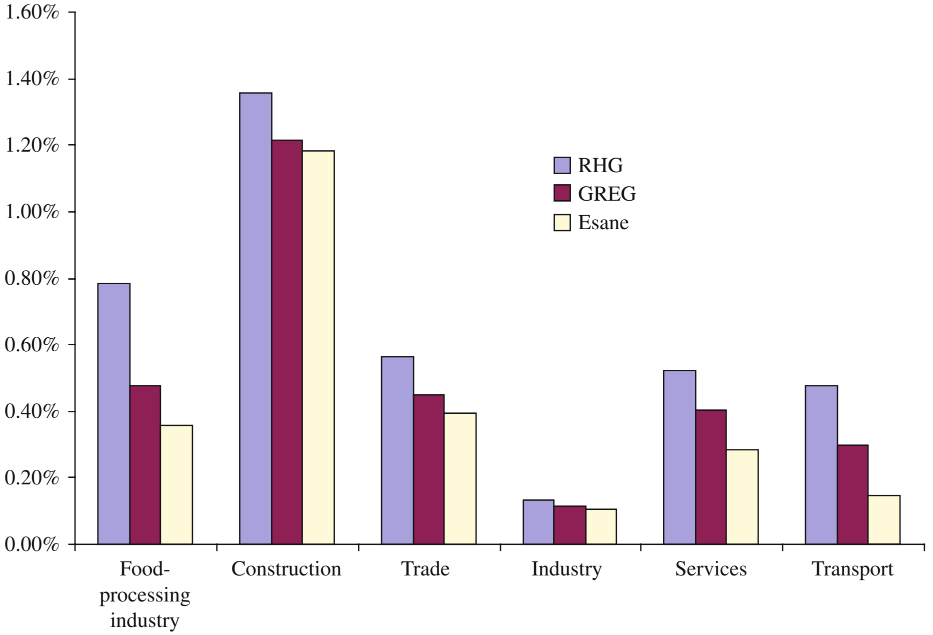


Fig. 4. Coefficients of variation of the three estimators (RHG, GREG, ESANE) for the estimation of the total of the value added by main economic sector

problematic, especially when it concerns variables for which negative aggregates make no economic sense, like turnover or salary.

In practice, this kind of problematic situation appears only when the estimation is relating to too small a domain, either because very few enterprises are concerned by the

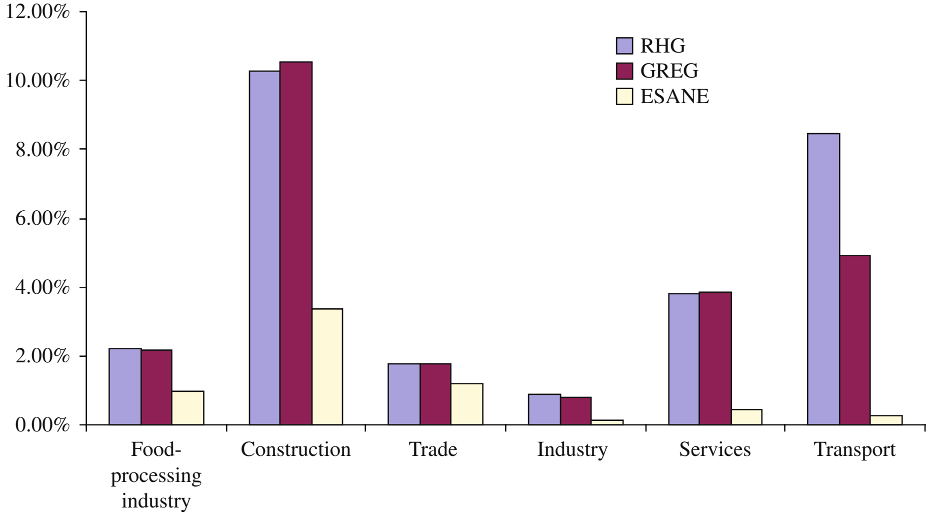


Fig. 5. Coefficients of variation of the three estimators (RHG, GREG, ESANE) for the estimation of the total of the gross investments in tangible goods by main economic sector

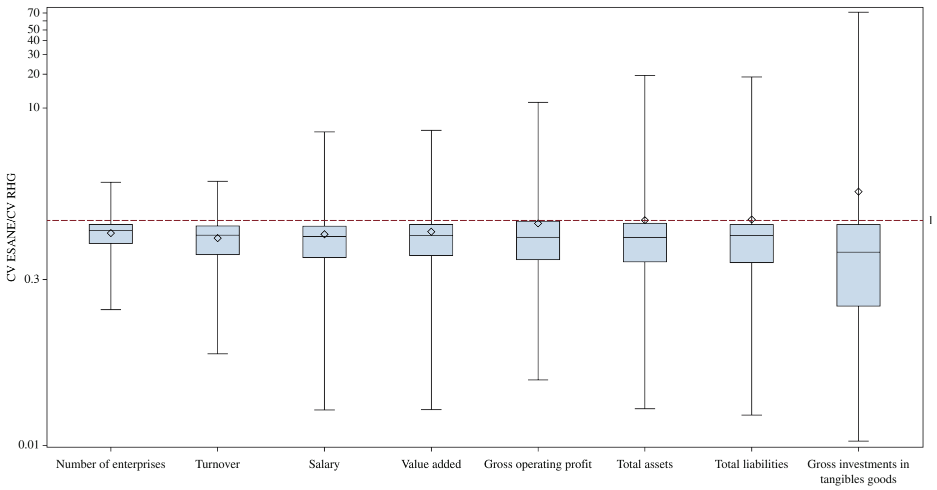


Fig. 6. Box plots of the ratio between sector-based Esane estimators' CVs and CVs relating to sector-based RHG estimators at the "group" (three-digit) level (using Esane's 2010 data)
Note: for a given variable, the diamond refers to the mean of the ratios.

variable of interest (like the variable "Sumptuary costs and expenses"), or mostly because the estimation is performed at fine levels of industry disaggregation. Indeed, as the industry disaggregation becomes finer, the amount of misclassification becomes larger, and simultaneously, the sample size available in finer-level cells to estimate this misclassification becomes smaller. Under these conditions, the difference estimators are not robust, and the change of the APE code of a single enterprise with a large value of one variable and/or a big sampling weight may create problems in the above formula, leading to negative values.

From a theoretical perspective, these negative estimates are not really problematic. Indeed, they merely reveal direct estimates' lack of precision when domain sample sizes are too small, a problem that would not necessarily appear so obviously when using classical methods: for such small domains, the RHG or calibrated estimators would have a very large variance, and when using administrative data directly with approximate values of the APE code coming from the register – available for all units but not necessarily up-to-date –, we would have a large bias.

However, these negative estimates constitute a practical drawback for the production of results at fine levels of industry disaggregation. To avoid being faced with a lot of potentially negative estimates for small domains, it has been decided to adjust the strategy concerning the estimators:

- For sector-based estimations at the "group" level (three digits of the NACE Rev.2 classification) and higher levels, the difference estimator presented in 3.3 is used. Indeed, at these relatively highly aggregated levels, we have very few "wrongly" negative estimates – less than 0,1% of all the group estimates – and they concern only variables of minor interest, such as the "Sumptuary costs and expenses". So we can deal with this problem by not publishing these rare negative estimates.

- For sector-based estimations at more detailed levels, we differentiate the "elementary" variables – that is, variables which are only components and never the result of accounting relationships – from the other variables:
  - For a given elementary variable Y, the group-level estimate is prorated to a finer level according to the structure of the elementary variable stemming from the survey. More precisely, for a group G and a finer area D ⊂ G, the total of Y on the area D is estimated by:

$$\hat{Y}^{D}_{\text{prorated}} = \hat{Y}^{G}_{\text{diff}} \frac{\sum_{i\in R} w_i Y_i 1I_{\text{area\_survey}=D}(i)}{\sum_{i\in R} w_i Y_i 1I_{\text{Group\_survey}=G}(i)}$$

  - For the other variables, the estimates result from the accounting relationships applied to the appropriate elementary variables estimates (see Gros 2012a for more details).

By construction, such a strategy ensures both positive estimates and consistency between the different estimates in the ESANE device, and these "prorated estimators" remain asymptotically unbiased. On the other hand, they use the administrative data less intensively at an individual level than the difference estimators, so we can expect more mixed performances in terms of accuracy. This expectation is confirmed by the comparison of sector-based estimators' CV at the five-digit level of the French nomenclature, presented in Figure 7 (Table 3 in the Appendix gives the means and quintiles corresponding to these box plots).

As we can see, at this fine level of industry disaggregation, the prorated estimators indeed lead to mixed results in terms of accuracy: they perform better than the RHG estimators only half of the time. In fact, neither of the two estimators is statistically better than the other, but the prorated estimator has the advantage of preserving the consistency of group-level estimates and finer-level estimates.

4. Other Issues

The new system was implemented in 2009, and at the present time has produced results for five years. Besides the questions of estimators that have been presented above, some other issues were raised.

First, the data editing of this composite material is complex. It has been divided into subprocesses, each one dedicated to one source (administrative or survey): this choice was made mainly to keep some flexibility in case of changes in one source, for example if the content of the tax files is modified. Moreover, the calendar of the deliveries of the different files is not the same: the return of the statistical survey questionnaires is spread over a long period, between March and October $n + 1$ concerning data of year n, while for tax data there are only a few deliveries, each one containing a large number of enterprises. Then, a step comparing the survey and the administrative data helps to achieve a cross validation of each source. More precisely, the value of the turnover, as its "rough breakdown" (between production, sales and services), is available in the two sources

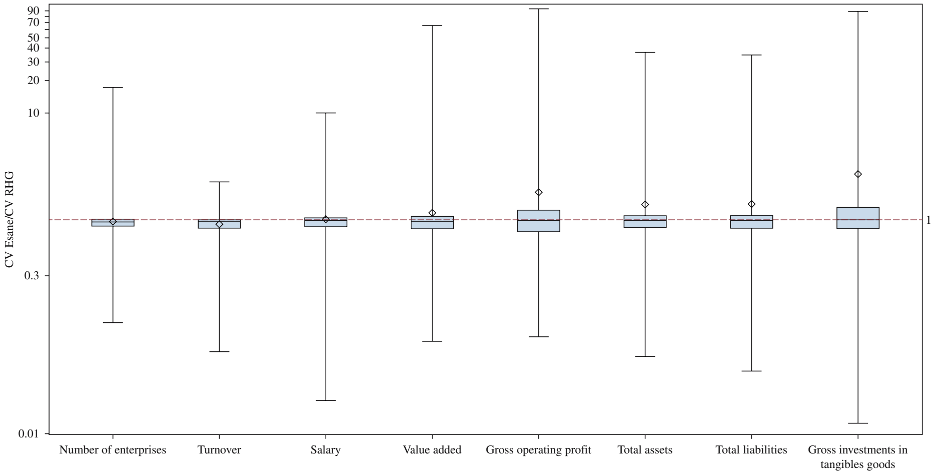


Fig. 7. Box plots of the ratio between sector-based Esane estimators' CVs and CVs relating to sector-based RHG estimators at the lower-class (5-digit) level (using Esane's 2010 data). Note: for a given variable, the diamond refers to the mean of the ratios

(survey and tax files), and the most important differences have to be checked by the clerks. This step is a very innovative part of the new system (Gros 2012b).

Questions were also raised concerning the scope of the business statistics. Using administrative and survey data jointly helped to revisit the choices made to define this scope. The scope is based on criteria available in the business register, such as the APE code and the legal status of the enterprise. Observing how the records of the tax files behaved relatively to the scope defined *a priori* helped to define choices concerning some specific categories of enterprises more precisely (Brion 2012b).

Mainly, the questions raised came back to the definition of the enterprise. In the European definition, an enterprise is the "smallest combination of legal units, that is, an organisational unit producing goods or services, which benefits from a certain degree of autonomy in decision making, especially for the allocation of its current resources". At the present moment, using a device mainly based on the legal units shows some limitations, and Insee is working to take the concept of enterprise into account better in the device: a second step concerning the renewing of the structural business statistics will consist in integrating these aspects, and some studies have shown that it will have general consequences for the significance of the statistics (Béguin et al. 2012). What is presented here concerns only the national part of the enterprise (sometimes named truncated enterprise) in the case of a multinational enterprise.

To conclude, we think that combining administrative and survey data leads to a strengthening of the quality of the produced statistics through the mutual improvement of the two kinds of sources. Moreover, in the presented device, the combined statistical estimators are intended to use every kind of information as much as possible. They show better statistical characteristics than other estimators, but in some cases this may go hand in hand with more complexity than in the case of the use of a single source.

Appendix

Table 2. Means and quintiles of the ratio between sector-based Esane estimators' CVs and CVs relating to sector-based RHG estimators at the "group" (three-digit) level (using Esane's 2010 data)

	Number of enterprises	Turnover	Salary	Added value	Gross operating profit	Total assets	Total liabilities	Gross investments in tangible goods
Mean	0.78	0.68	0.65	0.78	0.94	1.00	0.98	1.70
Max	2.20	2.26	3.36	6.34	12.37	19.61	18.84	71.30
Q99	1.89	1.97	3.04	5.13	7.81	13.92	18.06	48.11
Q95	1.02	1.02	1.37	1.80	1.99	3.04	3.10	5.10
Q90	0.99	0.97	1.07	1.12	1.45	1.46	1.36	1.88
Q75	0.93	0.89	0.89	0.92	0.97	0.94	0.92	0.92
Median	0.81	0.73	0.68	0.72	0.78	0.71	0.72	0.52
Q25	0.63	0.47	0.41	0.46	0.45	0.43	0.41	0.17
Q10	0.46	0.24	0.20	0.19	0.20	0.17	0.12	0.05
Q5	0.37	0.11	0.13	0.08	0.03	0.08	0.05	0.02
Q1	0.17	0.00	0.03	0.00	0.00	0.02	0.00	0.01
Min	0.16	0.00	0.02	0.00	0.00	0.00	0.00	0.00

Table 3. Means and quintiles of the ratio between sector-based Esane estimators' CVs and CVs relating to sector-based RHG estimators at the "under-class" (five-digit) level (using Esane's 2010 data)

	Number of enterprises	Turnover	Salary	Added value	Gross operating profit	Total assets	Total liabilities	Gross investments in tangible goods
Mean	0.96	0.91	1.00	1.16	1.79	1.38	1.41	2.64
Max	17.34	2.26	10.01	65.83	93.75	37.01	34.79	88.32
Q99	1.50	1.72	3.56	4.43	19.09	14.10	17.32	41.70
Q95	1.11	1.14	1.50	1.60	3.57	2.82	2.96	7.74
Q90	1.06	1.06	1.15	1.35	2.32	1.70	1.72	3.40
Q75	1.01	1.00	1.04	1.07	1.23	1.09	1.10	1.31
Median	0.96	0.96	0.98	0.97	0.99	0.99	0.98	1.00
Q25	0.87	0.83	0.86	0.83	0.77	0.85	0.83	0.82
Q10	0.68	0.62	0.63	0.61	0.54	0.59	0.58	0.47
Q5	0.59	0.46	0.46	0.50	0.36	0.39	0.33	0.20
Q1	0.33	0.25	0.12	0.16	0.16	0.14	0.10	0.02
Min	0.11	0.06	0.02	0.07	0.08	0.05	0.04	0.01

5. References

Béguin, J.M., V. Hecquet, and J. Lemasson. 2012. "France's Economic Fabric More Concentrated Than it Seemed. New Definition and New Categories of Enterprises." *Insee-première*, 1399. Insee, Paris. Available at http://www.insee.fr/en/ffc/ipweb/ip1399/ip1399.pdf (latest access October 2015).

Brion, P. 2007. "Redesigning the French Structural Business Statistics, Using More Administrative Data." In Proceedings of the Third International Conference on Establishment Surveys, June 18–21, 2007, Montreal, Canada. 533–541. Alexandria, VA [CD-Rom]: American Statistical Association. Available at: https://www.amstat.org/meetings/ices/2007/Proceedings/ICES2007-000034.pdf (accessed October 2015).

Brion, P. 2011. "Esane, Le Dispositif Rénové de Production des Statistiques Structurelles D'entreprises." *Courrier des Statistiques* n°130, Insee, Paris. Available at http://www.insee.fr/fr/ffc/docs_ffc/cs130d.pdf (accessed October 2015).

Brion, P. 2012a. "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics." *Journal of Official Statistics* 28: 341–347.

Brion, P. 2012b. "The New French System of Production of Structural Business Statistics." In Proceedings of the Fourth International Conference on Establishment Surveys, June 2012, Montreal, Canada. Available at: http://www.amstat.org/meetings/ices/2012/papers/302161.pdf (accessed October 2015).

Chami, S. 2010. "Reengineering French Structural Business Statistics: an Extended Use of Administrative Data." In Proceedings of the Q2010 Conference, May 4–6, 2010, Helsinki. Available at: https://q2010.stat.fi/sessions/session-27 (accessed October 2015)

Costanzo, L. 2011. "An Overview of the Use of Administrative Data for Business Statistics in Europe." ESSnet Admin Data, workpackage 1, Eurostat. Available at: http://essnet.admindata.eu/Document/GetFile?objectId = 5358 (accessed October 2015)

Deroyon, T. 2013. "Missing Data Treatment in Administrative Fiscal Sources for the French Structural Business Statistics Production System." In Proceedings of the Third European Establishment Statistics Workshop, September 9–11, 2013, Nuremberg. Available at: http://enbes.wikispaces.com/file/view/Deroyon%202013.pdf/456103752/Deroyon%202013.pdf (accessed October 2015)

Deville, J.C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87: 376–382. Doi: http://dx.doi.org/10.2307/2290268.

ESSnet on Administrative Data. 2011. "Main Findings of the Information Collection on the Use of Administrative Data for Business Statistics in EU and EFTA Countries." Deliverable 1.1, Eurostat. Available at: http://essnet.admindata.eu/WorkPackage/ShowAllDocuments?objectid=4251 (accessed October 2015)

Grandjean, J.P. 1997. "The System of Enterprise Statistics." *Courrier des Statistiques*. English series n°3, Insee, Paris. Available at: http://www.epsilon.insee.fr/jspui/bitstream/1/14403/1/csa3.pdf (accessed October 2015)

Gros, E. 2012a. "Esane ou les Malheurs de l'Estimateur Composite." In Proceedings of the *Journées de Méthodologie Statistique*, Insee, Paris. Available at: http://jms.insee.fr/files/documents/2012/936_2-JMS2012_S23-2_GROS-ACTE.PDF (accessed October 2015).

Gros, E. 2012b. "First Assessment of the Combined Use of Administrative and Survey Data in the New System of French Structural Business Statistics." In Proceedings of the Fourth International Conference on Establishment Surveys, June 2012, Montreal, Canada. Available at: http://www.amstat.org/meetings/ices/2012/papers/301882.pdf (accessed October 2015)

Haag, O. 2010. "Redesigning French Structural Business Statistics: Redesign of the Annual Survey." In Proceedings of the Q2010 Conference, May 4–6, 2010, Helsinki. Available at: https://q2010.stat.fi/sessions/session-14 (accessed October 2015)

Kovar, J. and P. Whitridge. 1995. "Imputation of Business Survey Data." In *Business survey methods*, edited by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott. New York: John Wiley.

Kroese, A.H. and R.H. Renssen. 2000. "New Applications of Old Weighting Techniques – Constructing a Consistent Set of Estimates Based on Data from Different Sources." In Proceedings of the Second International Conference on Establishment Surveys, June 17–21, 2000, Buffalo, NY. 831–840. Available at: http://www.amstat.org/meetings/ices/2000/proceedings/INTRO.pdf (accessed October 2015)

Little, R.J. 2012. Rejoinder to the Discussion of his Paper: "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics." *Journal of Official Statistics* 28: 367–372.

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

Received November 2012
Revised May 2015
Accepted August 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 611–625, http://dx.doi.org/10.1515/JOS-2015-0037

First Impressions of Telephone Survey Interviewers

Jessica Broome[1]

Survey nonresponse may increase the chances of nonresponse error, and different interviewers contribute differentially to nonresponse. This article first addresses the relationship between initial impressions of interviewers in survey introductions and the outcome of these introductions, and then contrasts this relationship with current viewpoints and practices in telephone interviewing. The first study described here exposed judges to excerpts of interviewer speech from actual survey introductions and asked them to rate twelve characteristics of the interviewer. Impressions of positive traits such as friendliness and confidence had no association with the actual outcome of the call, while higher ratings of "scriptedness" predicted lower participation likelihood. At the same time, a second study among individuals responsible for training telephone interviewers found that when training interviewers, sounding natural or unscripted during a survey introduction is not emphasized. This article concludes with recommendations for practice and further research.

Key words: Survey; telephone; nonresponse; interviewer.

1. Introduction and Background

Survey nonresponse has the potential to bias survey estimates (Groves et al. 2004). It has been demonstrated that telephone interviewers vary substantially in their response rates (Oksenberg and Cannell 1988). Identifying vocal characteristics and techniques of more successful telephone interviewers (i.e., those with higher overall response rates) may impact data quality by allowing for more targeted screening and training of interviewers with the aim of reducing nonresponse.

Literature from both survey methodology (Oksenberg et al. 1986) and telemarketing (Ketrow 1990) has found that a pleasing or attractive voice in the initial seconds of a phone call is imperative in extending the interaction. Further, Ketrow (1990) discusses the importance of giving an initial impression of competence, and Oksenberg and colleagues (Oksenberg et al. 1986; Oksenberg and Cannell 1988) found that judges' ratings of phone-interviewer competence based on brief recorded excerpts were positively associated with the interviewers' success. This is not to imply that in survey interview introductions, having a pleasing, competent-sounding voice in the opening statement is enough to guarantee success. However, an interviewer voice that gives listeners a positive first

[1] University of Michigan – Institute for Social Research, 426 Thompson Street, Ann Arbor, Michigan 48104, U.S.A. Email: jsbroome@umich.edu

Acknowledgments: The author would like to acknowledge the Charles Cannell Fund in Survey Methodology, the University of Michigan Center for the Education of Women, the A. Regula Herzog Young Investigators Fund, and the Harvey G. and Joyce H. Behner Graduate Fellowship for their support as well as Frederick Conrad, Norbert Schwarz, Steve Heeringa, and Susan Brennan for their input.

impression may lead to a longer conversation, thus increasing the likelihood of participation.

Nonresponse to telephone surveys has been increasing steadily over the past 25 years (Curtin et al. 2005). Declining response rates have the potential to increase nonresponse error (Groves et al. 2004; Teitler et al. 2003). Further, nonresponse rates vary by interviewer (Morton-Williams 1993; Oksenberg and Cannell 1988; O'Muircheartaigh and Campanelli 1999; Snijkers et al. 1999). Uncovering the characteristics and tactics of successful interviewers can help to reduce nonresponse, either by using vocal and personality characteristics as hiring criteria or by training interviewers to adopt characteristics or tactics which have been shown to lead to increased success.

In contrast to face-to-face interviewers, telephone survey interviewers have just two primary tools that are under their control in their efforts to persuade answerers to participate: what they say (speech) and how they say it (vocal characteristics). A small body of literature (e.g., Sharf and Lehman 1984; Oksenberg et al. 1986; Oksenberg and Cannell 1988; Groves et al. 2007; Conrad et al. 2013) finds relationships between vocal characteristics of interviewers in telephone-survey introductions and interviewer success in obtaining interviews. In general, successful interviewers have been ones who spoke louder (Oksenberg et al. 1986; Oksenberg and Cannell 1988; van der Vaart et al. 2005) and with more falling intonation (Sharf and Lehman 1984; Oksenberg and Cannell 1988). In addition, success has been shown to be correlated both with higher mean fundamental frequency (Sharf and Lehman 1984) and higher perceived pitch (Oksenberg et al. 1986), as well as variable fundamental frequency (Sharf and Lehman 1984; Groves et al. 2007) and variable pitch (Oksenberg et al. 1986). The terms "pitch" and "fundamental frequency" are often used interchangeably, but a necessary distinction is that fundamental frequency is an acoustic measure of vocal-chord vibrations, while pitch is a listener's perception of frequency or how "high" or "low" a voice sounds.

Three recent studies have found nonlinear relationships between success and rate of speech (Groves et al. 2007; Steinkopf et al. 2010; Benkí, Broome, Conrad, Groves and Kreuter 2011): contacts with speech that is either overly slow or overly fast tend to be less successful. Benkí et al. (2011) found that contacts with interviewer speech in the range of 3.34–3.68 words per second were the most likely to be successful.

One critical question concerns what underlies these associations; what is it about an interviewer who speaks at a particular rate or with more variable pitch that leads to success, especially given the limited amount of exposure an answerer has to the interviewer's voice before deciding whether or not to participate? Oksenberg et al. (1986, 99) emphasized the importance for an interviewer to have a voice that potential respondents find appealing in the first few seconds of a survey interview introduction context, stating that "if vocal characteristics lead the respondent to perceive the interviewer as unappealing, cooperation will be less likely."

Two dimensions of person perception, warmth and competence, have been shown to be relevant to the development of first impressions of others across a range of contexts (Asch 1946; Fiske, Cuddy, and Glick 2007; Kelley 1950; Rosenberg, Nelson, and Vivekanathan 1968). Several studies in the literature on interviewer vocal characteristics (Oksenberg et al. 1986; van der Vaart et al. 2005) suggest that ratings of personal characteristics on these dimensions of person perception are associated with both interviewer response rates and vocal characteristics. These studies involved collecting ratings of several interviewer personality characteristics, which were then successfully reduced to two dimensions interpretable as "warmth" and "competence." Characteristics on the "warmth" dimension included being cheerful, friendly, enthusiastic, polite, interested in her task, and pleasant to listen to. Oksenberg et al. (1986) and van der Vaart et al. (2005) found correlations between high ratings on the warmth dimension and vocal characteristics, including variation in pitch, higher pitch, and a faster rate of speech, suggesting that listeners' impressions of interviewer personality are based, at least in part, on physical (acoustic) attributes of interviewers' voices. Characteristics composing the "competence" dimension included being self-assured, educated, intelligent, and professional. Van der Vaart et al. (2005) found that interviewers rated highly on "competence" characteristics tended to have lower pitch.

Importantly, Oksenberg et al. (1986) and Van der Vaart et al. (2005) found that high ratings on a "warmth" dimension correlated with ratings of judges' willingness to participate. This aligns with Morton-Williams's (1993) finding that warm or "likable" interviewers increased perceived benefits to potential respondents and improved participation rates, and also with Cialdini's (1984) "Liking" Principle of Compliance: people are more likely to comply with a request from someone they like.

Further, Cialdini (1984) suggests a compliance heuristic based on the principle of authority; requests from an authoritative speaker are more likely to be honored than requests with less authority. Impressions of authoritative characteristics such as competence and confidence, in turn, have been shown to be associated with interviewer success (Oksenberg et al. 1986; Oksenberg and Cannell 1988; Steinkopf et al. 2010).

While a small body of literature explores the relationship between interviewer vocal characteristics and impressions, there are clearly challenges to conducting research in this area. For example, the independent variables used are judges' ratings of an interviewer's vocal characteristics. When such ratings are collected in person, small sample sizes tend to be the norm; for example, two early studies (Oksenberg et al. 1986; Oksenberg and Cannell 1988) were each based on six recordings. Studies with larger numbers of judges, such as those by Huefken and Schaefer (2003), with 51 judges, Steinkopf et al. (2010), which used 56 judges, and Van der Vaart et al. (2005), with twelve judges, were based on the work of student (rather than professional) interviewers, limiting the applicability of findings. Finally, dependent variables assessed in existing studies are either interviewers' historical response rates, judges' own willingness to comply, or judges' beliefs that "someone" will comply; no study has yet associated vocal characteristics with actual contact outcomes.

The aim of the exploratory studies described in this article was to see whether first impressions, formed in the initial seconds of a telephone interviewer's introduction, are an important component in determining the outcome of a survey introduction. This article will address several questions concerning first impressions of telephone interviewers:

- Which first impressions are most predictive of a successful outcome?
- How do relationships between first impressions and success compare with practitioners' ideas about what makes a successful interviewer?

The first hypothesis addressed in this article (h1) is that interviewers who are perceived more positively and less negatively in the initial seconds of a contact by judges will have greater success, as measured by contact outcome.

The second hypothesis (h2) is that the characteristics that practitioners perceive as important to interviewer success will parallel those characteristics that predict actual cooperation.

This article reports results from two studies. The "Listeners' Study" elicited ratings of interviewer personality characteristics in audio-recorded telephone introductions from five surveys. The raters (or "listeners") were internet-survey panel members who answered questions after listening to brief excerpts of interviewer speech from real (not staged) telephone-survey introductions. Having a large number of raters, combined with the use of real contacts conducted by professional interviewers for which the actual outcome is known, is unique in studies on interviewer voice. Another novel element of this study is that, in order to explore whether practitioners focus on the attributes of interviewer speech that are most related to the outcome of survey invitations, results from the Listeners' Study are contrasted with those from a web survey of practitioners who hire and train telephone interviewers. In this "Practitioners' Study," practitioners were asked which characteristics of interviewers they consider in hiring and training.

This article concludes with a discussion of implications for practice, as well as suggestions for future research in this area.

2. Data and Methods

The data described in this section are drawn from two web surveys. The "Listeners' Study" was a survey among 3,403 adult, English-speaking members of an internet-survey panel. The "Practitioners' Study" was a smaller survey of 44 survey practitioners who are responsible for the hiring and training of survey interviewers in academic, government, and for-profit survey organizations.

2.1. Listeners' Study: Selection of Contacts

The recordings used in the listeners' study were selected from 1,380 audio-recorded telephone-survey introductions. These introductions, conducted by 100 interviewers, were from five telephone surveys that were audio recorded for another project. In this project, all contacts associated with selected households, regardless of who the interviewer was, were included in the dataset (Benkí et al. 2011; Conrad et al. 2013). Contacts by 49 different interviewers with ranging lengths of tenure and response rates varying over the course of their tenure at University of Michigan from .07–.21 are included in this dataset.

The recordings were classified into five outcomes: "agree," where the answerer cooperates and agrees to participate; "refuse," where there is an explicit refusal (for example, "I will not take the survey. Please do not call again"); "scheduled callback," where the interviewer either schedules a time to call back or asserts that she will call again; "hang up," where the answerer hangs up but never clearly refuses; and "other."

The listeners' study uses excerpts from these recorded introductions (referred to hereafter as "contacts") from three of the five studies. To facilitate comparisons (particularly in analyses of vocal characteristics such as pitch), only introductions by female interviewers were selected. Contacts in which the answerer hangs up during or directly following the interviewer's first speaking turn were excluded, using the rationale that these are "hard-core nonrespondents" who are determined not to become respondents, and the interviewer has no opportunity to use her voice or speech to convince them otherwise.

Because listeners were asked to make judgments about the interviewer's personality, contacts had to contain enough speech to make these determinations. The minimum amount of speech required for inclusion was a statement of name and affiliation. Finally, contacts were omitted if the interviewer asked for a particular person by name, indicating that the interviewer had already spoken at length to someone in the household, and the persuasion process was likely to be quite different.

Applying these criteria to the 1,380 contacts resulted in 283 recordings from the Survey of Consumer Attitudes, or SCA ($n = 168$); the National Study on Medical Decisions, or NSMD ($n = 110$); and the Mississippi Community Study, or MCS ($n = 5$).

These 283 contacts form the basis of the listeners' study. 118 (42 percent) had an outcome of "agree" and 165 (58 percent) had an outcome of "refuse." Listeners were not told the likelihood of either outcome.

2.2. Listeners' Study: Description of Stimulus and Lines of Questioning

The listeners' study used online presentation of audio recordings of telephone-survey invitations to elicit listeners' judgments about telephone interviewers' personality characteristics. In this study, 3,403 members of an online survey panel listened to interviewer speech from the contacts described above. The stimuli to which listeners were exposed consisted of brief introductory statements by the interviewer, such as: "Hello, my name is ________ and I'm calling from the University of Michigan about our survey on________."

Each listener heard excerpts from five contacts randomly selected from the corpus described above, which contained 283 introductions by 49 different interviewers. It was possible for some listeners to hear multiple introductions by one interviewer, and for others to hear five different interviewers. Interviewers had between 1 and 23 contacts in the dataset, with an average of 5.8 contacts. While the same group of five contacts could conceivably be heard by multiple listeners, assignment and order of excerpts were random so as to avoid context effects from presenting excerpts in set groups or a set order.

For each of the five contacts, listeners were asked to rate the interviewer on twelve characteristics (confident, competent, professional, knowledgeable, enthusiastic, pleasant to listen to, friendly, natural sounding, genuine, scripted, irritating, and uncertain), using a scale from 1 (not at all) to 6 (extremely). These are referred to as "characteristic ratings" below. Many of these traits have been shown to be related to interviewer success in the literature (Oksenberg and Cannell 1988; van der Vaart et al. 2005).

2.3. Listeners' Study: Preparation of Contacts

After selecting the interviewer speech to be used, the recording was amplified to use the full range of sounds that a recorded voice would make. Amplification was maintained at the same level for all contacts, thus making all contacts comparable in volume. Finally, to preserve interviewers' anonymity, the interviewer's name in each contact was replaced with a quarter-second-long tone. For consistency, this was done even in the few cases where the interviewer only said her first name.

2.4. *Listeners' Study: Data Collection and Respondent Descriptives*

Data collection was conducted by a commercial vendor, Lightspeed Research (http://www.lightspeedresearch.com/). 15,000 invitations were sent to a stratified random sample of members of Lightspeed's own 1.3 million-member volunteer online panel. (All sample members had a known chance of being invited; the list of invitees was stratified by gender, age and region in an attempt to attain representativeness of the US population on these variables.) This panel is recruited using a variety of sources, including opt-in email, co-registration, e-newsletter campaigns, and placements of banner advertisements. Panelists receive regular survey invitations. While the panel does not perfectly mirror the gender distribution of the US population according to Census data (32% of panelists are male, compared to 49% in the population), the respondents to the listeners' survey were more representative in terms of gender. Respondents were evenly divided between males (49%) and females (51%). One-third (33%) were aged 60 or older, while 20% were 50–59, 18% were 40–49, 17% were 30–39, and 12% were 18–29. 88% of respondents were white (compared to 68% in the general population), and 81% had at least some college education (compared to 55% of the national population, according to the 2010 US Census). These discrepancies can be considered a limitation of the study.

This study was fielded August 12–18, 2011. Panel members were screened to ensure that they were 18 years of age or older (as would be any eligible respondents to the surveys represented by these contacts), and that they characterized their ability to understand spoken English as "excellent" or "good." This screening criterion was deemed necessary for listeners to be expected to make personality judgments about the interviewer based on brief speech clips.

After their eligibility for the study was determined, listeners were exposed to an "introductory" audio clip and asked to identify one of the words in the clip. The purpose of this exercise was threefold: first, to ensure that listeners were using a computer with working audio; second, to familiarize them with the type of audio they would be hearing during the survey; and third, as a quality-control check to ensure that listeners could sufficiently distinguish words in the contact. 126 potential listeners were screened out at this stage; 3,403 listeners completed the survey.

While the mean exposure length of contacts was 10.32 seconds, the range was wide: from 2.3 to 49.2 seconds. To roughly match the burden on listeners and ensure that none received multiple long contacts, contacts were stratified into five groups based on logical length categories. There were between 45 and 70 contacts in each length category, resulting in more ratings for some contacts than others; the mean number of ratings by length category ranged from 49 to 76. Each listener was exposed to a set of five introductions, each consisting of one randomly selected contact from each length category. For each contact, listeners rated the interviewer on the twelve characteristics discussed above.

2.5. Practitioners' Study

As mentioned above, a second study was conducted among individuals responsible for hiring and training telephone interviewers. A questionnaire was created to assess these practitioners' ratings of the importance of various behaviors and attributes to telephone interviewers' success, as well as to understand their current focuses in hiring and training telephone interviewers. The final survey was programmed in the online survey tool Qualtrics.

A sampling frame was developed that relied on personal contacts of the author, as well as on a list of all members of the Association of Academic Survey Research Organizations (AASRO), which, although not a complete listing of academic survey institutions, was fairly comprehensive and readily accessible. The final sample consisted of 113 individuals at 108 organizations, including two government, 92 academic, three non-profit and eleven for-profit organizations. Two weeks after the initial invitation was sent, a reminder email was sent to all members of the original frame with working email addresses, with the exception of those participants who had already provided their email addresses (respondents were given the option to provide their email addresses if they wished to receive a copy of the results), and those sample members who had requested no further contact.

The survey was completed by 44 respondents between June 5 and July 12, 2011, resulting in a 42% response rate. This response rate is sufficient for the purposes of this study, which was not to uncover precise estimates but rather general trends among practitioners. Further, variation in the organizational characteristics (number of CATI stations and number of interviews conducted in 2010) of those who did respond reduces the chances of nonresponse bias.

Respondents represented a wide range of organizations in size and workload. The median number of CATI stations in respondent organizations was 25, but the number of stations ranged from nine to 450. Close to half (42%) of respondents reported that their organization had conducted fewer than 5,000 telephone interviews in 2010, while an equal percentage reported that their organization had conducted 10,000 or more interviews. Respondents reported that, on average, 81% of the interviews their organizations conducted were for government, academic, or non-profit organizations, 15% were for for-profit organizations, and 2% were for "other" organizations (2% were not sure).

To qualify for the study, practitioners had to have responsibility for hiring and/or training telephone interviewers. Of the 44 respondents, 41 indicated responsibility for hiring interviewers and 40 indicated responsibility for training interviewers.

3. Results

3.1. Listeners' Study: Characteristic Ratings

3.1.1. Description of Ratings

The listeners' study asked for ratings of twelve characteristics of interviewers from five contacts per listener on a six-point scale (1 = not at all to 6 = extremely). Each contact was rated by at least 30 listeners. Analyses were conducted at the contact level; for each contact, a mean rating across all listeners who heard it was calculated for each characteristic. The mean of all contact-level means, as well as the minimum and maximum mean for each characteristic, are reported. The mean ratings for each characteristic across contacts ranged from 2.50 to 3.89, as shown in Table 1.

Ratings for all nine positive characteristics were highly correlated, as shown in Table 2, indicating that an overall impression of positivity drives judgments.

In addition, a factor analysis, using ratings of all characteristics except scriptedness, revealed that only one factor, explaining 86% of total variance, had extremely high loadings for all nine positive characteristics. The method of principal factors was used to extract factor scores. The overall Kaiser-Meyer-Olkin measure of sampling adequacy, which gives a measure of how much each item is correlated with the others, was 0.92.

Ratings of "uncertain" and "irritating" were highly correlated with each other (.71), but ratings of "scripted" were not highly correlated with ratings of any other characteristic. The mean correlation between scripted and positive characteristics was .01.

3.1.2. Characteristic Ratings As Predictors of Contact Outcome

It was hypothesized (h1) that when ratings of nine positive interviewer characteristics (enthusiastic, friendly, natural, genuine, pleasant to listen to, confident, professional, competent, and knowledgeable) were high and ratings of three negative characteristics (irritating, uncertain, and scripted) were low, a contact would be more likely to result in cooperation than when the positive characteristics were rated lower and the negative characteristics were rated higher.

This hypothesis was partially supported. Twelve bivariate logistic models were constructed, all accounting for the multilevel structure of this dataset (contacts nested within interviewers). For all of these models, the dependent variable was coded as $y = 1$ (agree) or $y = 0$ (refusal). The equation for these models can be written as $\log(\pi_{\text{agree}}/1 - \pi_{\text{agree}}) = a + b_1x_1 + u_j$, where π_{agree} denotes the probability of cooperation; a is an intercept; x_1 represents the mean rating of a characteristic

Table 1. Description of characteristic ratings

Characteristic	Mean of contact-level means (sd)	Minimum contact- level mean rating	Maximum contact-level mean rating	Spread
Confident	3.62 (.56)	1.77	4.63	2.86
Professional	3.70 (.52)	1.85	4.73	2.88
Pleasant to listen to	3.54 (.46)	2.13	4.49	2.36
Competent	3.67 (.50)	1.90	4.67	2.77
Knowledgeable	3.61 (.49)	2.13	4.75	2.62
Natural sounding	3.65 (.41)	2.35	4.47	2.12
Enthusiastic	3.43 (.50)	2.25	4.51	2.26
Genuine	3.59 (.39)	2.41	4.45	2.04
Scripted	3.78 (.31)	2.70	4.67	1.97
Friendly	3.89 (.38)	2.84	4.62	1.78
Uncertain	2.70 (.50)	1.83	4.55	2.72
Irritating	2.50 (.35)	1.73	3.51	1.78

Table 2. Correlations between characteristic ratings

	CON	FRI	PRO	PLE	COM	KNO	NAT	ENT	GEN	SCR	UNC	IRR
Confident	1.00											
Friendly	0.77	1.00										
Professional	0.94	0.72	1.00									
Pleasant	0.86	0.86	0.88	1.00								
Competent	0.97	0.76	0.97	0.88	1.00							
Knowledgeable	0.92	0.70	0.92	0.82	0.94	1.00						
Natural	0.84	0.81	0.88	0.93	0.88	0.83	1.00					
Enthusiastic	0.78	0.88	0.66	0.72	0.73	0.68	0.69	1.00				
Genuine	0.87	0.84	0..88	0.91	0.90	0.89	0.93	0.75	1.00			
Scripted	0.13	−0.13	0.16	0.01	0.11	0.13	−0.11	−0.10	−0.09	1.00		
Uncertain	−0.89	−0.63	−0.85	−0.72	−0.85	−0.77	−0.69	−0.65	−0.70	−0.17	1.00	
Irritating	−0.70	−0.63	−0.76	−0.84	−0.76	−0.66	−0.78	−0.49	−0.73	−0.01	0.71	1.00

(for example, scriptedness) across all listeners who rated the contact; and u_j represents the random, unobserved effects of interviewers.

Of these twelve models, only the model for scripted had a significant coefficient ($b = -1.05$, standard error = .40, $z = -2.59$, $p = 0.010$), indicating that perceptions of the interviewer as more scripted decrease the likelihood of a contact's success. These results persisted when the models controlled for the length of exposure, and also when only the contacts with the longest exposure lengths (at least ten seconds) were analyzed.

A model was constructed which predicted contact outcome using the factor score described above and the contact's mean scriptedness rating, while controlling for recording length and accounting for clustering by interviewer. The equation for this model was $\log(\pi_{\mathrm{agree}}/1 - \pi_{\mathrm{agree}}) = a + b_1x_1 + b_2x_2 + b_3x_3 + u_{\mathrm{j}}$, where π_{agree} denotes the probability of cooperation; a is an intercept; x_1 represents the contact's factor score; x_2 represents the mean rating of scriptedness across all listeners who rated the contact; x_3 represents the length of the recording; and u_{j} represents the random, unobserved effects of interviewers. As Table 3 shows, only scriptedness was a significant predictor in this model ($z = -2.65, p = 0.008$); the factor score was not, indicating that initial impressions of scriptedness, but not of any other characteristic, are important to a contact's outcome.

In summary, there was no support for the hypothesis that positive characteristics would predict a successful outcome. But a negative characteristic, scriptedness, was negatively associated with success, with contacts where interviewers are less scripted being more successful than those who were rated as more scripted. Agreement with the survey request exhibits almost no variation across interviewers after accounting for other factors including scriptedness, but substantial variation across contacts; in contacts where the interviewer is not considered scripted, agreement is more likely.

3.2. Comparison Between Listeners' and Practitioners' Surveys

Hypothesis 2 was that practitioners' views of what makes a successful interviewer would align with the characteristics which were found to predict contact success in the listeners' study. To some degree, this is the case; practitioners recognize "how genuine the interviewer sounds" and "the ability to respond to concerns expressed by potential respondents" as important to an interviewer's response rate, acknowledging that interviewers should not sound robotic during their introductions. However, practitioners appear conflicted; while they recognize the need for a genuine-sounding interaction between interviewer and answerer, they also emphasize the need for interviewers to follow a script, with 48% saying "an interviewer's ability to follow a script during an

Table 3. Coefficients in model predicting cooperation

	Coefficient	SE	Z	P
Factor score	−.017	.12	−0.14	0.89
Scriptedness rating	−1.07	.40	−2.65	0.008
Length of recording	.016	.02	.98	0.33

Standard deviation of random interviewer effects: .0000227
Variance of random interviewer effects: 5.0625E-10
Intraclass correlation coefficient: 1.53881E-10

introduction" is extremely important, and 38% rating it as somewhat important. Practitioners rank "an interviewer's ability to 'ad lib' or deviate from a script during an introduction" as slightly less important to an interviewers' success. On the other hand, results from the listeners' study indicated that impressions of scriptedness are, in fact, detrimental to the success of contacts, with lower ratings of scriptedness found in successful contacts; indeed, scriptedness is the *only* characteristic that matters to listeners.

Further, of the 18 elements tested (shown in Table 4), the one judged by survey practitioners as most important to an interviewer's success in obtaining interviews was "the initial impression an interviewer gives to sample members." This "initial impression" may well include an interviewer's degree of scriptedness; however, this finding shows that emphases differ between practitioners and listeners, as findings from the listeners' survey indicate that, aside from scriptedness, no ratings of interviewer characteristics based on early impressions can predict success on a given contact.

Far more important than an ability to "ad lib", according to practitioners, were traits such as competence, professional demeanor, and confidence – ratings of which were not predictive of actual contact-level outcome.

Among practitioners responsible for training telephone interviewers, just 15% reported that "developing a personalized or nonscripted introduction" is a primary focus of their organization's interviewer training, while 44% reported that it is not a focus at all. (Practitioners were shown a list of 13 items and given the instruction, "For each of the following, please indicate if it is a primary focus, a secondary focus, or not a focus at all in telephone-interviewer training.") "Following introductory scripts," by contrast, was a primary training focus in the vast majority (78%) of organizations surveyed. This emphasis on following introductory scripts contrasts with the finding in the listeners' study that higher ratings of scriptedness predict less success at the contact level.

These results demonstrate a disconnect between listeners and practitioners. While listeners' judgments of scriptedness are predictive of a contact's success (indeed, this is the only rated characteristic associated with success), practitioners place less emphasis on reducing scriptedness and more on other impressions conveyed by interviewers.

4. Conclusions, Applications, and Discussion

This exploratory research has found that survey practitioners believe that initial impressions of an interviewer are important to that interviewer's success. By contrast, most ratings of interviewer traits such as competence, confidence, and professionalism based on a brief exposure are not predictive of the ultimate outcome of the conversation. One exception to this is ratings of scriptedness, which are significant predictors of contact outcome; indeed, scriptedness is the only rated component of a first impression to predict success. This may be attributable to scriptedness being the most noticeable of all characteristics rated, overwhelming characteristics such as friendly and professional in the initial seconds of an introduction. However, practitioners emphasize the importance of "following a script," even though this practice might actually harm interviewers' chances of obtaining a completed interview.

This can be applied to survey practice, as an emphasis on decreasing the scripted or unnatural nature of survey introductions may well serve to improve interviewer

Table 4. Practitioners' ratings of importance to interviewer's success

	Scale: 1 (not at all important) – 4 (very important)	Mean (sd)	% Extremely important	% Somewhat/ extremely important	% Not very/ not at all important
1	The initial impression an interviewer gives to sample members.	3.88 (.33)	88%	100%	0%
2	The ability to address concerns expressed by potential respondents.	3.84 (.38)	83%	100%	0%
3	How competent the interviewer sounds to potential respondents.	3.84 (.38)	83%	100%	0%
4	Professional demeanor when talking to potential respondents.	3.81 (.44)	86%	98%	2%
5	How confident the interviewer sounds to potential respondents.	3.81 (.45)	83%	98%	2%
6	The ability to convey knowledge about the study.	3.72 (.46)	71%	100%	0%
7	How genuine the interviewer sounds to potential respondents.	3.70 (.47)	69%	100%	0%
8	An interviewer's voice that does not sound monotonous (has pitch variability).	3.58 (.5)	57%	100%	0%
9	How friendly the interviewer sounds to potential respondents.	3.53 (.55)	57%	98%	2%
10	The interviewer's speech rate.	3.42 (.59)	48%	95%	5%
11	How enthusiastic the interviewer sounds to potential respondents.	3.42 (63)	50%	93%	7%
12	A pleasant-sounding voice.	3.41 (.59)	45%	95%	5%
13	The interviewer's ability to follow a script during an introduction.	3.30 (.78)	48%	86%	14%
14	The interviewer speaks without any "um's" or "uh's."	3.16 (.71)	36%	81%	19%
15	The interviewer's ability to "ad lib" or deviate from a script during an introduction.	2.95 (.89)	29%	72%	28%
16	How high or low the interviewer's voice sounds (pitch).	2.70 (.74)	12%	63%	37%
17	The interviewer emphasizes the length of the survey.	2.67 (.67)	12%	65%	35%
18	The interviewer emphasizes the incentive.	2.47 (.74)	10%	51%	49%

performance. Currently, most practitioners train and encourage interviewers to follow a script and, to a lesser degree, emphasize "ad libbing" during an introduction. It appears that practitioners recognize the importance of not sounding scripted; however, unscriptedness is admittedly difficult to train and measure.

Recommendations for interviewer training can be made based on these results. Specifically, interviewer response rates may benefit from an emphasis in interviewer training on speech that is as natural and unscripted as possible, through the use of intonation patterns and word selection. Interviewers can be exposed to contacts with both high and low ratings of scriptedness to make the difference clear.

While interviewers may be required to mention particular points in their introduction or even to follow a verbatim introductory script, they should be trained to sound as conversational as possible, particularly at the start of their introduction. Both Houtkoop-Steenstra and van den Bergh (2000) and Morton-Williams (1993) found that interviewers who were allowed to adapt their introductory script had greater success.

Beyond the introduction, the issue of standardized interviewing, and what departures from verbatim interview scripts can mean for data quality, is the subject of much debate. Schober and Conrad (1997; Conrad and Schober 2000) found clear evidence that "conversational" interviewing, or allowing interviewers to use any means necessary to convey question meaning, can enhance data accuracy. Nevertheless, "reading the questions exactly as worded" is a tenet of interview administration which is upheld and enforced in most survey organizations, and it is clear from results of the practitioners' study that standardized interviewing skills are a high priority in nearly all organizations. Because emphasizing the need to read questions in a standardized manner may seem to conflict with emphasis on less-scripted delivery of introductions, interviewers need to be trained to "wear two hats." In training, it needs to be made explicit to interviewers that there are two distinct (but, arguably, equally important) elements of the phone component of their job, each requiring a different style of speech and interaction. In the introductory or persuasive portion, scriptedness may be a liability, and the ability to "think on one's feet" to respond to answerers is an asset. In contrast, in the interviewing portion, deviating from a script may have ramifications for data quality, or at the very least will represent a lack of adherence to the organization's procedures. Interviewers should be trained to "switch gears" between these two speech styles, and perhaps even be encouraged to acknowledge to respondents that their delivery of the questions will sound different from their introduction.

Finally, additional research could further these findings. Interviewer-level analyses, such as a larger study collecting ratings of characteristics for a greater number of contacts per interviewer to measure the impact of ratings on overall success rates, is recommended. Replicating the listeners' study using a greater number of contacts by the same interviewers may shed light on those interviewer characteristics or behaviors across multiple contacts that lead to greater success.

5. References

Asch, S.E. 1946. "Forming Impressions of Personality." *Journal of Abnormal and Social Psychology* 9: 258–290. Doi: http://dx.doi.org/10.1037/h0055756.

Benkí, J., J. Broome, F. Conrad, R. Groves, and F. Kreuter. 2011. "Effects of Speech Rate, Pitch, and Pausing on Survey Participation Decisions." Paper presented at the 66th annual conference of the American Association for Public Opinion Research, Phoenix, AZ, May 14. Available at: http://www.amstat.org/sections/srms/proceedings/42011/files/400189.pdf (accessed September 29, 2015).

Cialdini, R.B. 1984. *Influence: Science and Practice*. New York: Harper Collins.

Conrad, F., J. Broome, J. Benkí, R. Groves, F. Kreuter, D. Vannette, and C. McClain. 2013. "Interviewer Speech and the Success of Survey Invitations." *Journal of the Royal Statistical Society* 176: 191–210. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01064.x.

Conrad, F.G. and M.F. Schober. 2000. "Clarifying Question Meaning in a Household Telephone Survey." *Public Opinion Quarterly* 64: 1–28. Doi: http://dx.doi.org/10.1086/316757.

Curtin, R., S. Presser, and E. Singer. 2005. "Changes in Telephone Survey Nonresponse over the Past Quarter Century." *Public Opinion Quarterly* 69: 87–98.

Fiske, S.T., A.J.C. Cuddy, and P. Glick. 2007. "Universal Dimensions of Social Cognition: Warmth and Competence." *Trends in Cognitive Science* 11: 77–83. Doi: http://dx.doi.org/10.1016/j.tics.2006.11.005.

Groves, R.M., B.C. O'Hare, D. Gould-Smith, J. Benkí, and P. Maher. 2007. "Telephone Interviewer Voice Characteristics and the Survey Participation Decision." In *Advances in Telephone Survey Methodology*, edited by J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, and R.L. Sangster. (pp. 385–400). New York: Wiley.

Groves, R.M., S. Presser, and S. Dipko. 2004. "The Role of Topic Interest in Survey Participation Decisions." *Public Opinion Quarterly* 68: 2–31. Doi: http://dx.doi.org/10.1093/poq/nfh002.

Houtkoop-Steenstra, H., and H. van den Bergh. 2000. "Effects of Introductions in Large-Scale Telephone Survey Interviews." *Sociological Methods and Research* 28: 281–300. Doi: http://dx.doi.org/10.1177/0049124100028003002.

Huefken, V. and A. Schaefer. 2003. "Zum Einfluss Stimmlicher Merkmale und Ueberzeugungsstrategien der Interviewer auf die Teilnahme in Telefonumfragen." *Koelner Zeitschrift fuer soziologie und sozial psychologiei* 55: 321–339.

Kelley, H.H. 1950. "The Warm-Cold Variable in First Impressions of Persons." *Journal of Personality* 18: 431–439. Doi: http://dx.doi.org/10.1111/j.1467-6494.1950.tb01260.x.

Ketrow, S.M. 1990. "Attributes of a Telemarketer's Voice and Persuasiveness: A Review and Synthesis of the Literature." *Journal of Direct Marketing* 4: 7–21. Doi: http://dx.doi.org/10.1002/dir.4000040304.

Morton-Williams, J. 1993. *Interviewer Approaches*. Cambridge: Cambridge University Press.

Oksenberg, L. and C. Cannell. 1988. "Effects of Interviewer Vocal Characteristics on Nonresponse." In *Telephone Survey Methodology*, edited by R.M. Groves, P.B. Biemer, L.E. Lyberg, J.T. Massey, W.L. II, Nichols, and J. Waksberg. (pp. 257–272). New York: John Wiley and Sons.

Oksenberg, L., L. Coleman, and C. Cannell. 1986. "Interviewers' Voices and Refusal Rates in Telephone Surveys." *Public Opinion Quarterly* 50: 97–111. Doi: http://dx.doi.org/10.1086/268962.

O'Muircheartaigh, C. and P. Campanelli. 1999. "A Multilevel Exploration of the Role of Interviewers in Survey Nonresponse." *Journal of the Royal Statistical Society Series A* 162: 437–446. Doi: http://dx.doi.org/10.1111/1467-985X.00147.

Rosenberg, S., C. Nelson, and P.S. Vivekananthan. 1968. "A Multidimensional Approach to the Structure of Personality Impressions." *Journal of Personality and Social Psychology* 9: 283–294. Doi: http://dx.doi.org/10.1037/h0026086.

Schober, M.F. and F.G. Conrad. 1997. "Does Conversational Interviewing Reduce Survey Measurement Error?" *Public Opinion Quarterly* 61: 576–602.

Sharf, D.J., and M.E. Lehman. 1984. "Relationship Between the Speech Characteristics and Effectiveness of Telephone Interviewers." *Journal of Phonetics* 12: 219–228.

Snijkers, G., J. Hox, and E.D. de Leeuw. 1999. "Interviewers' Tactics for Fighting Survey Nonresponse." *Journal of Official Statistics* 15: 185–198.

Steinkopf, L., G. Bauer, and H. Best. 2010. "Nonresponse in CATI-Surveys." *Methods, Data, and Analysis* 4: 3–26. Available at: http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/mda/Vol.4_Heft_1/01_Best.pdf (accessed September 29, 2015).

Teitler, J.O., N.E. Reichman, and S. Sprachman. 2003. "Costs and Benefits of Improving Response Rates for a Hard-to-Reach Population." *Public Opinion Quarterly* 67: 126–138.

United States Census Bureau. 2010. "Educational Attainment." Available at: http://www.census.gov/hhes/socdemo/education/data/cps/2010/tables.html.

Van der Vaart, W., Y. Ongena, A. Hoogendoorn, and W. Dijkstra. 2005. "Do Interviewers' Voice Characteristics Influence Cooperation Rates in Telephone Surveys?" *International Journal of Opinion Research* 18: 488–499. Doi: http://dx.doi.org/10.1093/ijpor/edh117.

Received November 2012
Revised February 2015
Accepted March 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 627–647, http://dx.doi.org/10.1515/JOS-2015-0038

Quarterly Regional GDP Flash Estimates by Means of Benchmarking and Chain Linking

Ángel Cuevas[1], *Enrique M. Quilis*[2], *and Antoni Espasa*[3]

In this article we propose a methodology for estimating the GDP of a country's different regions, providing quarterly profiles for the annual official observed data. Thus the article offers a new instrument for short-term monitoring that allows the analysts to quantify the degree of synchronicity among regional business cycles. Technically, we combine time-series models with benchmarking methods to process short-term quarterly indicators and to estimate quarterly regional GDPs ensuring their temporal and transversal consistency with the National Accounts data. The methodology addresses the issue of nonadditivity, explicitly taking into account the transversal constraints imposed by the chain-linked volume indexes used by the National Accounts, and provides an efficient combination of structural as well as short-term information. The methodology is illustrated by an application to the Spanish economy, providing real-time quarterly GDP estimates, that is, with a minimum compilation delay with respect to the national quarterly GDP. The estimated quarterly data are used to assess the existence of cycles shared among the Spanish regions.

Key words: Benchmarking; chain linking; national accounts; regional accounts; GDP flash estimates.

1. Introduction

Business cycle analysis and the short-term monitoring of a national economy can be substantially improved if an explicit regional dimension is taken into consideration. In this way, the diffusion of the aggregate (or national) cycle can be analyzed in detail: identifying leading/coincident/lagged regions, detecting common and specific shocks and so on. The relevance of this added geographical dimension is especially important both for large or medium-sized countries as well as for countries with decentralized systems that allow specific economic policies to be applied. Of course, the quarterly regional estimates that we present below are also very useful for regional governments.

The Regional Accounts (RA) are annual data and in this context we here propose a methodology for estimating quarterly Gross Domestic Product (GDP) time series at the regional level, providing a new instrument for short-term monitoring that allows us to gauge the degree of synchronicity and the identification of shared and idiosyncratic shocks to different regions.

[1] Macroeconomic Research Department, Independent Authority for Fiscal Responsibility, José Abascal n° 2, 2ª planta 28003, Madrid, Spain. Email: angel.cuevas@airef.es
[2] Macroeconomic Research Department, Independent Authority for Fiscal Responsibility, José Abascal n° 2, 2ª planta 28003, Madrid, Spain. Email: enrique.quilis@airef.es
[3] Department of Statistics and Instituto Flores de Lemus, Universidad, Carlos III de Madrid, Calle Madrid 126, 28903 Madrid, Spain. Email: antoni.espasa@uc3m.es

Our methodology ensures the consistency of these quarterly regional GDPs with the national quarterly GDP, taking into account the chain-linking procedures that underlie its compilation. Note that the same principles of volume estimation using chain-linked indices have been used in our analysis and we have applied the same procedures of seasonal and calendar adjustment used by the Quarterly National Accounts (QNA).

Structural consistency is also ensured, since the quarterly regional GDPs are consistent with their annual RA counterparts. The fact that both QNA and RA share the same National Accounts (NA) framework provides the base for the consistency obtained in our analysis. In this way, we can use the quarterly regional estimates to derive structural measures at the regional level.

The modeling approach is highly reliant on a set of regional high-frequency indicators. These indicators provide the ultimate basis used by the model to generate GDP according to time-series techniques ranging from univariate ARIMA models to multivariate dynamic-factor models. The set of indicators and models are homogeneous across regions, ensuring the comparability of the results.

The methodology has three main stages:

1. Processing of the high-frequency indicators available at the regional level and estimation, for each region, of a synthetic index that combines the available short-term information.
2. Temporal disaggregation and interpolation of annual regional GDP using the indicators processed in Step 1.
3. Balancing of these initial quarterly estimates in order to ensure transversal consistency with national quarterly GDP, at the same time preserving the temporal consistency achieved in the previous stage.

It is worth emphasizing that, from an operational perspective, early estimates of quarterly regional GDPs may be available with a minimum delay with respect to the national quarterly GDP release, the so-called "GDP flash estimate". Thus the national figure may have timely regional counterparties, enhancing the informational content of analysis carried out at the aggregate level.

The main contributions of our article are:

- A methodology for obtaining quarterly estimates of GDP for all the regions in a country, derived in a consistent way with the official available data provided by the NA, both RA and QNA.
- Early (or flash) estimates of quarterly GDP at the regional level that may be released at the same time as the national GDP.
- Transversal consistency is compliant with the chain-linking methodology, circumventing its nonadditive features in the balancing step.

The article is organized as follows. The second section outlines the modeling approach, going into detail on its main steps. A complete and in-depth application of the methodology using Spanish data appears in section three. Finally, in the fourth section, we present our conclusions and future lines of research.

2. Modeling Approach

In this section we present the main steps of the proposed methodology. The modeling approach consists of three basic steps: (i) seasonal adjustment of regional short-term raw indicators and construction of synthetic indicators for each region by means of factor analysis, (ii) initial quarterly estimates of regional GDP provided by benchmarking and (iii) enforcement of the transversal constraint that links the regional quarterly GDPs with their national counterpart.

This aggregation constraint must be consistent with the chain-linking procedure used to compile quarterly GDP at the national level, dealing with the nonadditivity issue in an appropriate way. We now turn to examine the three stages in more detail; however, to simplify the exposition, we first present the required information set.

2.1. Information Set

The model requires as input three elements that vary according to their sampling frequency (annual or quarterly), their spatial coverage (regional or national) and their method of compilation (NA or short-term indicators).

The variables of the system are: regional GDPs (y), national GDP (z), and regional short-term indicators in their original or raw form (xr). Upper-case letters refer to annual variables, while lower-case letters refer to quarterly variables. Let $T = 1,..,N$ be the annual (low-frequency) index, $t = 1,..,4$ the quarterly index within a natural year and $j = 1,..,M$ the regional (cross-section) index.

Hence, $Y = \{Y_{T,j}: T = 1,..,N; j = 1,..,M\}$ is a NxM matrix comprising the annual regional GDPs that play the role of temporal benchmarks of the system. Aggregation of the regional GDPs generates the GDP at the national level. Note that, aggregation is performed according to the chain-linking methodology.

Variable z is a $nx1$ vector comprising the observed quarterly GDP provided by the QNA, being n the number of quarterly observations satisfying $n \geq 4N$. This figure is available more timely than the regional data and shares with them the corresponding annual GDP volume index:

$$Z_T = \frac{1}{4}\sum_{t \in T} z_{t,T} \tag{1}$$

For example, taking 2011 as a reference, the QNA released its first estimate of 2010:Q4 on February, 11 while the RA released its first estimate of 2010 on March, 24. Both estimates share the annual figure for 2010 implicitly provided by the QNA by means of temporal aggregation of the four quarters of 2010.

Finally, xr is a nxM matrix comprising the observed raw quarterly indicators that operate as high-frequency proxies for the regional aggregates Y. As will be explained later, we work with the seasonally and calendar-adjusted indicators (x) instead of the raw indicators (xr).

Only the indicators x are observed at the three dimensions of the system: T (annual index), t (quarterly index) and j (regional index). Therefore, they provide the interpolation basis for Y (across the quarterly dimension t) and z (across the regional dimension j).

Table 1. Information set of the model: Quarterly GDP tracker (x), Annual Regional GDP (Y), quarterly National GDD (Z) and Quarterly Regional GDP (y), which is the variable to be estimated.

		Region 1			Region 2			Nation
Year	Quarter	x_1	y_1	Y_1	x_2	y_2	Y_2	Z
1	1	$x_{1,1,1}$	$y_{1,1,1}$	$\mathbf{Y_{1,1}}$	$x_{2,1,1}$	$y_{2,1,1}$	$\mathbf{Y_{2,1}}$	$\mathbf{z_{2,1,1}}$
	2	$x_{1,2,1}$	$y_{1,2,1}$		$x_{2,2,1}$	$y_{2,2,1}$		$\mathbf{z_{2,2,1}}$
	3	$x_{1,3,1}$	$y_{1,3,1}$		$x_{2,3,1}$	$y_{2,3,1}$		$\mathbf{z_{2,3,1}}$
	4	$x_{1,4,1}$	$y_{1,4,1}$		$x_{2.4,1}$	$y_{2,4,1}$		$\mathbf{z_{2,4,1}}$
2	1	$x_{1,1,2}$	$y_{1,1,2}$		$x_{2,1,2}$	$y_{2,1,2}$		$\mathbf{z_{2,1,2}}$
	2	$x_{1,2,2}$	$y_{1,2,2}$		$x_{2,2,2}$	$y_{2,2,2}$		$\mathbf{z_{2,2,2}}$
	3							
	4							

Note: bold variables are temporal constraints (Y) or transversal constraints (z).

In other words, our objective is to estimate y using x as interpolators and consistently with both Y and z.

Table 1 sets out the relationship among the inputs (Y, z and x) and the output (y) of the system for a simplified case with two regions ($M = 2$) and two years ($T = 2$). The first year is complete while the second year is incomplete (i.e., the last two quarters are not available for x and z and the annual figure for Y is not available either).

In this simplified example, we want to estimate the first year's quarterly regional GDPs $y_{j,t,1}$ consistently with their annual counterparts $Y_{j,1}$ and satisfying the transversal constraint that links the regional GDPs with the national GDP $z_{t,1}$ each quarter. The annual constraints do not apply during the second year since $Y_{j,2}$ are not available. Thus the only binding constraint is the transversal constraint.

2.2. *Processing Short-Term Indicators*

Typically, short-term regional economic indicators are compiled in raw form by the statistical agencies. However, the volume GDP used for short-term monitoring at the national level is calculated in two ways: using raw indicators or using seasonal and calendar-adjusted indicators. Since seasonal and calendar effects could be quite different between indicators and the macroeconomic aggregates, the second procedure for the calculation of the GDP seems more reliable. Usually these GDP figures are referred to as seasonal and calendar adjusted.

In order to ensure the homogeneity between both sources of information, regional raw indicators and seasonally adjusted quarterly national GDP, we apply an ARIMA model-based correction that filters out the raw data from seasonal and calendar effects, if they are present. The procedure has been implemented using the TRAMO-SEATS program, see Gómez and Maravall (1996) and Caporello and Maravall (2004). Formally:

$$x_{j,t,T} = V(B, F; \psi_j)\, xr_{j,t,T} \tag{2}$$

where $xr_{j,t,T}$ is the raw short-term indicator; $V()$ is the Wiener-Kolmogorov filter symmetrically defined on the backward and forward operators B and F and ψ_j are the parameters of the filter derived consistently with those of the ARIMA model for $xr_{j,t,T}$, see Gómez and Maravall (1998a, 1998b) for a detailed exposition of the model-based approach used by TRAMO-SEATS.

If the indicators are available at the monthly frequency, seasonal adjustment is performed on the monthly series. The resulting series are temporally aggregated to the quarterly frequency.

We have used TRAMO-SEATS because it is the method used by the Spanish National Statistical Institute (NSI) to adjust GDP to seasonal effects. Of course, the choice of the seasonal adjustment procedure depends on the official method used by the NSI to produce the GDP figures. In countries where X12-ARIMA is the official procedure, this should be also the choice for seasonally adjusting the short-term indicators.

In practice, several short-term economic indicators are used to monitor and estimate regional GDPs. These indicators are individually processed according to (2) and then linearly combined, producing a composite indicator that will be used as the high-frequency proxy for regional GDPs. As will be explained in the third section, we use factor analysis to estimate a synthetic indicator for each region because it provides an objective and simple way to combine the available indicators.

2.3. *Initial Quarterly Regional GDP Estimation*

Preliminary estimates of quarterly GDP at the regional level are compiled using benchmarking techniques (see Di Fonzo 1987, 2002 and Proietti 2006 for an in-depth exposition). These techniques play an important role in the compilation practices of QNA around the world (see Eurostat 1998 and Bloem et al. 2001).

We have considered several benchmarking procedures for deriving the preliminary GDP estimates: Chow and Lin (1971), Fernández (1981), Santos-Silva and Cardoso (2001) and Proietti (2006). All of them hinge upon a dynamic linear model that links the (observable) high-frequency indicator with the (unobservable) regional GDP. (To keep the notation simple we have omitted the regional index j).

$$y_t = \phi y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t \tag{3}$$

The innovation u follows an AR(1) process:

$$u_t = \rho u_{t-1} + a_t \tag{4}$$

Finally, the random shock that drives the innovation u is the Gaussian white-noise process:

$$a_t \sim iid\, N(0, v_a) \tag{5}$$

The model includes a temporal constraint that makes y quantitatively consistent with its annual counterpart Y:

$$Y = Cy \tag{6}$$

C is the temporal aggregation-extrapolation matrix defined as:

$$C = (I_N \otimes c | O_{N,n-sN}) \tag{7}$$

where N is the number of low-frequency observations, $\otimes$ stands for the Kronecker product,

Table 2. Benchmarking methods

	Parameter		
Method	ϕ	β_1	ρ
Chow-Lin	0	0	(0,1)
Fernández	0	0	1
Santos Silva-Cardos	(0,1)	0	0
Proietti	(0,1)	$\neq 0$	0

c is a row vector of size s which defines the type of temporal aggregation and s is the number of high-frequency data points for each low-frequency data point. If $c = [1,1, . . .,1]$ we would have the case of the temporal aggregation of a flow, if $c = [1/s,1/s, . . .,1/s]$ the case of the average of an index, and if $c = [0,0, . . .,1]$, an interpolation would be obtained. In our case, $s = 4$.

Extrapolation arises when $n > sN$. In this case, the problem can be solved easily by simply extending the temporal aggregation matrix by considering new columns of zeroes which do not distort the temporal aggregation relationship and that do not pose any difficulty to the inclusion of the last n-sN data points of the indicators in the process of estimating y.

The different benchmarking methods depend on the values of the parameters in (3) and (4) according to Table 2.

The methods of Chow-Lin and Fernández place the dynamics in the innovation, which may follow a stationary AR(1) process (Chow-Lin) or a nonstationary I(1), random-walk process (Fernández). Litterman (1983) proposes a methodology close to those of Chow-Lin and Fernández. However, the empirical and Monte Carlo evidence show that its performance is sometimes disappointing. This is due to the flatness of the implied likelihood profile and, therefore, the corresponding observational equivalence in a wide range of values for its dynamical parameter, see Proietti (2006). On the other hand, the methods of Santos Silva-Cardoso and Proietti place the dynamics in the variables y and x, treating the innovation as a purely random shock. Gregoir (1994) and Salazar et al. (1994) also propose methods in which the dynamics of y and x play an explicit role.

The estimation of the parameters and the unobserved time series y is performed by maximizing the implied log-likelihood profile of the low-frequency model. The low-frequency model incorporates the temporal aggregation constraints [2.6] and [2.7]. This optimization is performed by means of a grid search on the stationary domain of ϕ or ρ and pinning down the values of β and σ that maximize the log-likelihood function conditioned on the selected value for ϕ or ρ (see Bournay and Laroque 1979 for an in-depth exposition). The computations have been carried out using the functions written in Matlab by Abad and Quilis (2005).

2.4. *Balancing in a Chain-Linking Setting*

The estimates derived in the previous step do not verify the transversal constraint that should relate them to the national quarterly GDP, satisfying the same type of relationship that links annual regional GDPs and annual national GDP. We solve the problem by applying a multivariate balancing procedure, in particular a multivariate extension of the Denton (1971) method. This extension can be expressed in matrix form (as in Di Fonzo 1990 and Di Fonzo and Marini 2003), as well as in state-space form (see Proietti 2011). In this article we have adopted the former approach, using the functions written in Matlab by Abad and Quilis (2005).

This balancing method depends on the formulation of additive constraints. However, volume indexes compiled according to the chain-linking methodology are nonadditive, see Bloem et al. (2001) and Abad et al. (2007). Fortunately, we can transform the chain-linked measures in order to write them in an additive form and then use the powerful machinery of balancing procedures to ensure transversal and temporal consistency. Finally, we can express the results in the initial chain-linked format by reversing the transformation.

The constraint that links regional and national quarterly volume GDP is:

$$z_{t,T} = \left(\sum_{j} W_{j,T-1} \frac{y_{j,t,T}}{Y_{j,T-1}} \right) Z_{T-1} \tag{8}$$

where $z_{t,T}$ is the national quarterly volume GDP, $W_{j,T-1}$ is the weight of region j in year $T-1$ and $y_{j,t,T}$ is the quarterly volume GDP of the jth region. Weights are computed using GDPs valued at current prices, see Abad et al. (2007) for a complete derivation. Finally, Z_T and $Y_{j,T}$ are the annual counterparts $z_{t,T}$ of and $y_{j,t,T}$.

After some algebraic manipulations, we can express the constraint in additive form:

$$\underbrace{\frac{z_{t,T}}{Z_{T-1}}}_{r_{t,T}} = \underbrace{\sum_{j} W_{j,T-1} \frac{y_{j,t,T}}{Y_{j,T-1}}}_{wr_{j,t,T}} = \sum_{j} wr_{j,t,T} \tag{9}$$

In (9), the relationship between the national ratio $r_{t,T}$ and the weighted regional ratios $\mathrm{w}r_{j,t,T}$ is additive.

Plugging the initial estimates derived according to (3)–(7) into (9), we obtain the preliminary, unbalanced estimates:

$$wr^{*}_{j,t,T} = W_{j,T-1} \frac{\hat{y}_{j,t,T}}{Y_{j,T-1}} \tag{10}$$

The balanced and temporally consistent time series $\mathrm{w}r^{**}_{j,t,T}$ are the output of the following constrained quadratic optimization program:

$$\underset{wr^{*}}{MIN}\,(wr^{**} - wr^{*})' D' D (wr^{**} - wr^{*}) \quad s.t. \quad H\,wr^{**} = R_e \tag{11}$$

being:

$$H = \begin{bmatrix} 1^{‘}_{M} \otimes I_n \\ I_M \otimes C \end{bmatrix} \text{ and } R_e = \begin{bmatrix} z \\ WR \end{bmatrix}$$

where 1_M is a column vector of ones and WR is the annual counterpart of the weighted regional ratios written in matrix form.

In program (11), the objective function reflects the volatility of the discrepancies between the quarter-to-quarter growth rates of the balanced series and those of the unbalanced ones. After some mathematical manipulation, an explicit expression can be derived:

$$wr^{**} = wr^{*} + (D'D)^{-1}H'[H(D'D)^{-1}H']^{-1}(R_e - Hwr^{*}) \tag{12}$$

The interpretation of Equation (12) is straightforward: the quarterly balanced series are the result of adding a correction factor to the unbalanced series. This correction factor derives from the distribution of the discrepancy between the preliminary unbalanced estimates and the constraint series R_e.

Once we have obtained the consistent weighted ratios, we can reverse the transformation (9) to derive the final estimates of the quarterly regional GDP in volume terms:

$$y^{**}_{j,t,T} = wr^{**}_{j,t,T}\frac{Y_{j,T-1}}{W_{j,T-1}} \tag{13}$$

In this way, the estimates of quarterly GDP derived in the previous equation are quantitatively consistent in their time dimension (taking as benchmark their annual regional counterparts) and in their cross-section dimension (generating the GDP provided by the QNA by regional aggregation). We should also emphasize that the consistency extends to the methodological dimension too, since the chain-linking procedures currently used by the NA have been properly taken into account. Finally, using time-series methods to project the basic short-term indicators, we can derive nowcasts (or flash estimates) of regional quarterly GDP in a timely manner.

As a summary, Figure 1 presents a picture of the complete procedure. The diagram emphasizes the binding constraints and the homogeneous processing of information at the regional level. Note that the box labeled "balancing" embeds the dechaining and

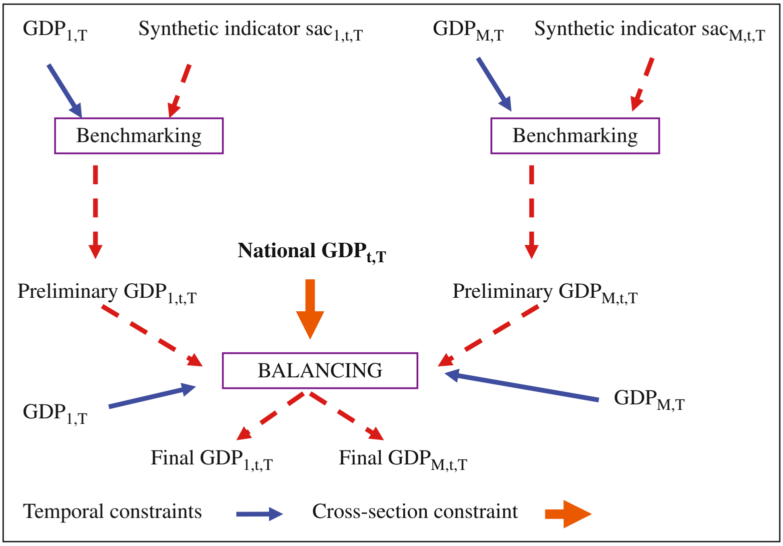


Fig. 1. Schedule of Steps 2 (Benchmarking) and 3 (Balancing). Note: national variables in bold. Quarterly index t goes from 1 to 4; annual index T goes from 1 to N and regional index j goes from 1 to M

rechaining steps required to circumvent the nonadditive features of the chain-linked volume indexes.

2.5. Comparison with Other Approaches

Table 3 compares our methodology with related approaches along six dimensions: high-frequency model, role of constraints (temporal and transversal), explicit consideration of chain linking, mixing data frequencies (e.g., annual and quarterly data) and computational approach.

Di Fonzo (1990) presents a methodology closely related to ours. We have expanded his approach to cope with the issue of chain linking and focus the results upon flash estimation and benchmarking. Di Fonzo and Marini (2005) may be considered a variant of Di Fonzo (1990) in which balancing plays also a critical role.

In addition, Proietti (2011) is also a close reference. He generalizes the Di Fonzo (1990) model to take into account integrated random-walk innovations and deals with the issue of nonadditivity posed by the chain-linking volume indexes implicitly, arranging the measurement equations to consider a statistical discrepancy. His computational approach relies on Kalman filtering of the state-space representation of the model. By contrast, our approach is matrix-oriented, following Di Fonzo (1990).

Spatial correlation plays an important role due to the fact that short-term regional indicators are closely related and the estimation of regional GDPs at the quarterly frequency depends also on the national quarterly GDP (Step 3: balancing).

However, our procedure is oriented towards the temporal disaggregation of regional aggregates, at the same time preserving the cross-section consistency with the national quarterly GDP rather than the spatial disaggregation of national totals taking the information contained in the regional indicators as the basis for interpolation. The last approach is used by the so-called spatial Chow-Lin procedure that adapts the Chow-Lin method to the spatial nature of the data and may be used to distribute a grand total into its spatial components at a given point in time (see Vidoli and Mazziotta 2012 and Polasek and Séllner 2010 among others). This procedure is very flexible and can be used to disaggregate national, regional or provincial totals into their spatial components (regions, provinces or areas), but does not consider explicitly the temporal constraints that are the hallmark of the NA, both regional and quarterly, and of our procedure.

Finally, we want to emphasize that our approach is focused on the estimation of (unobservable) quarterly regional GDPs rather than on the forecasting of the (observable) annual regional GDPs. To ensure the comparability and homogeneity of those estimates, our procedure hinges upon the temporal and cross-section consistency in the same way as implemented in the NA. The reliance on mimicking the NA limits the selection of indicators as well as the modeling approach. Lehmann and Wohlrabe (2012) present a detailed forecasting exercise at the regional level, using a variety of models and a large set of indicators with different spatial coverage.

3. Case Study: A System of Flash Regional Quarterly GDP Estimates for Spain

In this section we present the main results of a system of regional quarterly GDP flash estimates for the Spanish economy, following the modeling approach previously outlined.

Table 3. Comparison with other methodological approaches

	Di Fonzo (1990)	Di Fonzo & Marini (2003)	Proietti (2011)	Ours
High-frequency model	Static model + I(1) innovations	Unspecified	Static model + I(1) or I(2) innovations	Static or dynamic model + AR(1)/I(1) innovations
Temporal constraints	Yes	Yes	Yes	Yes
Transversal constraints	Yes	Yes	Yes	Yes
Chain- linking constraints	No	No	No	Yes
Mixing frequencies	Yes	Yes	Yes	Yes
Computational approach	Matrix oriented	Matrix oriented	State space	Matrix oriented

3.1. Selection of Monthly Regional Indicators

This subsection details the indicators that have been selected for model estimation. The selection process was carried out under the premise that indicators should be available in a timely fashion and should provide a synthetic measure of each of the regional economies.

The criterion for choosing these variables is the consideration of the regional counterpart of all the indicators used in the compilation of the QNA (see Álvarez 1989, Martínez and Melis 1989, INE 1993 and Álvarez 2005). To fulfil this goal, we have prepared a set of monthly regional indicators that provides a fairly comprehensive basis for analyzing and monitoring GDP at the regional level. This set offers a high-frequency approximation to the behavior of the main macroeconomic aggregates: gross added value (industry, construction, and services), consumption, external trade and employment. The selected indicators, with a brief description of them, are:

- IPI: Index of Industrial Production.
  - Units: Index number.
  - Source: National Statistical Institute (*Instituto Nacional de Estadística, INE*).
  - Starting date: 1995.01.
  - Back-calculation: combining data from 1990 base (1995.01–2002.01) and 2005 base (2002.01–2011.12), using the oldest period-on-period rates of growth to retropolate the newest base.
- LIC: Municipal construction licenses. Total area to build.
  - Units: square meters.
  - Source: Ministry of Public Works (*Ministerio de Fomento*).
  - Starting date: 1995.01.
  - Back-calculation: Data for Basque Country (País Vasco) during the period 1995.01–1997.12 have been back calculated using the average of the remaining regions as indicator. Some specific missing data (Basque Country -2008.08- and Navarra -2009.12-) have been interpolated using the program TRAMO.
- PER: Overnight stays in hotel establishments.
  - Units: Number of overnight stays.
  - Source: National Statistical Institute (*Instituto Nacional de Estadística, INE*).
  - Starting date: 1995.01.
  - Back-calculation: The series have been homogenized since 1998.12 by means of univariate intervention analysis in order to correct the methodological change introduced in 1999.01.
- IAS: Services sector activity indicator.
  - Units: Index number. Valuation at current prices.
  - Source: National Statistical Institute (*Instituto Nacional de Estadística, INE*).
  - Starting date: 2005.01.
  - Deflated using the Consumer Price Index (CPI) for services (house rentals excluded).
  - Missing data since 1995.01 have been estimated using the static factor derived from the indicators that start in 1995.01 as regressor.

- ICM: Retail sales index.
  - Units: Index number. Valuation at current prices, gas stations excluded.
  - Source: National Statistical Institute (*Instituto Nacional de Estadística, INE*).
  - Starting date: 2001.01.
  - Deflated using the CPI for services (house rentals excluded).
  - Missing data since 1995.01 have been estimated using the static factor derived from the indicators that start in 1995.01 as regressor.
- MAT: Car registrations.
  - Units: Registrations.
  - Source: Traffic department (*Dirección General de Tráfico, Ministerio del Interior*).
  - Starting date: 1995.01.
- EXP: Exports of goods.
  - Units: Euros, valuation at current prices.
  - Source: External trade statistics, Ministry of Economy and Competitiveness.
  - Starting date: 1995.01.
  - Deflated using the national exports unit value index.
- IMP: Imports of goods.
  - Units: Euros, valuation at current prices.
  - Source: External trade statistics, Ministry of Economy and Competitiveness.
  - Starting date: 1995.01.
  - Deflated using the national imports unit value index.
- AFI: Social security system: registered workers.
  - Units: persons.
  - Source: Labor department (*Ministerio de Empleo y Seguridad Social*).
  - Starting date: 1995.01.

The short-term indicators, in order to be consistent with the QNA data (as mentioned in Section 2), have been seasonally and calendar adjusted.

3.2. Regional Synthetic Indexes

To combine the information contained in the individual monthly indicators in an efficient and operative way, we have calculated a synthetic indicator for each region. In order to convey an idea of the correlation between the individual indicators and the estimated synthetic indicator (common factor), Table 4 shows the loading vectors, estimated by means of principal components factor analysis.

We have to note how loadings vary depending on the predominant activities in which each region specializes. Since two of the indicators (IAS and ICM) have been completed using the common factor estimated from the remaining indicators, their correlations with the common factor estimated with the balanced panel are overestimated to a certain extent. This fact complicates the exact quantification of their role. However, their economic relevance (IAS for the whole services sector and ICM for private consumption) recommends their inclusion in the estimation of the regional GDP trackers.

The corresponding monthly regional synthetic indicators are temporally aggregated to the quarterly frequency.

Table 4. Regional synthetic indexes: loading structure

	AFI	EXP	IMP	IPI	LIC	MAT	PER	ICM	IAS
Andalucía (AND)	0.54	0.28	0.05	0.45	0.01	0.77	0.21	0.73	0.90
Aragón (ARA)	0.31	0.63	0.29	0.79	0.04	0.63	0.01	0.51	0.65
Asturias (AST)	0.42	0.41	0.25	0.31	0.17	0.63	0.25	0.74	0.87
Baleares (BAL)	0.29	0.24	0.19	0.33	0.09	0.74	0.07	0.37	0.78
Canarias (CAN)	0.63	0.01	0.01	0.50	0.10	0.54	0.23	0.78	0.84
Cantabria (CANT)	0.35	0.56	0.36	0.57	0.07	0.56	0.06	0.06	0.74
Castilla La Mancha (CLM)	0.50	0.32	0.31	0.57	0.39	0.48	0.03	0.69	0.88
Castilla León (CYL)	0.31	0.55	0.50	0.61	0.01	0.68	0.01	0.08	0.82
Cataluña (CAT)	0.38	0.62	0.45	0.77	0.12	0.69	0.01	0.68	0.90
Extremadura (EXT)	0.41	0.30	0.14	0.14	0.30	0.75	0.01	0.42	0.76
Galicia (GAL)	0.32	0.62	0.24	0.45	0.01	0.70	0.23	0.66	0.89
Madrid (MAD)	0.41	0.43	0.32	0.62	0.01	0.48	0.28	0.75	0.69
Murcia (MUR)	0.45	0.24	0.01	0.37	0.04	0.76	0.19	0.79	0.87
Navarra (NAV)	0.35	0.61	0.54	0.72	0.01	0.32	0.21	0.09	0.64
País Vasco (PV)	0.14	0.58	0.49	0.76	0.08	0.57	0.01	0.62	0.86
La Rioja (RIO)	0.18	0.66	0.44	0.54	0.25	0.43	0.19	0.67	0.88
Valencia (VAL)	0.43	0.53	0.25	0.75	0.06	0.64	0.01	0.72	0.91

3.3. National Accounts Data: Regional Accounts and Quarterly National Accounts

Apart from the monthly regional indicators mentioned above, regional annual GDPs in chained-volume indices are provided by the RA according to ESA-95 conventions and they are available for the time span 1995–2011. The cross-section dimension includes 17 regions (*Comunidades Autónomas*) plus two autonomous cities that will be jointly considered, giving M = 18, a NUTS-2 regional breakdown according to Eurostat's classification.

Finally, the quarterly transversal constraint is the Spanish quarterly volume GDP provided by the QNA. This variable is compiled seasonally and calendar adjusted.

3.4. Empirical Results

Using the abovementioned data for the period 1995.01 – 2012.12 we can compare now the final results obtained using the different benchmarking techniques mentioned in section two (Fernandez, Chow-Lin, Santos Silva-Cardoso (SSC for brevity), Proportional Denton and Proietti) in order to select the most appropriate in terms of correlation and volatility.

Table 5 shows the summary results obtained with the different methods. Starting with the composite indicators derived by factor analysis for each region in the first stage, we apply different benchmarking methods and compare the different results obtained after final balancing. In order to summarize the results, we present the average correlation of the quarterly growth rate of GDP finally estimated by region with the initial composite indicator and the average standard deviation of the quarterly growth rate of GDP finally estimated by region.

This table shows that there seems to be a trade-off relationship between correlation and volatility (except in proportional Denton, which shows high volatility and low correlation). The Fernández and Chow-Lin methods are closest to the evolution of the indicator, without assuming a more complex structure in the errors, as is the case with SSC and Proietti.

Based on these results, we have decided to choose either the Fernández or the Chow-Lin method, because we think it is more important to be as faithful as possible to the information contained in the indicators, despite having higher volatility. Additionally, this is the method currently suggested for the compilation of the Spanish QNA (see Quilis 2005).

Regarding the distinction between the Fernández or Chow-Lin method, the results of the exercise show an innovational parameter with Chow-Lin close to 1 (approximately 0.98–0.99 in most cases), so under this situation both methods are practically equivalent.

With the aim of analyzing both the duration and the date of entry and exit of the recession in each region, Table 6 presents the evolution of the estimated year-on-year rates

Table 5. Comparison of methods (quarterly rates of growth)

	Fernandez	Chow-Lin	SSC	Denton Prop.	Proietti
Average Standard Deviation	0.821	0.858	0.731	0.843	0.744
Average Correlation	0.767	0.776	0.683	0.670	0.736

Table 6. Dating recession in quarterly GDP (year-on-year rates of growth)

	2008				2009				2010				2011				2012				
	T I	T II	T III	T IV	T I	T II	T III	T IV	T I	T II	T III	T IV	T I	T II	T III	T IV	T I	T II	T III	T IV	
Spain	2.7	1.9	0.3	–1.4	–3.4	–4.4	–4.0	–3.1	–1.5	–0.2	0.0	0.4	0.5	0.5	0.6	0.0	–0.7	–1.4	–1.6	–1.5	**Spain**
Andalucía	2.9	1.7	–0.4	–1.8	–2.8	–3.8	–3.8	–3.4	–2.3	–1.0	–0.5	–0.1	–0.2	–0.4	–0.1	0.2	–0.2	–0.7	–1.4	–1.7	**Andalucía**
Aragón	3.5	2.2	0.7	–2.9	–4.4	–4.9	–4.8	–1.8	–1.3	–1.5	–0.5	0.1	–0.1	1.0	0.8	–1.5	–0.7	–2.1	–2.0	–0.2	**Aragón**
Asturias	2.9	2.3	0.2	–1.0	–3.9	–5.6	–5.8	–4.6	–2.0	–0.7	0.0	0.3	0.1	0.4	0.3	–0.6	–1.3	–1.9	–2.0	–2.1	**Asturias**
Baleares	2.6	2.3	0.7	–0.5	–2.1	–5.1	–4.3	–3.8	–2.5	–0.9	–0.6	–0.8	–0.6	2.2	2.5	2.0	1.3	–1.1	–0.9	0.1	**Baleares**
Canarias	1.6	1.4	–0.3	–1.5	–3.0	–4.5	–4.7	–4.5	–3.1	–2.0	1.1	1.4	2.3	2.8	1.2	1.1	–0.3	–1.0	–2.0	–0.6	**Canarias**
Cantabria	2.1	1.9	0.7	–0.5	–2.1	–3.7	–4.5	–4.1	–2.4	–1.3	–1.2	–0.6	–0.1	0.2	1.2	0.7	0.0	–0.6	–0.8	–0.1	**Cantabria**
Castilla La Mancha	3.8	2.5	0.6	–0.8	–2.8	–3.5	–4.4	–4.1	–3.1	–2.4	–0.4	–0.2	–0.2	0.4	–0.7	–0.5	–1.3	–1.7	–1.3	–1.2	**Castilla La Mancha**
Castilla León	3.2	2.1	0.5	–2.3	–3.2	–3.4	–3.3	–1.4	0.2	1.6	0.2	0.5	0.9	0.3	2.0	0.8	–0.3	–1.6	–1.9	–2.1	**Castilla León**
Cataluña	1.9	0.9	–0.5	–1.5	–3.7	–4.2	–3.9	–3.1	–1.0	0.2	0.7	1.0	0.6	0.5	0.9	0.0	–0.1	–0.6	–0.9	–0.6	**Cataluña**
Extremadura	4.4	3.6	0.5	–1.1	–3.1	–3.4	–3.0	–2.0	–0.8	0.0	–2.3	–1.3	–1.5	–1.1	1.4	–0.7	–1.0	–2.7	–2.9	–2.7	**Extremadura**
Galicia	3.6	2.1	1.1	–0.1	–2.2	–3.8	–3.8	–3.9	–1.9	0.5	0.3	0.7	0.7	0.3	–0.3	–0.8	–0.9	–1.5	–1.5	–1.9	**Galicia**
Madrid	2.4	2.0	0.6	–1.1	–2.7	–3.8	–2.4	–1.8	–0.5	0.4	–0.1	0.0	0.8	0.6	0.8	0.0	–1.7	–2.5	–2.8	–2.9	**Madrid**
Murcia	3.7	2.8	1.1	–1.2	–3.6	–5.4	–4.6	–4.7	–2.7	–0.6	–0.3	0.0	–0.5	–0.3	–0.3	–0.1	–0.3	–0.7	–1.7	–2.4	**Murcia**
Navarra	2.8	3.7	1.0	0.0	–3.8	–4.9	–3.3	–2.4	0.1	0.3	0.4	0.8	1.5	2.0	1.0	0.5	–1.3	–2.4	–2.0	–1.5	**Navarra**
País Vasco	2.6	2.5	1.1	–0.8	–3.4	–5.1	–4.7	–3.2	–0.8	0.9	1.2	1.4	1.7	1.4	0.8	0.2	–1.0	–1.6	–1.3	–1.0	**País Vasco**
La Rioja	3.8	2.3	0.7	–1.0	–3.8	–4.5	–5.2	–5.2	–2.8	–2.7	–1.8	–0.5	–0.7	0.8	1.5	1.3	0.6	0.0	–0.4	–1.0	**La Rioja**
Valencia	3.4	1.9	0.5	–2.7	–6.1	–7.0	–6.0	–4.4	–2.1	–0.2	–0.6	0.1	0.7	0.2	0.2	–0.8	–1.1	–1.3	–1.0	–1.1	**Valencia**

▨	Negative rates
🟥	Minimum rate
🟩	Positive rates

of growth in the quarterly frequency; for the exercise performed with the Chow-Lin method, for example.

The table shows how the crisis has affected regions unevenly. For example, we can place the bulk of the recession between the fourth quarter of 2008 and the first quarter of 2010. Most of the regions fell into recession at the same time but not all of them left it simultaneously; this is the case of regions such as Andalucía, where the contractionary period is particularly long. We can see that many regions fall back into recession after the first quarter of 2012.

In relation to the variance of these results, Figure 2 shows the different box plots of the year-on-year rates of growth in the quarterly frequency for the different regions:

We observe a greater presence of outliers in periods of recession than in periods of expansion. This is partly due to the longer duration of the latter, rendering the median less representative for recessionary quarters. At the same time, the highest rate of variability is not linked to the larger size (GDP weight) of the region (see Appendix 1).

The temporal dimension of the data allows us to appreciate a reduction in volatility after 2003, although this is a property inherited from the annual data published by the RA (see Figure 3):

Finally, in order to clarify the importance of the balancing procedure on the final estimate, an exercise on two regions has been carried out: one with a large size (Cataluña) and other with a small size (La Rioja). This exercise is trying to reveal whether a small region can seriously change its initial estimate of quarterly GDP with the final balancing.

Initial or preliminary estimates do not take into account the information contained in the national quarterly GDP. Those initial estimates are modified to be consistent each quarter with the quarterly national GDP, reflecting the fact that the national data is the transversal aggregation of the regions.

The difference between the initial and the final estimates reflects the balancing procedure that ensures the transversal constraint and preserves, for each region, the temporal consistency with the Regional Accounts.

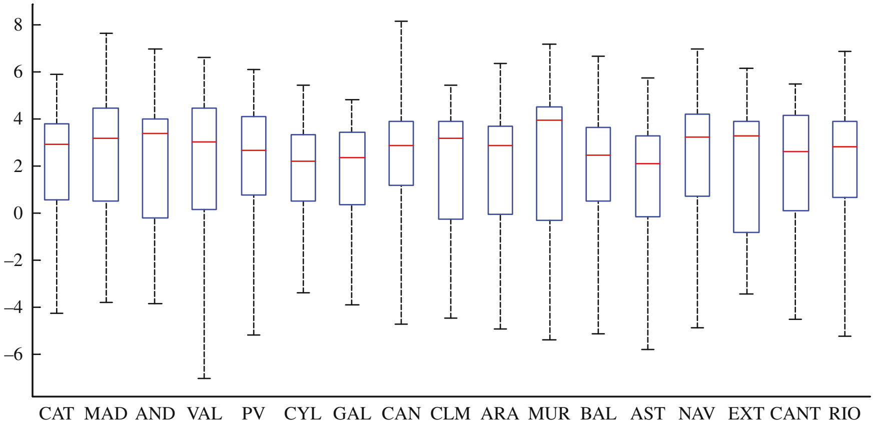


Fig. 2. Box plot: annual growth rates by region in quarterly frequency, sorted according to weight on Spanish GDP. Note: Central line stands for median values, the box represents 50% of the central part of the data and the whiskers are the minimum and maximum of the data

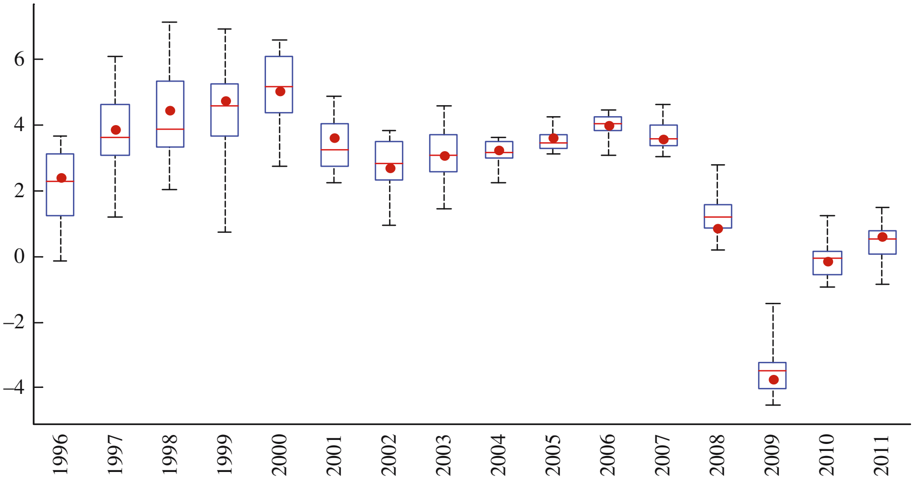


Fig. 3. Box plot: year-on-year rates of growth (annual data). Note: Dot is the aggregate data for Spain

Figure 4 shows, firstly, the initial quarterly regional GDP estimation (distribution of annual regional GDP according to the indicator) against the evolution of the indicator and, secondly, the initial quarterly estimation against the final quarterly GDP.

It is easy to see how the first step of estimating quarterly GDP depending on the evolution of the indicator is even more crucial to the subsequent balancing procedure. Furthermore, the small region does not have its initial estimate changed substantially compared with that of the large region. This fact shows the robustness of the balancing

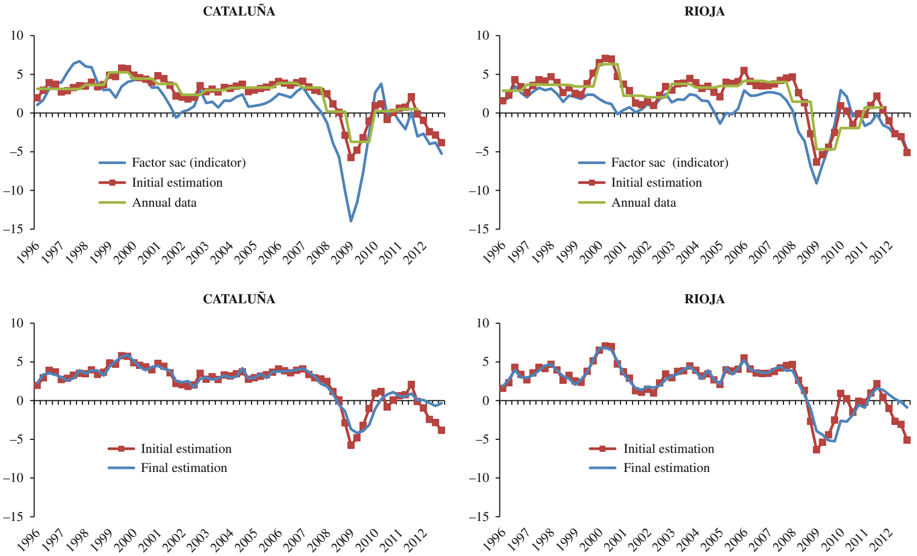


Fig. 4. Initial quarterly estimation vs. final balanced estimation. Small vs. large regions, year-on-year rates of growth

procedure, revealing that the variability in the final estimate is driven by the variability of the selected indicator.

4. Conclusions

In this article we have presented a feasible way to add a regional dimension to the short-term macroeconomic analysis, satisfying the temporal and cross-section constraints imposed by the NA. Our procedure generates results that are comparable across regions, are based on meaningful short-term information, and may be updated at the same time as the GDP flash national estimates, providing a solid basis for specific regional estimates.

In summary, the major outcomes of the model are:

- It solves the lack of quarterly GDP at the regional level, providing estimates consistent with the official available data published by the NA (RA and QNA). These estimates are a stand-alone product that may be used as input in regional econometric models.
- It provides a regional breakdown of the early estimates of the quarterly national volume GDP that may be released simultaneously, providing flash estimates at the regional level.

There are several promising lines of research that may broaden the scope of the article. The use of dynamic-factor models to estimate the regional high-frequency synthetic indexes may provide a more complete description of the economic conditions at the regional level.

The modeling approach can be extended easily to accommodate several types of extrapolations. For example, the transversal benchmark of the model (the national quarterly GDP) may be an official release made by the NSI or a forecast made by an analyst (e.g., the research department of an investment bank). In the latter case, we can combine these forecasts with the projected path for the underlying short-term quarterly regional indicators to generate the corresponding regional quarterly GDPs. The resulting conditional extrapolations can be used to assess the expected cyclical position of each region with respect to the nation.

Finally, the estimated regional quarterly GDPs can be used to analyze issues related to the synchronicity of the regional business cycles as well as their pattern of co-movements.

Appendix 1: Main Features of the Spanish Regions (2011)

	Population (thousand)	Population weight	GDP weight	Employment weight
Andalucía	8,270.5	17.9%	13.5%	14.7%
Aragón	1,315.5	2.9%	3.2%	3.1%
Asturias	1,054.5	2.3%	2.1%	2.1%
Baleares	1,092.5	2.4%	2.5%	2.6%
Canarias	2,107.0	4.6%	3.9%	4.1%
Cantabria	578.3	1.3%	1.2%	1.2%
Castilla La Mancha	2,045.4	4.4%	3.5%	3.9%
Castilla León	2,483.8	5.4%	5.3%	5.3%
Cataluña	7,303.1	15.8%	18.6%	17.8%
Extremadura	1,083.1	2.3%	1.6%	1.9%
Galicia	2,732.0	5.9%	5.3%	5.7%
Madrid	6,371.6	13.8%	18.0%	16.8%
Murcia	1,471.4	3.2%	2.6%	3.0%
Navarra	622.8	1.4%	1.7%	1.6%
País Vasco	2,127.9	4.6%	6.2%	5.3%
La Rioja	312.7	0.7%	0.8%	0.7%
Valencia	5,001.2	10.8%	9.5%	9.8%
Ceuta y Melilla	151.7	0.3%	0.3%	0.3%
Spain	46,125.0	100.0%	100.0%	100.0%

5. References

Abad, A. and E.M. Quilis. 2005. "Software to Perform Temporal Disaggregation of Economic Time Series." Eurostat, Working Papers and Series. Available at: http://ec.europa.eu/eurostat/documents/4187653/5774917/LN-SR012007-EN.PDF/c83eb69e-e3a9-4fdd-923d-76c09fea6f7b (accessed October 2015).

Abad, A., A. Cuevas, and E.M. Quilis. 2007. "Chain-Linked Volume Indexes: a Practical Guide." Universidad Carlos III de Madrid, Instituto Flores de Lemus, *Boletín de Inflación y Análisis Macroeconómico* 157: 72–85. Available at http://e-archivo.uc3m.es/handle/10016/20332#preview (accessed October 2015).

Álvarez, F. 1989. "Base Estadística en España de la Contabilidad Nacional Trimestral." *Revista Española de Economía* 6: 59–84.

Álvarez, R. 2005. "Notas Sobre Fuentes Estadísticas." In Servicio de Estudios del Banco de España, *El análisis de la economía española*, Alianza Editorial, Madrid, Spain.

Bloem, A.M., R.J. Dippelsman, and N.O. Mæhle. 2001. *Quarterly National Accounts Manual. Concepts, Data Sources, and Compilation*. International Monetary Fund. Available at: https://www.imf.org/external/pubs/ft/qna/2000/Textbook/ch1.pdf (accessed October 2015).

Bournay, J. and G. Laroque. 1979. "Réflexions sur la Méthode D'elaboration des Comptes Trimestriels." *Annales de l'INSEE* 36: 3–30. Available at: http://www.jstor.org/stable/20075332.

Caporello, G. and A. Maravall. 2004. “Program TSW. Revised Manual.” Bank of Spain, Occasional Paper no. 0408. http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/04/Fic/do0408e.pdf (accessed October 2015).

Chow, G. and A.L. Lin. 1971. “Best Linear Unbiased Distribution and Extrapolation of Economic Time Series by Related Series.” *Review of Economic and Statistics* 53: 372–375. Available at: http://www.jstor.org/stable/1928739.

Denton, F.T. 1971. “Adjustment of Monthly or Quarterly Series to Annual Totals: an Approach Based on Quadratic Minimization.” *Journal of the American Statistical Society* 66: 99–102. Doi: http://dx.doi.org/10.1080/01621459.1971.10482227.

Di Fonzo, T. 1987. *La Stima Indiretta di Serie Economiche Trimestrali*. Cleup Editore, Padova, Italy.

Di Fonzo, T. 1990. “The Estimation of *M* Disaggregate Time Series when Contemporaneous and Temporal Aggregates are Known.” *Review of Economics and Statistics* 72: 178–182. Doi: http://dx.doi.org/10.2307/2109758.

Di Fonzo, T. 2002. “Temporal Disaggregation of Economic Time Series: Towards a Dynamic Extension.” European Commission (Eurostat) Working Papers and Studies, Theme 1, General Statistics (pp. 41). Available at: http://ec.europa.eu/eurostat/documents/3888793/5816173/KS_AN-03-035-EN.PDF/21c4417c-dbec-45ec-b440-fe8bf95661b7?version=1.0 (accessed October 2015).

Di Fonzo, T. and M. Marini. 2003. “Benchmarking Systems of Seasonally Adjusted Time Series According to Denton’s Movement Preservation Principle.” Dipartimento di Scienze Statistiche, Università di Padova, Working Paper no. 2003–09. Available at: http://www.oecd.org/std/21778574.pdf, (accessed October 2015).

Eurostat. 1998. *Handbook of Quarterly National Accounts*. Luxembourg: Statistical Office of the EC.

Fernández, R.B. 1981. “Methodological Note on the Estimation of Time Series.” *Review of Economic and Statistics* 63: 471–478. Doi: http://dx.doi.org/10.2307/1924371.

Gómez, V. and A. Maravall. 1996. “Programs TRAMO and SEATS.” Bank of Spain, Working Paper no. 9628. Available at: http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/96/Fich/dt9628e.pdf (accessed October 2015).

Gómez, V. and A. Maravall (1998a) “Guide for using the programs TRAMO and SEATS”, Bank of Spain, Working Paper no. 9805. Available at: http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/98/Fic/dt9805e.pdf (accessed October 2015).

Gómez, V. and A. Maravall (1998b) “Automatic modeling methods for univariate series”, Bank of Spain, Working Paper no. 9808. Available at: http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/98/Fic/dt9808e.pdf (accessed October 2015).

Gregoir, S. 1994. “Propositions Pour une Désagrégation Temporelle Basée sur des Modèles Dynamiques Simples.” In *Workshop on Quarterly National Accounts*, ed. Eurostat. Luxembourg: Statistical Office of the EC. Available at: http://ec.europa.eu/eurostat/documents/3888793/5815741/KS-AN-03-014-EN.PDF/284f1001-fd36-4999-b007-a22033e8aaf9 (accessed October 2015).

INE. 1993. *Contabilidad Nacional Trimestral de España (CNTR). Metodología y serie trimestral 1970–1992*. Instituto Nacional de Estadística.

Litterman, R.B. 1983. "A random walk, Markov model for the distribution of time series." *Journal of Business and Economic Statistics* 1: 169–173. Available at: http://www.jstor.org/stable/1391858.

Lehmann, R. and K. Wohlrabe. 2012. "Forecasting GDP at the Regional Level with Many Predictors." CESIFO, Working Paper no. 3956. Available at: http://www-sre.wu.ac.at/ersa/ersaconfs/ersa13/ERSA2013_paper_00015.pdf (accessed October 2015).

Martínez, A. and F. Melis. 1989. "La Demanda y la Oferta de Estadísticas Coyunturales." *Revista Española de Economía* 6: 7–58.

Polasek, W. and R. Séllner. 2010. "Spatial Chow-Lin Methods for Data Completion in Econometric Flow Models." Institut für Höhere Studien (HIS), Economic Series no. 255. Available at: https://www.ihs.ac.at/publications/eco/es-255.pdf (accessed October 2015).

Proietti, T. 2006. "Temporal Disaggregation by State Space Methods: Dynamic Regression Methods Revisited." *Econometrics Journal* 9: 357–372. Doi: http://dx.doi.org/10.1111/j.1368-423X.2006.00189.

Proietti, T. 2011. "Multivariate Temporal Disaggregation with Cross-Sectional Constraints." *Journal of Applied Statistics* 38: 1455–1466. Doi: http://dx.doi.org/10.1080/02664763.2010.505952.

Quilis, E.M. 2005. "Benchmarking Techniques in the Spanish Quarterly National Accounts." European Commission, Working papers and studies (Eurostat-OECD Workshop on Frontiers in Benchmarking Techniques and Their Application to Official Statistics, Luxembourg, April 7–8, 2005). Available at: http://ec.europa.eu/eurostat/documents/4187653/5774917/LN-SR012007-EN.PDF/c83eb69e-e3a9-4fdd-923d-76c09fea6f7b (accessed October 2015).

Salazar, E., R. Smith, S. Wright, and M. Weale. 1994. "Indicators of Monthly National Accounts." In *Workshop on Quarterly National Accounts*, ed. Eurostat. Luxembourg: Statistical Office of the EC. Available at: http://ec.europa.eu/eurostat/documents/3888793/5815741/KS-AN-03-014-EN.PDF/284f1001-fd36-4999-b007-a22033e8aaf9 (accessed October 2015).

Santos-Silva, J.M.C. and F. Cardoso. 2001. "The Chow-Lin Method Using Dynamic Models." *Economic Modelling* 18: 269–280. Doi: http://dx.doi.org/10.1016/S0264-9993(00)00039-0.

Vidoli, F. and C. Mazziotta. 2012. "Spatial Composite and Disaggregate Indicators: Chow-Lin Methods and Applications." *Real Estate* 2: 9–19. Available at: http://fvidoli.weebly.com/uploads/2/3/0/8/23088460/eng_spatialcompositeanddisaggregate.pdf (accessed October 2015).

Received February 2013
Revised January 2015
Accepted January 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 627–647, http://dx.doi.org/10.1515/JOS-2015-0038

Quarterly Regional GDP Flash Estimates by Means of Benchmarking and Chain Linking

Ángel Cuevas[1], *Enrique M. Quilis*[2], *and Antoni Espasa*[3]

In this article we propose a methodology for estimating the GDP of a country's different regions, providing quarterly profiles for the annual official observed data. Thus the article offers a new instrument for short-term monitoring that allows the analysts to quantify the degree of synchronicity among regional business cycles. Technically, we combine time-series models with benchmarking methods to process short-term quarterly indicators and to estimate quarterly regional GDPs ensuring their temporal and transversal consistency with the National Accounts data. The methodology addresses the issue of nonadditivity, explicitly taking into account the transversal constraints imposed by the chain-linked volume indexes used by the National Accounts, and provides an efficient combination of structural as well as short-term information. The methodology is illustrated by an application to the Spanish economy, providing real-time quarterly GDP estimates, that is, with a minimum compilation delay with respect to the national quarterly GDP. The estimated quarterly data are used to assess the existence of cycles shared among the Spanish regions.

Key words: Benchmarking; chain linking; national accounts; regional accounts; GDP flash estimates.

1. Introduction

Business cycle analysis and the short-term monitoring of a national economy can be substantially improved if an explicit regional dimension is taken into consideration. In this way, the diffusion of the aggregate (or national) cycle can be analyzed in detail: identifying leading/coincident/lagged regions, detecting common and specific shocks and so on. The relevance of this added geographical dimension is especially important both for large or medium-sized countries as well as for countries with decentralized systems that allow specific economic policies to be applied. Of course, the quarterly regional estimates that we present below are also very useful for regional governments.

The Regional Accounts (RA) are annual data and in this context we here propose a methodology for estimating quarterly Gross Domestic Product (GDP) time series at the regional level, providing a new instrument for short-term monitoring that allows us to gauge the degree of synchronicity and the identification of shared and idiosyncratic shocks to different regions.

[1] Macroeconomic Research Department, Independent Authority for Fiscal Responsibility, José Abascal n° 2, 2ª planta 28003, Madrid, Spain. Email: angel.cuevas@airef.es
[2] Macroeconomic Research Department, Independent Authority for Fiscal Responsibility, José Abascal n° 2, 2ª planta 28003, Madrid, Spain. Email: enrique.quilis@airef.es
[3] Department of Statistics and Instituto Flores de Lemus, Universidad, Carlos III de Madrid, Calle Madrid 126, 28903 Madrid, Spain. Email: antoni.espasa@uc3m.es

Our methodology ensures the consistency of these quarterly regional GDPs with the national quarterly GDP, taking into account the chain-linking procedures that underlie its compilation. Note that the same principles of volume estimation using chain-linked indices have been used in our analysis and we have applied the same procedures of seasonal and calendar adjustment used by the Quarterly National Accounts (QNA).

Structural consistency is also ensured, since the quarterly regional GDPs are consistent with their annual RA counterparts. The fact that both QNA and RA share the same National Accounts (NA) framework provides the base for the consistency obtained in our analysis. In this way, we can use the quarterly regional estimates to derive structural measures at the regional level.

The modeling approach is highly reliant on a set of regional high-frequency indicators. These indicators provide the ultimate basis used by the model to generate GDP according to time-series techniques ranging from univariate ARIMA models to multivariate dynamic-factor models. The set of indicators and models are homogeneous across regions, ensuring the comparability of the results.

The methodology has three main stages:

1. Processing of the high-frequency indicators available at the regional level and estimation, for each region, of a synthetic index that combines the available short-term information.
2. Temporal disaggregation and interpolation of annual regional GDP using the indicators processed in Step 1.
3. Balancing of these initial quarterly estimates in order to ensure transversal consistency with national quarterly GDP, at the same time preserving the temporal consistency achieved in the previous stage.

It is worth emphasizing that, from an operational perspective, early estimates of quarterly regional GDPs may be available with a minimum delay with respect to the national quarterly GDP release, the so-called “GDP flash estimate”. Thus the national figure may have timely regional counterparties, enhancing the informational content of analysis carried out at the aggregate level.

The main contributions of our article are:

- A methodology for obtaining quarterly estimates of GDP for all the regions in a country, derived in a consistent way with the official available data provided by the NA, both RA and QNA.
- Early (or flash) estimates of quarterly GDP at the regional level that may be released at the same time as the national GDP.
- Transversal consistency is compliant with the chain-linking methodology, circumventing its nonadditive features in the balancing step.

The article is organized as follows. The second section outlines the modeling approach, going into detail on its main steps. A complete and in-depth application of the methodology using Spanish data appears in section three. Finally, in the fourth section, we present our conclusions and future lines of research.

2. Modeling Approach

In this section we present the main steps of the proposed methodology. The modeling approach consists of three basic steps: (i) seasonal adjustment of regional short-term raw indicators and construction of synthetic indicators for each region by means of factor analysis, (ii) initial quarterly estimates of regional GDP provided by benchmarking and (iii) enforcement of the transversal constraint that links the regional quarterly GDPs with their national counterpart.

This aggregation constraint must be consistent with the chain-linking procedure used to compile quarterly GDP at the national level, dealing with the nonadditivity issue in an appropriate way. We now turn to examine the three stages in more detail; however, to simplify the exposition, we first present the required information set.

2.1. Information Set

The model requires as input three elements that vary according to their sampling frequency (annual or quarterly), their spatial coverage (regional or national) and their method of compilation (NA or short-term indicators).

The variables of the system are: regional GDPs (y), national GDP (z), and regional short-term indicators in their original or raw form (xr). Upper-case letters refer to annual variables, while lower-case letters refer to quarterly variables. Let $T = 1,..,N$ be the annual (low-frequency) index, $t = 1,..,4$ the quarterly index within a natural year and $j = 1,..,M$ the regional (cross-section) index.

Hence, $Y = \{Y_{T,j}: T = 1,..,N; j = 1,..,M\}$ is a NxM matrix comprising the annual regional GDPs that play the role of temporal benchmarks of the system. Aggregation of the regional GDPs generates the GDP at the national level. Note that, aggregation is performed according to the chain-linking methodology.

Variable z is a $nx1$ vector comprising the observed quarterly GDP provided by the QNA, being n the number of quarterly observations satisfying $n \geq 4N$. This figure is available more timely than the regional data and shares with them the corresponding annual GDP volume index:

$$Z_T = \frac{1}{4}\sum_{t\in T} z_{t,T} \tag{1}$$

For example, taking 2011 as a reference, the QNA released its first estimate of 2010:Q4 on February, 11 while the RA released its first estimate of 2010 on March, 24. Both estimates share the annual figure for 2010 implicitly provided by the QNA by means of temporal aggregation of the four quarters of 2010.

Finally, xr is a nxM matrix comprising the observed raw quarterly indicators that operate as high-frequency proxies for the regional aggregates Y. As will be explained later, we work with the seasonally and calendar-adjusted indicators (x) instead of the raw indicators (xr).

Only the indicators x are observed at the three dimensions of the system: T (annual index), t (quarterly index) and j (regional index). Therefore, they provide the interpolation basis for Y (across the quarterly dimension t) and z (across the regional dimension j).

Table 1. Information set of the model: Quarterly GDP tracker (x), Annual Regional GDP (Y), quarterly National GDD (Z) and Quarterly Regional GDP (y), which is the variable to be estimated.

		Region 1			Region 2			Nation
Year	Quarter	x_1	y_1	Y_1	x_2	y_2	Y_2	Z
1	1	$x_{1,1,1}$	$y_{1,1,1}$	$\mathbf{Y_{1,1}}$	$x_{2,1,1}$	$y_{2,1,1}$	$\mathbf{Y_{2,1}}$	$\mathbf{z_{2,1,1}}$
	2	$x_{1,2,1}$	$y_{1,2,1}$		$x_{2,2,1}$	$y_{2,2,1}$		$\mathbf{z_{2,2,1}}$
	3	$x_{1,3,1}$	$y_{1,3,1}$		$x_{2,3,1}$	$y_{2,3,1}$		$\mathbf{z_{2,3,1}}$
	4	$x_{1,4,1}$	$y_{1,4,1}$		$x_{2.4,1}$	$y_{2,4,1}$		$\mathbf{z_{2,4,1}}$
2	1	$x_{1,1,2}$	$y_{1,1,2}$		$x_{2,1,2}$	$y_{2,1,2}$		$\mathbf{z_{2,1,2}}$
	2	$x_{1,2,2}$	$y_{1,2,2}$		$x_{2,2,2}$	$y_{2,2,2}$		$\mathbf{z_{2,2,2}}$
	3							
	4							

Note: bold variables are temporal constraints (Y) or transversal constraints (z).

In other words, our objective is to estimate y using x as interpolators and consistently with both Y and z.

Table 1 sets out the relationship among the inputs (Y, z and x) and the output (y) of the system for a simplified case with two regions ($M = 2$) and two years ($T = 2$). The first year is complete while the second year is incomplete (i.e., the last two quarters are not available for x and z and the annual figure for Y is not available either).

In this simplified example, we want to estimate the first year's quarterly regional GDPs $y_{j,t,1}$ consistently with their annual counterparts $Y_{j,1}$ and satisfying the transversal constraint that links the regional GDPs with the national GDP $z_{t,1}$ each quarter. The annual constraints do not apply during the second year since $Y_{j,2}$ are not available. Thus the only binding constraint is the transversal constraint.

2.2. *Processing Short-Term Indicators*

Typically, short-term regional economic indicators are compiled in raw form by the statistical agencies. However, the volume GDP used for short-term monitoring at the national level is calculated in two ways: using raw indicators or using seasonal and calendar-adjusted indicators. Since seasonal and calendar effects could be quite different between indicators and the macroeconomic aggregates, the second procedure for the calculation of the GDP seems more reliable. Usually these GDP figures are referred to as seasonal and calendar adjusted.

In order to ensure the homogeneity between both sources of information, regional raw indicators and seasonally adjusted quarterly national GDP, we apply an ARIMA model-based correction that filters out the raw data from seasonal and calendar effects, if they are present. The procedure has been implemented using the TRAMO-SEATS program, see Gómez and Maravall (1996) and Caporello and Maravall (2004). Formally:

$$x_{j,t,T} = V(B, F; \psi_j)\, xr_{j,t,T} \tag{2}$$

where $xr_{j,t,T}$ is the raw short-term indicator; *V()* is the Wiener-Kolmogorov filter symmetrically defined on the backward and forward operators B and F and ψ_j are the parameters of the filter derived consistently with those of the ARIMA model for $xr_{j,t,T}$, see Gómez and Maravall (1998a, 1998b) for a detailed exposition of the model-based approach used by TRAMO-SEATS.

If the indicators are available at the monthly frequency, seasonal adjustment is performed on the monthly series. The resulting series are temporally aggregated to the quarterly frequency.

We have used TRAMO-SEATS because it is the method used by the Spanish National Statistical Institute (NSI) to adjust GDP to seasonal effects. Of course, the choice of the seasonal adjustment procedure depends on the official method used by the NSI to produce the GDP figures. In countries where X12-ARIMA is the official procedure, this should be also the choice for seasonally adjusting the short-term indicators.

In practice, several short-term economic indicators are used to monitor and estimate regional GDPs. These indicators are individually processed according to (2) and then linearly combined, producing a composite indicator that will be used as the high-frequency proxy for regional GDPs. As will be explained in the third section, we use factor analysis to estimate a synthetic indicator for each region because it provides an objective and simple way to combine the available indicators.

2.3. *Initial Quarterly Regional GDP Estimation*

Preliminary estimates of quarterly GDP at the regional level are compiled using benchmarking techniques (see Di Fonzo 1987, 2002 and Proietti 2006 for an in-depth exposition). These techniques play an important role in the compilation practices of QNA around the world (see Eurostat 1998 and Bloem et al. 2001).

We have considered several benchmarking procedures for deriving the preliminary GDP estimates: Chow and Lin (1971), Fernández (1981), Santos-Silva and Cardoso (2001) and Proietti (2006). All of them hinge upon a dynamic linear model that links the (observable) high-frequency indicator with the (unobservable) regional GDP. (To keep the notation simple we have omitted the regional index j).

$$y_t = \phi y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t \tag{3}$$

The innovation u follows an AR(1) process:

$$u_t = \rho u_{t-1} + a_t \tag{4}$$

Finally, the random shock that drives the innovation u is the Gaussian white-noise process:

$$a_t \sim iid\, N(0, v_a) \tag{5}$$

The model includes a temporal constraint that makes y quantitatively consistent with its annual counterpart Y:

$$Y = Cy \tag{6}$$

C is the temporal aggregation-extrapolation matrix defined as:

$$C = (I_N \otimes c | O_{N,n-sN}) \tag{7}$$

where N is the number of low-frequency observations, $\otimes$ stands for the Kronecker product,

Table 2. Benchmarking methods

Method	Parameter ϕ	 β_1	 ρ
Chow-Lin	0	0	(0,1)
Fernández	0	0	1
Santos Silva-Cardos	(0,1)	0	0
Proietti	(0,1)	$\neq 0$	0

c is a row vector of size s which defines the type of temporal aggregation and s is the number of high-frequency data points for each low-frequency data point. If $c = [1,1,\ldots,1]$ we would have the case of the temporal aggregation of a flow, if $c = [1/s,1/s,\ldots,1/s]$ the case of the average of an index, and if $c = [0,0,\ldots,1]$, an interpolation would be obtained. In our case, $s = 4$.

Extrapolation arises when $n > sN$. In this case, the problem can be solved easily by simply extending the temporal aggregation matrix by considering new columns of zeroes which do not distort the temporal aggregation relationship and that do not pose any difficulty to the inclusion of the last $n\text{-}sN$ data points of the indicators in the process of estimating y.

The different benchmarking methods depend on the values of the parameters in (3) and (4) according to Table 2.

The methods of Chow-Lin and Fernández place the dynamics in the innovation, which may follow a stationary AR(1) process (Chow-Lin) or a nonstationary I(1), random-walk process (Fernández). Litterman (1983) proposes a methodology close to those of Chow-Lin and Fernández. However, the empirical and Monte Carlo evidence show that its performance is sometimes disappointing. This is due to the flatness of the implied likelihood profile and, therefore, the corresponding observational equivalence in a wide range of values for its dynamical parameter, see Proietti (2006). On the other hand, the methods of Santos Silva-Cardoso and Proietti place the dynamics in the variables y and x, treating the innovation as a purely random shock. Gregoir (1994) and Salazar et al. (1994) also propose methods in which the dynamics of y and x play an explicit role.

The estimation of the parameters and the unobserved time series y is performed by maximizing the implied log-likelihood profile of the low-frequency model. The low-frequency model incorporates the temporal aggregation constraints [2.6] and [2.7]. This optimization is performed by means of a grid search on the stationary domain of ϕ or ρ and pinning down the values of β and σ that maximize the log-likelihood function conditioned on the selected value for ϕ or ρ (see Bournay and Laroque 1979 for an in-depth exposition). The computations have been carried out using the functions written in Matlab by Abad and Quilis (2005).

2.4. Balancing in a Chain-Linking Setting

The estimates derived in the previous step do not verify the transversal constraint that should relate them to the national quarterly GDP, satisfying the same type of relationship that links annual regional GDPs and annual national GDP. We solve the problem by applying a multivariate balancing procedure, in particular a multivariate extension of the Denton (1971) method. This extension can be expressed in matrix form (as in Di Fonzo 1990 and Di Fonzo and Marini 2003), as well as in state-space form (see Proietti 2011). In this article we have adopted the former approach, using the functions written in Matlab by Abad and Quilis (2005).

This balancing method depends on the formulation of additive constraints. However, volume indexes compiled according to the chain-linking methodology are nonadditive, see Bloem et al. (2001) and Abad et al. (2007). Fortunately, we can transform the chain-linked measures in order to write them in an additive form and then use the powerful machinery of balancing procedures to ensure transversal and temporal consistency. Finally, we can express the results in the initial chain-linked format by reversing the transformation.

The constraint that links regional and national quarterly volume GDP is:

$$z_{t,T} = \left(\sum_j W_{j,T-1} \frac{y_{j,t,T}}{Y_{j,T-1}} \right) Z_{T-1} \tag{8}$$

where $z_{t,T}$ is the national quarterly volume GDP, $W_{j,T-1}$ is the weight of region j in year $T-1$ and $y_{j,t,T}$ is the quarterly volume GDP of the jth region. Weights are computed using GDPs valued at current prices, see Abad et al. (2007) for a complete derivation. Finally, Z_T and $Y_{j,T}$ are the annual counterparts $z_{t,T}$ of and $y_{j,t,T}$.

After some algebraic manipulations, we can express the constraint in additive form:

$$\underbrace{\frac{z_{t,T}}{Z_{T-1}}}_{r_{t,T}} = \underbrace{\sum_j W_{j,T-1} \frac{y_{j,t,T}}{Y_{j,T-1}}}_{wr_{j,t,T}} = \sum_j wr_{j,t,T} \tag{9}$$

In (9), the relationship between the national ratio $r_{t,T}$ and the weighted regional ratios $wr_{j,t,T}$ is additive.

Plugging the initial estimates derived according to (3)–(7) into (9), we obtain the preliminary, unbalanced estimates:

$$wr^*_{j,t,T} = W_{j,T-1} \frac{\hat{y}_{j,t,T}}{Y_{j,T-1}} \tag{10}$$

The balanced and temporally consistent time series $wr^{**}_{j,t,T}$ are the output of the following constrained quadratic optimization program:

$$\underset{wr^*}{MIN}\,(wr^{**} - wr^*)' D' D (wr^{**} - wr^*) \quad s.t. \quad H\,wr^{**} = R_e \tag{11}$$

being:

$$H = \begin{bmatrix} 1'_M \otimes I_n \\ I_M \otimes C \end{bmatrix} \text{ and } R_e = \begin{bmatrix} z \\ WR \end{bmatrix}$$

where 1_M is a column vector of ones and WR is the annual counterpart of the weighted regional ratios written in matrix form.

In program (11), the objective function reflects the volatility of the discrepancies between the quarter-to-quarter growth rates of the balanced series and those of the unbalanced ones. After some mathematical manipulation, an explicit expression can be derived:

$$wr^{**} = wr^{*} + (D'D)^{-1}H'[H(D'D)^{-1}H']^{-1}(R_e - Hwr^{*}) \tag{12}$$

The interpretation of Equation (12) is straightforward: the quarterly balanced series are the result of adding a correction factor to the unbalanced series. This correction factor derives from the distribution of the discrepancy between the preliminary unbalanced estimates and the constraint series R_e.

Once we have obtained the consistent weighted ratios, we can reverse the transformation (9) to derive the final estimates of the quarterly regional GDP in volume terms:

$$y^{**}_{j,t,T} = wr^{**}_{j,t,T}\frac{Y_{j,T-1}}{W_{j,T-1}} \tag{13}$$

In this way, the estimates of quarterly GDP derived in the previous equation are quantitatively consistent in their time dimension (taking as benchmark their annual regional counterparts) and in their cross-section dimension (generating the GDP provided by the QNA by regional aggregation). We should also emphasize that the consistency extends to the methodological dimension too, since the chain-linking procedures currently used by the NA have been properly taken into account. Finally, using time-series methods to project the basic short-term indicators, we can derive nowcasts (or flash estimates) of regional quarterly GDP in a timely manner.

As a summary, Figure 1 presents a picture of the complete procedure. The diagram emphasizes the binding constraints and the homogeneous processing of information at the regional level. Note that the box labeled "balancing" embeds the dechaining and

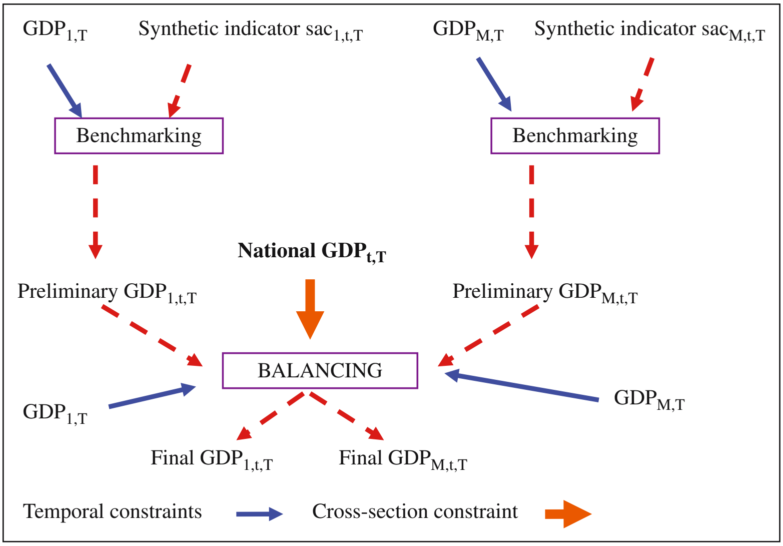


Fig. 1. Schedule of Steps 2 (Benchmarking) and 3 (Balancing). Note: national variables in bold. Quarterly index t goes from 1 to 4; annual index T goes from 1 to N and regional index j goes from 1 to M

rechaining steps required to circumvent the nonadditive features of the chain-linked volume indexes.

2.5. *Comparison with Other Approaches*

Table 3 compares our methodology with related approaches along six dimensions: high-frequency model, role of constraints (temporal and transversal), explicit consideration of chain linking, mixing data frequencies (e.g., annual and quarterly data) and computational approach.

Di Fonzo (1990) presents a methodology closely related to ours. We have expanded his approach to cope with the issue of chain linking and focus the results upon flash estimation and benchmarking. Di Fonzo and Marini (2005) may be considered a variant of Di Fonzo (1990) in which balancing plays also a critical role.

In addition, Proietti (2011) is also a close reference. He generalizes the Di Fonzo (1990) model to take into account integrated random-walk innovations and deals with the issue of nonadditivity posed by the chain-linking volume indexes implicitly, arranging the measurement equations to consider a statistical discrepancy. His computational approach relies on Kalman filtering of the state-space representation of the model. By contrast, our approach is matrix-oriented, following Di Fonzo (1990).

Spatial correlation plays an important role due to the fact that short-term regional indicators are closely related and the estimation of regional GDPs at the quarterly frequency depends also on the national quarterly GDP (Step 3: balancing).

However, our procedure is oriented towards the temporal disaggregation of regional aggregates, at the same time preserving the cross-section consistency with the national quarterly GDP rather than the spatial disaggregation of national totals taking the information contained in the regional indicators as the basis for interpolation. The last approach is used by the so-called spatial Chow-Lin procedure that adapts the Chow-Lin method to the spatial nature of the data and may be used to distribute a grand total into its spatial components at a given point in time (see Vidoli and Mazziotta 2012 and Polasek and Séllner 2010 among others). This procedure is very flexible and can be used to disaggregate national, regional or provincial totals into their spatial components (regions, provinces or areas), but does not consider explicitly the temporal constraints that are the hallmark of the NA, both regional and quarterly, and of our procedure.

Finally, we want to emphasize that our approach is focused on the estimation of (unobservable) quarterly regional GDPs rather than on the forecasting of the (observable) annual regional GDPs. To ensure the comparability and homogeneity of those estimates, our procedure hinges upon the temporal and cross-section consistency in the same way as implemented in the NA. The reliance on mimicking the NA limits the selection of indicators as well as the modeling approach. Lehmann and Wohlrabe (2012) present a detailed forecasting exercise at the regional level, using a variety of models and a large set of indicators with different spatial coverage.

3. Case Study: A System of Flash Regional Quarterly GDP Estimates for Spain

In this section we present the main results of a system of regional quarterly GDP flash estimates for the Spanish economy, following the modeling approach previously outlined.

Table 3. Comparison with other methodological approaches

	Di Fonzo (1990)	Di Fonzo & Marini (2003)	Proietti (2011)	Ours
High-frequency model	Static model + I(1) innovations	Unspecified	Static model + I(1) or I(2) innovations	Static or dynamic model + AR(1)/I(1) innovations
Temporal constraints	Yes	Yes	Yes	Yes
Transversal constraints	Yes	Yes	Yes	Yes
Chain- linking constraints	No	No	No	Yes
Mixing frequencies	Yes	Yes	Yes	Yes
Computational approach	Matrix oriented	Matrix oriented	State space	Matrix oriented

3.1. Selection of Monthly Regional Indicators

This subsection details the indicators that have been selected for model estimation. The selection process was carried out under the premise that indicators should be available in a timely fashion and should provide a synthetic measure of each of the regional economies.

The criterion for choosing these variables is the consideration of the regional counterpart of all the indicators used in the compilation of the QNA (see Álvarez 1989, Martínez and Melis 1989, INE 1993 and Álvarez 2005). To fulfil this goal, we have prepared a set of monthly regional indicators that provides a fairly comprehensive basis for analyzing and monitoring GDP at the regional level. This set offers a high-frequency approximation to the behavior of the main macroeconomic aggregates: gross added value (industry, construction, and services), consumption, external trade and employment. The selected indicators, with a brief description of them, are:

- IPI: Index of Industrial Production.
  - Units: Index number.
  - Source: National Statistical Institute (*Instituto Nacional de Estadística, INE*).
  - Starting date: 1995.01.
  - Back-calculation: combining data from 1990 base (1995.01–2002.01) and 2005 base (2002.01–2011.12), using the oldest period-on-period rates of growth to retropolate the newest base.
- LIC: Municipal construction licenses. Total area to build.
  - Units: square meters.
  - Source: Ministry of Public Works (*Ministerio de Fomento*).
  - Starting date: 1995.01.
  - Back-calculation: Data for Basque Country (País Vasco) during the period 1995.01–1997.12 have been back calculated using the average of the remaining regions as indicator. Some specific missing data (Basque Country -2008.08- and Navarra -2009.12-) have been interpolated using the program TRAMO.
- PER: Overnight stays in hotel establishments.
  - Units: Number of overnight stays.
  - Source: National Statistical Institute (*Instituto Nacional de Estadística, INE*).
  - Starting date: 1995.01.
  - Back-calculation: The series have been homogenized since 1998.12 by means of univariate intervention analysis in order to correct the methodological change introduced in 1999.01.
- IAS: Services sector activity indicator.
  - Units: Index number. Valuation at current prices.
  - Source: National Statistical Institute (*Instituto Nacional de Estadística, INE*).
  - Starting date: 2005.01.
  - Deflated using the Consumer Price Index (CPI) for services (house rentals excluded).
  - Missing data since 1995.01 have been estimated using the static factor derived from the indicators that start in 1995.01 as regressor.

- ICM: Retail sales index.
  - Units: Index number. Valuation at current prices, gas stations excluded.
  - Source: National Statistical Institute (*Instituto Nacional de Estadística, INE*).
  - Starting date: 2001.01.
  - Deflated using the CPI for services (house rentals excluded).
  - Missing data since 1995.01 have been estimated using the static factor derived from the indicators that start in 1995.01 as regressor.
- MAT: Car registrations.
  - Units: Registrations.
  - Source: Traffic department (*Dirección General de Tráfico, Ministerio del Interior*).
  - Starting date: 1995.01.
- EXP: Exports of goods.
  - Units: Euros, valuation at current prices.
  - Source: External trade statistics, Ministry of Economy and Competitiveness.
  - Starting date: 1995.01.
  - Deflated using the national exports unit value index.
- IMP: Imports of goods.
  - Units: Euros, valuation at current prices.
  - Source: External trade statistics, Ministry of Economy and Competitiveness.
  - Starting date: 1995.01.
  - Deflated using the national imports unit value index.
- AFI: Social security system: registered workers.
  - Units: persons.
  - Source: Labor department (*Ministerio de Empleo y Seguridad Social*).
  - Starting date: 1995.01.

The short-term indicators, in order to be consistent with the QNA data (as mentioned in Section 2), have been seasonally and calendar adjusted.

3.2. Regional Synthetic Indexes

To combine the information contained in the individual monthly indicators in an efficient and operative way, we have calculated a synthetic indicator for each region. In order to convey an idea of the correlation between the individual indicators and the estimated synthetic indicator (common factor), Table 4 shows the loading vectors, estimated by means of principal components factor analysis.

We have to note how loadings vary depending on the predominant activities in which each region specializes. Since two of the indicators (IAS and ICM) have been completed using the common factor estimated from the remaining indicators, their correlations with the common factor estimated with the balanced panel are overestimated to a certain extent. This fact complicates the exact quantification of their role. However, their economic relevance (IAS for the whole services sector and ICM for private consumption) recommends their inclusion in the estimation of the regional GDP trackers.

The corresponding monthly regional synthetic indicators are temporally aggregated to the quarterly frequency.

Table 4. Regional synthetic indexes: loading structure

	AFI	EXP	IMP	IPI	LIC	MAT	PER	ICM	IAS
Andalucía (AND)	0.54	0.28	0.05	0.45	0.01	0.77	0.21	0.73	0.90
Aragón (ARA)	0.31	0.63	0.29	0.79	0.04	0.63	0.01	0.51	0.65
Asturias (AST)	0.42	0.41	0.25	0.31	0.17	0.63	0.25	0.74	0.87
Baleares (BAL)	0.29	0.24	0.19	0.33	0.09	0.74	0.07	0.37	0.78
Canarias (CAN)	0.63	0.01	0.01	0.50	0.10	0.54	0.23	0.78	0.84
Cantabria (CANT)	0.35	0.56	0.36	0.57	0.07	0.56	0.06	0.06	0.74
Castilla La Mancha (CLM)	0.50	0.32	0.31	0.57	0.39	0.48	0.03	0.69	0.88
Castilla León (CYL)	0.31	0.55	0.50	0.61	0.01	0.68	0.01	0.08	0.82
Cataluña (CAT)	0.38	0.62	0.45	0.77	0.12	0.69	0.01	0.68	0.90
Extremadura (EXT)	0.41	0.30	0.14	0.14	0.30	0.75	0.01	0.42	0.76
Galicia (GAL)	0.32	0.62	0.24	0.45	0.01	0.70	0.23	0.66	0.89
Madrid (MAD)	0.41	0.43	0.32	0.62	0.01	0.48	0.28	0.75	0.69
Murcia (MUR)	0.45	0.24	0.01	0.37	0.04	0.76	0.19	0.79	0.87
Navarra (NAV)	0.35	0.61	0.54	0.72	0.01	0.32	0.21	0.09	0.64
País Vasco (PV)	0.14	0.58	0.49	0.76	0.08	0.57	0.01	0.62	0.86
La Rioja (RIO)	0.18	0.66	0.44	0.54	0.25	0.43	0.19	0.67	0.88
Valencia (VAL)	0.43	0.53	0.25	0.75	0.06	0.64	0.01	0.72	0.91

3.3. National Accounts Data: Regional Accounts and Quarterly National Accounts

Apart from the monthly regional indicators mentioned above, regional annual GDPs in chained-volume indices are provided by the RA according to ESA-95 conventions and they are available for the time span 1995–2011. The cross-section dimension includes 17 regions (*Comunidades Autónomas*) plus two autonomous cities that will be jointly considered, giving M = 18, a NUTS-2 regional breakdown according to Eurostat's classification.

Finally, the quarterly transversal constraint is the Spanish quarterly volume GDP provided by the QNA. This variable is compiled seasonally and calendar adjusted.

3.4. Empirical Results

Using the abovementioned data for the period 1995.01 – 2012.12 we can compare now the final results obtained using the different benchmarking techniques mentioned in section two (Fernandez, Chow-Lin, Santos Silva-Cardoso (SSC for brevity), Proportional Denton and Proietti) in order to select the most appropriate in terms of correlation and volatility.

Table 5 shows the summary results obtained with the different methods. Starting with the composite indicators derived by factor analysis for each region in the first stage, we apply different benchmarking methods and compare the different results obtained after final balancing. In order to summarize the results, we present the average correlation of the quarterly growth rate of GDP finally estimated by region with the initial composite indicator and the average standard deviation of the quarterly growth rate of GDP finally estimated by region.

This table shows that there seems to be a trade-off relationship between correlation and volatility (except in proportional Denton, which shows high volatility and low correlation). The Fernández and Chow-Lin methods are closest to the evolution of the indicator, without assuming a more complex structure in the errors, as is the case with SSC and Proietti.

Based on these results, we have decided to choose either the Fernández or the Chow-Lin method, because we think it is more important to be as faithful as possible to the information contained in the indicators, despite having higher volatility. Additionally, this is the method currently suggested for the compilation of the Spanish QNA (see Quilis 2005).

Regarding the distinction between the Fernández or Chow-Lin method, the results of the exercise show an innovational parameter with Chow-Lin close to 1 (approximately 0.98–0.99 in most cases), so under this situation both methods are practically equivalent.

With the aim of analyzing both the duration and the date of entry and exit of the recession in each region, Table 6 presents the evolution of the estimated year-on-year rates

Table 5. Comparison of methods (quarterly rates of growth)

	Fernandez	Chow-Lin	SSC	Denton Prop.	Proietti
Average Standard Deviation	0.821	0.858	0.731	0.843	0.744
Average Correlation	0.767	0.776	0.683	0.670	0.736

Table 6. *Dating recession in quarterly GDP (year-on-year rates of growth)*

	2008				2009				2010				2011				2012				
	T I	T II	T III	T IV	T I	T II	T III	T IV	T I	T II	T III	T IV	T I	T II	T III	T IV	T I	T II	T III	T IV	
Spain	2.7	1.9	0.3	–1.4	–3.4	–4.4	–4.0	–3.1	–1.5	–0.2	0.0	0.4	0.5	0.5	0.6	0.0	–0.7	–1.4	–1.6	–1.5	**Spain**
Andalucía	2.9	1.7	–0.4	–1.8	–2.8	–3.8	–3.8	–3.4	–2.3	–1.0	–0.5	–0.1	–0.2	–0.4	–0.1	0.2	–0.2	–0.7	–1.4	–1.7	**Andalucía**
Aragón	3.5	2.2	0.7	–2.9	–4.4	–4.9	–4.8	–1.8	–1.3	–1.5	–0.5	0.1	–0.1	1.0	0.8	–1.5	–0.7	–2.1	–2.0	–0.2	**Aragón**
Asturias	2.9	2.3	0.2	–1.0	–3.9	–5.6	–5.8	–4.6	–2.0	–0.7	0.0	0.3	0.1	0.4	0.3	–0.6	–1.3	–1.9	–2.0	–2.1	**Asturias**
Baleares	2.6	2.3	0.7	–0.5	–2.1	–5.1	–4.3	–3.8	–2.5	–0.9	–0.6	–0.8	–0.6	2.2	2.5	2.0	1.3	–1.1	–0.9	0.1	**Baleares**
Canarias	1.6	1.4	–0.3	–1.5	–3.0	–4.5	–4.7	–4.5	–3.1	–2.0	1.1	1.4	2.3	2.8	1.2	1.1	–0.3	–1.0	–2.0	–0.6	**Canarias**
Cantabria	2.1	1.9	0.7	–0.5	–2.1	–3.7	–4.5	–4.1	–2.4	–1.3	–1.2	–0.6	–0.1	0.2	1.2	0.7	0.0	–0.6	–0.8	–0.1	**Cantabria**
Castilla La Mancha	3.8	2.5	0.6	–0.8	–2.8	–3.5	–4.4	–4.1	–3.1	–2.4	–0.4	–0.2	–0.2	0.4	–0.7	–0.5	–1.3	–1.7	–1.3	–1.2	**Castilla La Mancha**
Castilla León	3.2	2.1	0.5	–2.3	–3.2	–3.4	–3.3	–1.4	0.2	1.6	0.2	0.5	0.9	0.3	2.0	0.8	–0.3	–1.6	–1.9	–2.1	**Castilla León**
Cataluña	1.9	0.9	–0.5	–1.5	–3.7	–4.2	–3.9	–3.1	–1.0	0.2	0.7	1.0	0.6	0.5	0.9	0.0	–0.1	–0.6	–0.9	–0.6	**Cataluña**
Extremadura	4.4	3.6	0.5	–1.1	–3.1	–3.4	–3.0	–2.0	–0.8	0.0	–2.3	–1.3	–1.5	–1.1	1.4	–0.7	–1.0	–2.7	–2.9	–2.7	**Extremadura**
Galicia	3.6	2.1	1.1	–0.1	–2.2	–3.8	–3.8	–3.9	–1.9	0.5	0.3	0.7	0.7	0.3	–0.3	–0.8	–0.9	–1.5	–1.5	–1.9	**Galicia**
Madrid	2.4	2.0	0.6	–1.1	–2.7	–3.8	–2.4	–1.8	–0.5	0.4	–0.1	0.0	0.8	0.6	0.8	0.0	–1.7	–2.5	–2.8	–2.9	**Madrid**
Murcia	3.7	2.8	1.1	–1.2	–3.6	–5.4	–4.6	–4.7	–2.7	–0.6	–0.3	0.0	–0.5	–0.3	–0.3	–0.1	–0.3	–0.7	–1.7	–2.4	**Murcia**
Navarra	2.8	3.7	1.0	0.0	–3.8	–4.9	–3.3	–2.4	0.1	0.3	0.4	0.8	1.5	2.0	1.0	0.5	–1.3	–2.4	–2.0	–1.5	**Navarra**
País Vasco	2.6	2.5	1.1	–0.8	–3.4	–5.1	–4.7	–3.2	–0.8	0.9	1.2	1.4	1.7	1.4	0.8	0.2	–1.0	–1.6	–1.3	–1.0	**País Vasco**
La Rioja	3.8	2.3	0.7	–1.0	–3.8	–4.5	–5.2	–5.2	–2.8	–2.7	–1.8	–0.5	–0.7	0.8	1.5	1.3	0.6	0.0	–0.4	–1.0	**La Rioja**
Valencia	3.4	1.9	0.5	–2.7	–6.1	–7.0	–6.0	–4.4	–2.1	–0.2	–0.6	0.1	0.7	0.2	0.2	–0.8	–1.1	–1.3	–1.0	–1.1	**Valencia**

▨	Negative rates
🟥	Minimum rate
🟩	Positive rates

of growth in the quarterly frequency; for the exercise performed with the Chow-Lin method, for example.

The table shows how the crisis has affected regions unevenly. For example, we can place the bulk of the recession between the fourth quarter of 2008 and the first quarter of 2010. Most of the regions fell into recession at the same time but not all of them left it simultaneously; this is the case of regions such as Andalucía, where the contractionary period is particularly long. We can see that many regions fall back into recession after the first quarter of 2012.

In relation to the variance of these results, Figure 2 shows the different box plots of the year-on-year rates of growth in the quarterly frequency for the different regions:

We observe a greater presence of outliers in periods of recession than in periods of expansion. This is partly due to the longer duration of the latter, rendering the median less representative for recessionary quarters. At the same time, the highest rate of variability is not linked to the larger size (GDP weight) of the region (see Appendix 1).

The temporal dimension of the data allows us to appreciate a reduction in volatility after 2003, although this is a property inherited from the annual data published by the RA (see Figure 3):

Finally, in order to clarify the importance of the balancing procedure on the final estimate, an exercise on two regions has been carried out: one with a large size (Cataluña) and other with a small size (La Rioja). This exercise is trying to reveal whether a small region can seriously change its initial estimate of quarterly GDP with the final balancing.

Initial or preliminary estimates do not take into account the information contained in the national quarterly GDP. Those initial estimates are modified to be consistent each quarter with the quarterly national GDP, reflecting the fact that the national data is the transversal aggregation of the regions.

The difference between the initial and the final estimates reflects the balancing procedure that ensures the transversal constraint and preserves, for each region, the temporal consistency with the Regional Accounts.

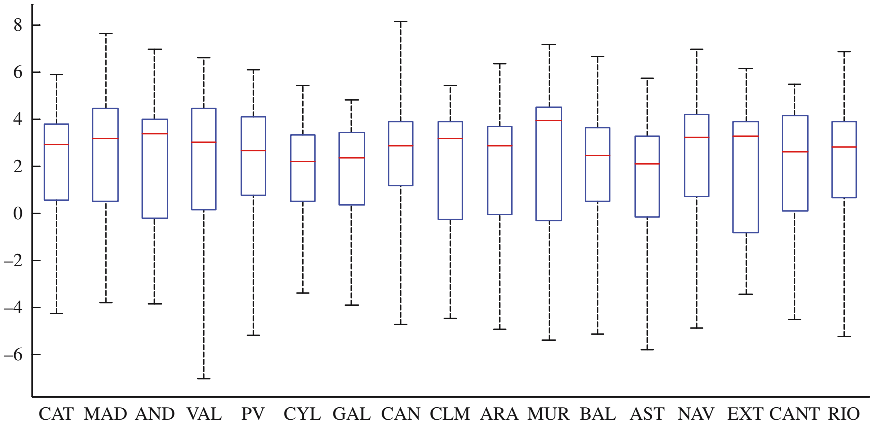


Fig. 2. Box plot: annual growth rates by region in quarterly frequency, sorted according to weight on Spanish GDP. Note: Central line stands for median values, the box represents 50% of the central part of the data and the whiskers are the minimum and maximum of the data

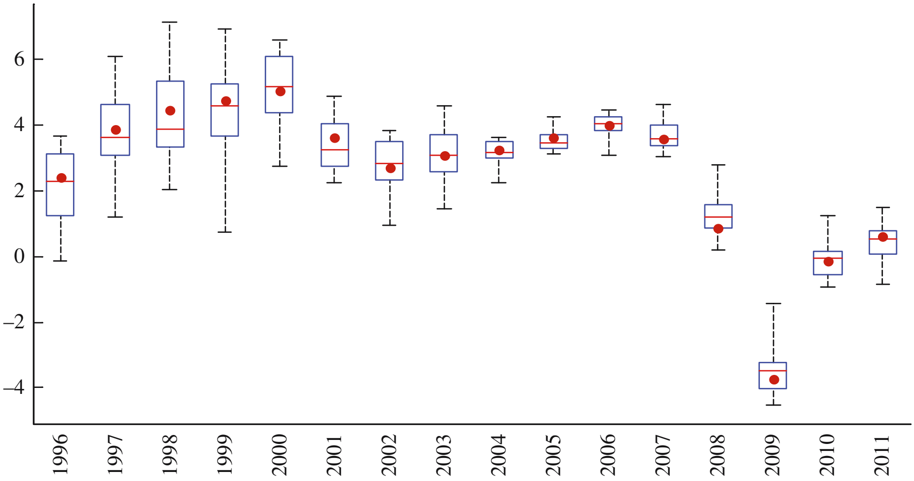


Fig. 3. Box plot: year-on-year rates of growth (annual data). Note: Dot is the aggregate data for Spain

Figure 4 shows, firstly, the initial quarterly regional GDP estimation (distribution of annual regional GDP according to the indicator) against the evolution of the indicator and, secondly, the initial quarterly estimation against the final quarterly GDP.

It is easy to see how the first step of estimating quarterly GDP depending on the evolution of the indicator is even more crucial to the subsequent balancing procedure. Furthermore, the small region does not have its initial estimate changed substantially compared with that of the large region. This fact shows the robustness of the balancing

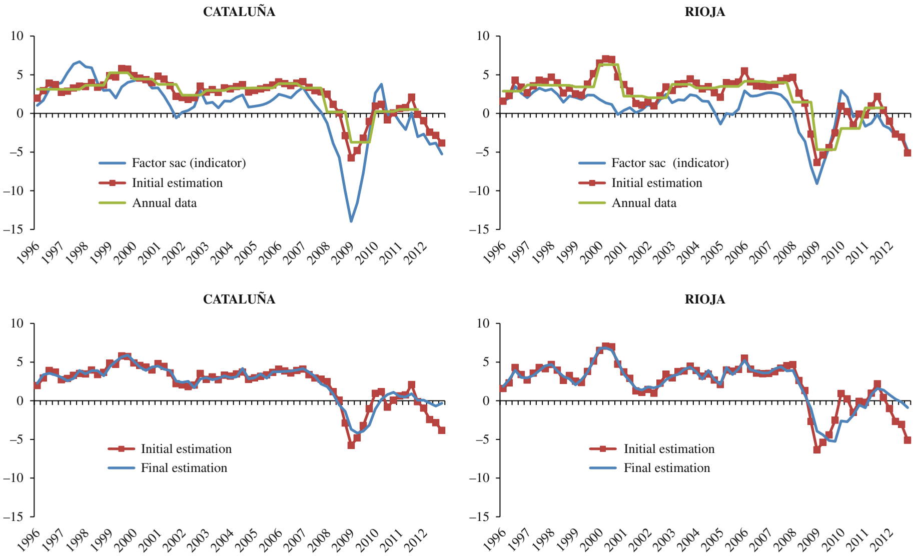


Fig. 4. Initial quarterly estimation vs. final balanced estimation. Small vs. large regions, year-on-year rates of growth

procedure, revealing that the variability in the final estimate is driven by the variability of the selected indicator.

4. Conclusions

In this article we have presented a feasible way to add a regional dimension to the short-term macroeconomic analysis, satisfying the temporal and cross-section constraints imposed by the NA. Our procedure generates results that are comparable across regions, are based on meaningful short-term information, and may be updated at the same time as the GDP flash national estimates, providing a solid basis for specific regional estimates.

In summary, the major outcomes of the model are:

- It solves the lack of quarterly GDP at the regional level, providing estimates consistent with the official available data published by the NA (RA and QNA). These estimates are a stand-alone product that may be used as input in regional econometric models.
- It provides a regional breakdown of the early estimates of the quarterly national volume GDP that may be released simultaneously, providing flash estimates at the regional level.

There are several promising lines of research that may broaden the scope of the article. The use of dynamic-factor models to estimate the regional high-frequency synthetic indexes may provide a more complete description of the economic conditions at the regional level.

The modeling approach can be extended easily to accommodate several types of extrapolations. For example, the transversal benchmark of the model (the national quarterly GDP) may be an official release made by the NSI or a forecast made by an analyst (e.g., the research department of an investment bank). In the latter case, we can combine these forecasts with the projected path for the underlying short-term quarterly regional indicators to generate the corresponding regional quarterly GDPs. The resulting conditional extrapolations can be used to assess the expected cyclical position of each region with respect to the nation.

Finally, the estimated regional quarterly GDPs can be used to analyze issues related to the synchronicity of the regional business cycles as well as their pattern of co-movements.

Appendix 1: Main Features of the Spanish Regions (2011)

	Population (thousand)	Population weight	GDP weight	Employment weight
Andalucía	8,270.5	17.9%	13.5%	14.7%
Aragón	1,315.5	2.9%	3.2%	3.1%
Asturias	1,054.5	2.3%	2.1%	2.1%
Baleares	1,092.5	2.4%	2.5%	2.6%
Canarias	2,107.0	4.6%	3.9%	4.1%
Cantabria	578.3	1.3%	1.2%	1.2%
Castilla La Mancha	2,045.4	4.4%	3.5%	3.9%
Castilla León	2,483.8	5.4%	5.3%	5.3%
Cataluña	7,303.1	15.8%	18.6%	17.8%
Extremadura	1,083.1	2.3%	1.6%	1.9%
Galicia	2,732.0	5.9%	5.3%	5.7%
Madrid	6,371.6	13.8%	18.0%	16.8%
Murcia	1,471.4	3.2%	2.6%	3.0%
Navarra	622.8	1.4%	1.7%	1.6%
País Vasco	2,127.9	4.6%	6.2%	5.3%
La Rioja	312.7	0.7%	0.8%	0.7%
Valencia	5,001.2	10.8%	9.5%	9.8%
Ceuta y Melilla	151.7	0.3%	0.3%	0.3%
Spain	46,125.0	100.0%	100.0%	100.0%

5. References

Abad, A. and E.M. Quilis. 2005. "Software to Perform Temporal Disaggregation of Economic Time Series." Eurostat, Working Papers and Series. Available at: http://ec.europa.eu/eurostat/documents/4187653/5774917/LN-SR012007-EN.PDF/c83eb69e-e3a9-4fdd-923d-76c09fea6f7b (accessed October 2015).

Abad, A., A. Cuevas, and E.M. Quilis. 2007. "Chain-Linked Volume Indexes: a Practical Guide." Universidad Carlos III de Madrid, Instituto Flores de Lemus, *Boletín de Inflación y Análisis Macroeconómico* 157: 72–85. Available at http://e-archivo.uc3m.es/handle/10016/20332#preview (accessed October 2015).

Álvarez, F. 1989. "Base Estadística en España de la Contabilidad Nacional Trimestral." *Revista Española de Economía* 6: 59–84.

Álvarez, R. 2005. "Notas Sobre Fuentes Estadísticas." In Servicio de Estudios del Banco de España, *El análisis de la economía española*, Alianza Editorial, Madrid, Spain.

Bloem, A.M., R.J. Dippelsman, and N.O. Mæhle. 2001. *Quarterly National Accounts Manual. Concepts, Data Sources, and Compilation*. International Monetary Fund. Available at: https://www.imf.org/external/pubs/ft/qna/2000/Textbook/ch1.pdf (accessed October 2015).

Bournay, J. and G. Laroque. 1979. "Réflexions sur la Méthode D'elaboration des Comptes Trimestriels." *Annales de l'INSEE* 36: 3–30. Available at: http://www.jstor.org/stable/20075332.

Caporello, G. and A. Maravall. 2004. “Program TSW. Revised Manual.” Bank of Spain, Occasional Paper no. 0408. http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/04/Fic/do0408e.pdf (accessed October 2015).

Chow, G. and A.L. Lin. 1971. “Best Linear Unbiased Distribution and Extrapolation of Economic Time Series by Related Series.” *Review of Economic and Statistics* 53: 372–375. Available at: http://www.jstor.org/stable/1928739.

Denton, F.T. 1971. “Adjustment of Monthly or Quarterly Series to Annual Totals: an Approach Based on Quadratic Minimization.” *Journal of the American Statistical Society* 66: 99–102. Doi: http://dx.doi.org/10.1080/01621459.1971.10482227.

Di Fonzo, T. 1987. *La Stima Indiretta di Serie Economiche Trimestrali*. Cleup Editore, Padova, Italy.

Di Fonzo, T. 1990. “The Estimation of *M* Disaggregate Time Series when Contemporaneous and Temporal Aggregates are Known.” *Review of Economics and Statistics* 72: 178–182. Doi: http://dx.doi.org/10.2307/2109758.

Di Fonzo, T. 2002. “Temporal Disaggregation of Economic Time Series: Towards a Dynamic Extension.” European Commission (Eurostat) Working Papers and Studies, Theme 1, General Statistics (pp. 41). Available at: http://ec.europa.eu/eurostat/documents/3888793/5816173/KS_AN-03-035-EN.PDF/21c4417c-dbec-45ec-b440-fe8bf95661b7?version=1.0 (accessed October 2015).

Di Fonzo, T. and M. Marini. 2003. “Benchmarking Systems of Seasonally Adjusted Time Series According to Denton’s Movement Preservation Principle.” Dipartimento di Scienze Statistiche, Università di Padova, Working Paper no. 2003–09. Available at: http://www.oecd.org/std/21778574.pdf, (accessed October 2015).

Eurostat. 1998. *Handbook of Quarterly National Accounts*. Luxembourg: Statistical Office of the EC.

Fernández, R.B. 1981. “Methodological Note on the Estimation of Time Series.” *Review of Economic and Statistics* 63: 471–478. Doi: http://dx.doi.org/10.2307/1924371.

Gómez, V. and A. Maravall. 1996. “Programs TRAMO and SEATS.” Bank of Spain, Working Paper no. 9628. Available at: http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/96/Fich/dt9628e.pdf (accessed October 2015).

Gómez, V. and A. Maravall (1998a) “Guide for using the programs TRAMO and SEATS”, Bank of Spain, Working Paper no. 9805. Available at: http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/98/Fic/dt9805e.pdf (accessed October 2015).

Gómez, V. and A. Maravall (1998b) “Automatic modeling methods for univariate series”, Bank of Spain, Working Paper no. 9808. Available at: http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/98/Fic/dt9808e.pdf (accessed October 2015).

Gregoir, S. 1994. “Propositions Pour une Désagrégation Temporelle Basée sur des Modèles Dynamiques Simples.” In *Workshop on Quarterly National Accounts*, ed. Eurostat. Luxembourg: Statistical Office of the EC. Available at: http://ec.europa.eu/eurostat/documents/3888793/5815741/KS-AN-03-014-EN.PDF/284f1001-fd36-4999-b007-a22033e8aaf9 (accessed October 2015).

INE. 1993. *Contabilidad Nacional Trimestral de España (CNTR). Metodología y serie trimestral 1970–1992*. Instituto Nacional de Estadística.

Litterman, R.B. 1983. "A random walk, Markov model for the distribution of time series." *Journal of Business and Economic Statistics* 1: 169–173. Available at: http://www.jstor.org/stable/1391858.

Lehmann, R. and K. Wohlrabe. 2012. "Forecasting GDP at the Regional Level with Many Predictors." CESIFO, Working Paper no. 3956. Available at: http://www-sre.wu.ac.at/ersa/ersaconfs/ersa13/ERSA2013_paper_00015.pdf (accessed October 2015).

Martínez, A. and F. Melis. 1989. "La Demanda y la Oferta de Estadísticas Coyunturales." *Revista Española de Economía* 6: 7–58.

Polasek, W. and R. Séllner. 2010. "Spatial Chow-Lin Methods for Data Completion in Econometric Flow Models." Institut für Höhere Studien (HIS), Economic Series no. 255. Available at: https://www.ihs.ac.at/publications/eco/es-255.pdf (accessed October 2015).

Proietti, T. 2006. "Temporal Disaggregation by State Space Methods: Dynamic Regression Methods Revisited." *Econometrics Journal* 9: 357–372. Doi: http://dx.doi.org/10.1111/j.1368-423X.2006.00189.

Proietti, T. 2011. "Multivariate Temporal Disaggregation with Cross-Sectional Constraints." *Journal of Applied Statistics* 38: 1455–1466. Doi: http://dx.doi.org/10.1080/02664763.2010.505952.

Quilis, E.M. 2005. "Benchmarking Techniques in the Spanish Quarterly National Accounts." European Commission, Working papers and studies (Eurostat-OECD Workshop on Frontiers in Benchmarking Techniques and Their Application to Official Statistics, Luxembourg, April 7–8, 2005). Available at: http://ec.europa.eu/eurostat/documents/4187653/5774917/LN-SR012007-EN.PDF/c83eb69e-e3a9-4fdd-923d-76c09fea6f7b (accessed October 2015).

Salazar, E., R. Smith, S. Wright, and M. Weale. 1994. "Indicators of Monthly National Accounts." In *Workshop on Quarterly National Accounts*, ed. Eurostat. Luxembourg: Statistical Office of the EC. Available at: http://ec.europa.eu/eurostat/documents/3888793/5815741/KS-AN-03-014-EN.PDF/284f1001-fd36-4999-b007-a22033e8aaf9 (accessed October 2015).

Santos-Silva, J.M.C. and F. Cardoso. 2001. "The Chow-Lin Method Using Dynamic Models." *Economic Modelling* 18: 269–280. Doi: http://dx.doi.org/10.1016/S0264-9993(00)00039-0.

Vidoli, F. and C. Mazziotta. 2012. "Spatial Composite and Disaggregate Indicators: Chow-Lin Methods and Applications." *Real Estate* 2: 9–19. Available at: http://fvidoli.weebly.com/uploads/2/3/0/8/23088460/eng_spatialcompositeanddisaggregate.pdf (accessed October 2015).

Received February 2013
Revised January 2015
Accepted January 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 673–697, http://dx.doi.org/10.1515/JOS-2015-0040

Cultural Variations in the Effect of Interview Privacy and the Need for Social Conformity on Reporting Sensitive Information

Zeina M. Mneimneh[1], *Roger Tourangeau*[2], *Beth-Ellen Pennell*[1], *Steven G. Heeringa*[1], *and Michael R. Elliott*[3]

Privacy is an important feature of the interview interaction mainly due to its potential effect on reporting information, especially sensitive information. Here we examine the effect of third-party presence on reporting both sensitive and relatively neutral outcomes. We investigate whether the effect of third-party presence on reporting sensitive information is moderated by the respondent's need for social conformity and the respondent's country of residence. Three types of outcomes are investigated: behavioral, attitudinal, and relatively neutral health events. Using data from 22,070 interviews and nine countries in the cross-national World Mental Health Survey Initiative, we fit multilevel logistic regression to study reporting effects on questions about suicidal behavior and marital ratings, and contrast these with questions about having high blood pressure, asthma, or arthritis. We find that there is an effect of third-party presence on reporting sensitive information and no effect on reporting of neutral information. Further, the effect of the interview privacy setting on reporting sensitive information is moderated by the need for social conformity and the cultural setting.

Key words: Privacy; cultural variability; interview variability.

1. Introduction

Many studies instruct interviewers to conduct their interviews in a private setting (with no third party present). The rationale is that the presence of a third party during the interview might interfere with the response process, possibly causing respondents to misreport information (especially that of a sensitive nature) or to rely on others present during the interview for answers to knowledge questions.

However, establishing interview privacy might not always be feasible, even when the study protocol calls for it. Most studies in different countries that report whether interviews were carried out in private reported rates of third-party presence higher than 35 percent (Mneimneh 2012).

The relatively common presence of a third party during the interview has led researchers to examine the effect of third-person presence on reporting answers to survey questions. In 1997, Aquilino proposed a framework that describes three factors affecting the size and

[1] University of Michigan - Survey Research Center, Ann Arbor, MI 48106-1248, U.S.A. Emails: zeinam@umich.edu, bpennell@umich.edu and sheering@umich.edu
[2] Westat, 1600 Research Boulevard, Rockville, MD 20850, U.S.A. Email: RogerTourangeau@westat.com
[3] University of Michigan - School of Public Health, Dept. of Biost., M4041 SPH II, 1420 Washington Heights, Ann Arbor, MI 48105, U.S.A. Email: mrelliot@umich.edu

direction of bystanders' effect on reporting sensitive information ("bystander" and "third person" are used interchangeably in this article). The first is whether the survey question asks about factual or subjective information. The second is whether the bystander knows the factual information requested. The third is the perceived likelihood that the respondent might experience negative consequences by revealing new and unwelcome information to the bystander. If the third party already knows the information to be reported, then their presence might not have an effect, or it might even lead to more truthful reporting. If the third party does not know the information requested, and the respondent perceives a high likelihood of negative consequences by revealing this information, then the presence of a third party might lead to misreporting. The last two factors are related to the type of relationship that exists between the third party and the respondent. The relationship must be significant to the respondent, and the respondent's answer must bear directly on the relationship for the respondents to change their answers in order to convey a more desirable image in the presence of the third person (Aquilino 1997; Pollner and Adams 1997). This has led a number of researchers to investigate the effect of specific types of relationships on reporting answers to sensitive factual and attitude questions.

The most commonly studied types of relationship between the third party and the respondent are parental and spousal relationships. Several studies have investigated the effect of parent or spouse presence during the interview on reporting sensitive *factual* information. The effect of parent presence on reporting substance use among youth and young adults is consistent. Youth and young adults interviewed in the presence of their parents were less likely to report substance use (Aquilino 1997; Aquilino et al. 2000; Gfroerer 1985; Hoyt and Chaloupka 1994; Moskowitz 2004). In a meta-analysis conducted by Tourangeau and Yan (2007), the authors concluded that parental presence significantly reduced reporting of socially undesirable information.

The effect of spouse presence has been less consistent. In one of the studies conducted by Aquilino (1997) among married couples less than 34 years old, spouse presence had no effect on reporting substance use. However, in another sample, Aquilino et al. (2000) found higher rates of reported substance use among respondents (less than 45 years old) who were interviewed in the presence of their spouse. Casterline and Chidambaram (1984) studied the effect of third-party presence on reporting contraceptive use in several countries in Latin America, the Caribbean, Asia, and Africa. The authors found that husband presence during the interview reduced the odds of reporting contraceptive use. Pollner and Adams (1994) found that spouse presence reduced the reporting of depression symptoms among adult respondents residing in Los Angeles, but it did not have an effect on reporting other mental-health symptoms.

The effect of third-party presence on reporting *subjective* information has also been mixed. In India, youth and young adults (15–29 years old) interviewed in the presence of their parents reported more positive attitudes toward family (Podmore et al. 1975). In the United States, respondents interviewed in the presence of their spouse reported a better quality of marital life (Aquilino 1993). On the other hand, Anderson and Silver (1987) found that partner presence had no effect on agreement between emigrant Soviet couples when asked about their satisfaction with housing and standard of living. Pollner and Adams (1997) and Smith (1997) also reported that spouse presence had no effect on respondents' attitudes toward spouse support and satisfaction with household arrangements and subjective questions on marriage and gender differences, respectively. Pollner and Adams (1997) concluded that the inconsistency of findings and the ambiguity surrounding the interview conditions of some studies indicated that a conclusive judgment about third-party effects on reporting was not possible and needed further investigation.

Several factors might have contributed to the inconsistency of these findings, including interviewer differences in reporting privacy measures and a failure to take into account respondent characteristics that are associated with both reporting of sensitive information and establishing a private interview. These respondent characteristics could moderate the effect of third-party presence on reporting.

The first factor, interviewer differences, is important since interviewers are relied on to request, achieve, and report on interview privacy. We have shown that between-interviewer variation in reporting privacy measures is large (even larger than between-country differences) and could possibly vary from one study to another depending on interviewer training protocols and the population studied (Mneimneh 2012). To date, none of the studies investigating the effect of third-party presence on reporting controlled for interviewer variation in interview privacy.

The second factor involves respondent characteristics that could moderate the effect of third-party presence on reporting sensitive information. These include respondents' need for social conformity and their cultural background. Respondent's need for social conformity is driven by the respondent's motivations and desire to obtain social approval from others (Cialdini and Goldstein 2004) and minimize possible negative evaluation by others (Johnson and van de Vijver 2003). Such conformity motivations could be activated and strengthened depending on contextual stimuli, such as the perception of threats to fitting in socially and the lack of anonymity of the interaction (Cialdini and Goldstein 2004). Therefore, the presence of others during the interview may intensify such motivations already held by certain respondents. Thus respondents with activated conformity motivations might be more likely to misreport sensitive information than those with low conformity motivations. To our knowledge, only one study controlled for a similar conformity construct when investigating the effect of third-party presence on reporting (Moskowitz 2004). The author did not investigate the possible moderating effect of the respondent's need for social conformity on interview privacy and reporting, however.

Respondents' cultural background is another characteristic associated with both interview privacy and social-desirability motivation. The respondent's cultural background could be defined in several ways. Throughout this article the term cultural background refers to the country where the respondent resides or originates from. Respondents who reside in collectivist and lower-income countries were more likely to be interviewed in the presence of a third party than those residing in individualistic and high-income countries, respectively (Mneimneh 2012). This is consistent with how collectivist cultures are structured. In collectivist cultures, the self is defined in terms of relationships with others. To maintain harmony and interdependence in such cultures, close attention is given to others in the social context, especially if they belong to the in-group circle–the individual's family and friends and others concerned with his or her welfare (Smith et al. 2006; Triandis 1995). Hofstede et al. (2010) discuss how a collectivist culture considers it normal for a member of one's in-group to invade one's privacy at any time. This stands in contrast to individualistic cultures, where ties between individuals are loose and the primary concern is independence (Hofstede et al. 2010; Triandis 1989). Cultural differences due to masculinity also seem to affect the presence of partners during the interview. Countries high in masculinity have distinct gender roles and are more focused on achievement and material success. The higher the masculinity of the country, the less likely a partner will be around during the interview (Mneimneh 2012).

The relationship between the interview privacy setting and wealth seems to be driven by the country's level of individualism and masculinity (Mneimneh 2012). As a country's wealth increases, its citizens give more attention to self-expression and personal choice. Moreover, as wealth increases, resources and commodities become more available, allowing citizens to become more independent rather than interdependent (Hofstede et al. 2010; Smith et al. 2006).

Respondents' cultural background is also associated with socially desirable reporting behavior. Most published literature has focused on the collectivism dimension and found it to be positively associated with general measures of social desirability (Bernardi 2006; Bond and Smith 1996; Triandis 1995) or specific components of social desirability, namely impression management (Lalwani et al. 2006). Triandis (1995, cited in Johnson and van de Vijver 2003), discussed how honesty in interactions with strangers is valued more in individualist societies, while saving face is more salient in collectivist societies. In collectivist societies, an individual's loss of face can also cause a loss of face for the group they belong to. Thus members that belong to the same in-group have a shared interest in avoiding any loss of face to maintain in-group harmony (Ting-Toomey 1999).

Other cultural dimensions found to be associated with social desirability are a country's level of uncertainty avoidance (Bernardi 2006) and wealth (van Hemert et al. 2002). The association between social desirability and these two cultural dimensions, however, work in opposite directions: whereas a country's level of uncertainty avoidance shows a positive association with social-desirability motivation, a country's level of wealth exhibits a negative association. Neither masculinity nor power-distance dimensions have been found to be associated with social desirability (Bernardi 2006).

In summary, the perception of one's role and status vis-à-vis the roles and statuses of others in any social interaction is unconsciously guided by one's cultural background (Hofstede et al. 2010); as a result, respondents in certain cultural settings – for example, those living in collectivist (vs. individualistic) and lower-income (vs. high-income) societies – may be more concerned about how they appear to others present during an interview, leading them to misreport information. These hypothesized associations are guided by the previously demonstrated relationship between a country's level of collectivism and wealth and social-desirability bias. In this article, we specifically focus on the cultural factors that have been shown to be associated with both interview privacy and social desirability (a country's wealth and its level of individualism).

To understand the moderating effect of respondents' need for social conformity and the cultural setting on the third-party presence and reporting of sensitive behaviors (suicidal behavior) and attitudes (marital rating), we tested the following hypotheses:

Suicidal behavior:

Hypothesis 1: There is an interaction between the respondent's need for social conformity and third-party presence. Reporting differences due to third-party presence are larger among respondents with a high need for social conformity.

Hypothesis 2: There is an interaction between the country's level of individualism and third-party presence. The effect of third-party presence on reporting suicidal behavior is reduced as the country's level of individualism increases.

Hypothesis 3: There is an interaction between the country's level of wealth and third-party presence. The effect of third-party presence on reporting suicidal behavior is accentuated in countries with a middle and low GNI per capita (compared to countries with a high GNI per capita).

Marital rating:

Hypothesis 4: There is an interaction between the respondent's need for social conformity and partner presence. Reporting differences due to partner presence are larger among respondents with a high need for social conformity.

Hypothesis 5: There is an interaction between the country's level of wealth and partner presence. The effect of partner presence on marital ratings is accentuated in countries with a middle and low GNI per capita (as compared to countries with a high GNI per capita).

For comparative purposes, we also tested the effect of third-party presence on relatively neutral health measures (having high blood pressure, arthritis, or asthma) using the following hypotheses:

Physical chronic conditions outcome.

Hypothesis 6: Third-party presence is not significantly related to the likelihood of reporting any of the physical chronic conditions.

Hypothesis 7: There is no significant interaction between the respondent's need for social conformity and third-party presence on reporting any of the physical chronic conditions.

Hypothesis 8: Neither the country's level of individualism nor the country's level of wealth significantly moderates the effect of third-party presence on reporting any of the physical chronic conditions.

From a theoretical point of view, testing for the above hypotheses will help survey researchers to better understand the effects of interview privacy on reporting sensitive information and shed the light on some of the inconsistencies in the literature. It emphasizes the importance of respondents' cultural and individual-level characteristics in moderating such effects and the need to account for them in future investigations. From a practical point of view, this research highlights the need to design well-defined interviewer privacy observations to capture the dynamics of the interview interaction and further investigate its effects on different types of questions.

2. Methods

Data from the World Mental Health (WMH) Survey Initiative were used to address these research questions.

The survey design, implementation, and data processing across all participating WMH countries were coordinated by two central organizations. All participating countries were required to follow a standard survey protocol that includes a probability-sample design, a fixed minimum sample size, a shared core instrument, a specific translation protocol, a set of quality-control measures, and specific interviewer-training protocols. However, countries were allowed to adapt certain features, including computerization of the interview, the contact protocol, respondent incentives, and the field-team structure. This mix of centralization and local control allowed establishing a survey protocol adapted to local conditions while maintaining comparability of the data collected (Pennell et al. 2010).

2.1. Sample Designs for the WMH Surveys

In the first twenty-four countries that completed the WMH surveys, nine – Brazil, Bulgaria, Japan, Lebanon, Mexico, Nigeria, the People's Republic of China, Romania, and the United States – collected data on the respondent's social-conformity motivations and interview privacy. The analyses focused on these nine countries.

All WMH surveys targeted the adult population and most of them featured nationally representative probability samples of individuals in households. One (Mexico) was representative of urban areas, one (Nigeria) of selected states, and four (Brazil, India, Japan, and the People's Republic of China) of selected metropolitan areas. Detailed information on the survey sample design is published elsewhere (Heeringa et al. 2008).

To reduce interview length, the WMH interviews were designed to be administered in two parts: Part 1 included core questionnaire sections and Part 2 included noncore sections. All respondents completed Part 1; Part 2 was administered to a subsample of Part 1 respondents. The current analyses focused on Part 2 respondents as the scale measuring the respondent's need for social conformity and the majority of key outcomes were collected only in Part 2. Table 1 presents the number of Part 2 interviews completed in each country and the number of field interviewers.

2.2. Questionnaire

The WHO Composite International Diagnostic Interview (CIDI) Version 3.0 was used in all WMH surveys. The CIDI 3.0 is a fully structured interview that generates diagnoses for a wide range of mental-health disorders. It also collects information on treatment, disability, and physical chronic conditions. Detailed questions on social and family life, employment history, finances, and childhood experiences are also included. The questionnaire was translated into each country's local language following the WHO translation guidelines (Harkness et al. 2008; Kessler et al. 2004; Kessler and Üstün 2004).

2.3. Questionnaire Administration

In all countries, trained interviewers conducted face-to-face interviews using either paper-and-pencil interviewing (PAPI) or computer-assisted personal interviewing (CAPI)

Table 1. Country-specific measures

Country	Number of Part 2 completed interviews	Number of interviewers	CIDI social-conformity scale Cronbach's alpha	Individualism score	Masculinity score	GNI per capita in nominal dollar values[a]
Brazil	2,942	129	0.7	−0.005	−0.520	9390 (M)
Bulgaria	2,233	97	0.7	−0.377	−1.039	6270 (M)
China	1,628	161	0.6	−0.841	+0.462	4270 (L)
Japan	1,682	191	0.7	+0.366	+2.136	41850 (H)
Lebanon	1,031	119	0.6	−0.005	−0.289	8880 (M)
Mexico	2,362	32	0.6	−0.377	+0.635	8890 (M)
Nigeria	2,143	63	0.8	−0.841	−0.693	1180 (L)
Romania	2,357	66	0.7	−0.377	−0.924	7840 (M)
United States of America	5,692	189	0.6	+2.455	+0.231	47390 (H)

[a] Calculated according to the Atlas Method for the year the data was collected; (L) = Low GNI per capita; (M) = Middle GNI per capita; (H) = High GNI per capita

methods. Interviewer training in each country was modeled on a five-day training that project staff from each country had attended. All data were collected before 2008, with the majority of the fieldwork taking place between 2001 and 2003. Detailed information on the specific years of data collection, mode used, response rates, and supervisor-to-interviewer ratio are published elsewhere (Pennell et al. 2008).

2.4. *Measures Studied*

As detailed below in the analysis section, a multilevel model was used to explore our research questions. Below we list the key binary outcomes, followed by predictors.

2.4.1. Key Outcomes

The key outcome measures fell into three categories, representing a variety of sensitive behaviors, attitudes, and relatively neutral measures.

Behavioral outcome: CIDI 3.0 included a section on suicidality. In this section, respondents were asked to report whether they had ever made a suicide plan or attempted suicide. This outcome was chosen based on the authors' perceived judgment that reporting on suicidal behavior is relatively undesirable across cultures.

Attitudinal outcome: Married respondents were asked to rate their relationship with their current partner by answering the following question: "Using a scale from 0 to 10 where 0 means the worst possible marriage, and 10 means the best, how would you rate your marriage?" Since we were interested in investigating the effect of partner presence on reporting high marital ratings (rather than an average increase of one point on the scale), the score was categorized into two groups: a high rating defined as a score above the midpoint of the scale (which is five) and low rating (five or below). Reporting high marital rating is judged to be desirable if the partner is present.

Possible cultural variations in the sensitivity of reporting suicidal behavior and high marital rating were investigated through testing for a moderation effect of cultural background on third-party/partner presence.

Physical Chronic Condition Outcomes: Respondents were asked to report whether they had ever been told by any health professional that they had the following conditions: high blood pressure, asthma, or arthritis. These outcomes were chosen based on our judgment that such chronic conditions are less sensitive than suicidal behavior and marital rating across cultures.

A number of substantive and sociodemographic predictors were investigated. These predictors are described below.

2.4.2. Respondent-Level Predictors

Interview privacy: Interviewer observations about the privacy of the interview setting were collected at the end of the interview. Interviewers were instructed to specify: 1) whether a third person was present at any time during the interview; 2) the relationship of the third party to the respondent (parent, partner, child, youth, or other adults); and 3) the duration of the third party's presence during the interview (all the time, most of the time, about half of the time, about one quarter of the time, or less than one quarter of the time). The current analyses focus on any third-party presence, excluding children under the age of six (for the suicidal behavior outcome) and partner presence (for the marital rating outcome). The duration of the third-party stay was divided into two categories: 1) all of the interview time and 2) some of the interview time.

Respondent's need for social conformity: CIDI 3.0 included an adapted version of the Marlowe-Crowne Social Desirability Scale (Crowne and Marlowe 1960). Crowne and Marlowe designed the scale to measure the construct they labeled as "the need for social approval." The CIDI adapted scale consisted of 10 true or false statements such as "I never met a person that I didn't like" and "I have always told the truth." Respondents were instructed to choose the answer that first came to their mind and not take too much time thinking before they answered. A value of one was assigned to each item the respondent endorsed. The sum of these values formed a score ranging from zero to ten. Within each country, the respondent's total score was standardized by the country's average and standard deviation. A score was considered high if it was at least one standard deviation above the national-level mean. This measure will be referred to as the CIDI social-conformity scale, and was used to investigate whether an interaction between third-party presence and a high score on this scale had an impact on reporting sensitive information. Country-level Cronbach alpha estimates are in the acceptable to good range (0.6–0.8) (Table 1).

Respondent sociodemographic predictors*:* These variables were treated as control variables and included the respondent's gender, age (18–34, 35–49, 50–64, older than 64 years), marital status (never married, currently married or cohabiting, previously married), education level (high, middle, low, very low relative to the rest of the country), income (high, middle, low, very low relative to the rest of the country), current employment status (employed, studying, taking care of home, other), and household size (fewer than two, two, three, more than three). Among those currently married or cohabiting, the partner's level of education (high, middle, low, very low) and type of occupation (not employed, have a low-skill job, low-to-average skill job, average-to-high skill job, and high-skill job) were also taken into account.

2.4.3. Interviewer-Level Predictors

Interviewer identification numbers were available for all nine countries. This information enabled the modeling of random effects (intercepts) for the individual members of the interviewing force in each country. No other interviewer-level covariates were available.

2.4.4. Country-Level Predictors

Two country-level cultural dimensions were included in the current analyses: individualism and masculinity. Country-specific scores for each of the two dimensions and more details on their assessment are available in Hofstede et al. (2010). For each dimension and for each country, a standardized score was calculated based on the average score and the standard deviation across all the countries included in the analyses. Higher scores indicated higher levels of the underlying dimension. The country-specific scores are presented in Table 1.

Finally, the countries' economic strength and standard of living was measured by their Gross National Income (GNI) per capita. GNI measures in nominal dollar values that were calculated according to the Atlas Method for the year the data was collected were used. According to the World Bank, the Atlas method "applies a conversion factor that averages the exchange rate for a given year and the two preceding years, adjusted for differences in rates of inflation" ("GNI per capita," Atlas method (current US$), accessed July 26, 2012, http://data.wordbank.org/indicator/NY.GNP.PCAP.CD). The limited number of countries and the vast differences in wealth leave big gaps in the GNI measure in this group of countries. The countries that are in this sample thus lend themselves to categories of wealth rather than a continuum of wealth (as evident in Table 1). Given these big gaps, the countries were categorized into three groups according to their GNI per capita: high, middle, and low GNI per capita. This categorization matches the World Bank classification of countries by income.

3. Analysis

Because of the limited number of countries, a two-level logistic regression with respondents nested within interviewers was used for all outcomes. Interviewers were treated as random effects. All predictors were treated as fixed effects, including the cultural factors. Indicator variables for individual countries were not included as fixed effects due to a resulting low-rank design matrix. Country effects were adjusted for through the three country-level variables – individualism, masculinity, and GNI per capita. The analyses were repeated using a three-level model with respondents at level 1, interviewers at level 2, and countries at level 3; the results with respect to our main hypotheses were *consistent* but less stable because of the limited number of countries in the analyses. Thus the analyses and the findings reported below are based on the two-level model.

For suicidal behaviors and physical chronic condition outcomes (high blood pressure, asthma, or arthritis), the main predictors included whether any third party was present during all of the interview time (vs. none of the time), any third party was present during some of the interview time (vs. none of the time), the respondent's level on the CIDI social-conformity scale (high vs. low), the country's level of individualism, and the country's level of GNI per capita (middle vs. high and low vs. high). Only cultural factors that have been demonstrated in the literature to be related to *both* interview privacy and social desirability were included as main predictors. The country's standardized masculinity score as well as respondent demographics and socioeconomic characteristics were also included in all the models as control variables. Interactions between third-party presence measures and a high score on the CIDI social-conformity scale, and between third-party presence measures and the country's level of individualism and wealth were tested. To maintain a parsimonious model, only significant interactions (with a *p*-value less than 0.05) were kept in the final model; nonsignificant interactions were removed.

The marital rating outcome was collected in only five out of the nine countries. Substantive predictors for the marital outcome rating included partner presence, the respondent's level on the CIDI social-conformity scale (high vs. low), and the country's level of GNI per capita (middle vs. high and low vs. high). The country's level of individualism was not included in the model as two of the five countries had the same level of individualism and there was not enough variation to explore. Respondent sociodemographic characteristics as well as the presence of any other third party (other than a partner) and country's level of masculinity were also included in all the models. Interactions between partner-presence measures and a high score on the CIDI social-conformity scale, and between partner-presence measures and the country's level of wealth were tested separately. To maintain a parsimonious model, only significant interactions (with a *p*-value less than 0.05) were kept in the final model; nonsignificant interactions were removed.

All multilevel models are unweighted. To explore the effect of weights on the findings, the analyses were replicated using weighted and unweighted single-level logistic regression models. The weights accounted for within-country differential probability of selection, post-adjustment to the country's sociodemographic distributions, and subsampling to specific questionnaire sections. Adjusted and unadjusted point estimates were consistent (with single-level models being more stable), suggesting that the use of weights to adjust for informative sample design was not necessary for the relationships investigated. All analyses were conducted using the PROC GLIMMIX procedure in SAS version 9.2 (SAS institute, NC).

4. Results

4.1. Outcome Rates and Interview Privacy Rates Across Countries

Table 2 presents the different outcome rates in each of the countries included in the analyses. Weighted rates for combined suicide plan or suicide attempt reports differed greatly across countries, with the lowest rates (1.0%) reported in Nigeria and Romania, and the highest rate (7.5%) reported in the United States. High marital ratings were reported by the majority of respondents in all five countries where such ratings were collected. All rates were higher than 90%.

All countries included questions on high blood pressure, asthma, and arthritis. Reported rates of asthma were generally low (mainly less than 5.5%), except in the United States where 11.6% of the respondents reported having asthma. Arthritis and high blood pressure were more common and rates were more variable across countries. Arthritis rates range from 7.0% (Lebanon) to 33.3% (Romania) and high blood pressure rates range from 4.3% (Lebanon) to 24.1% (United States).

On average, 37% of the interviews were conducted in the presence of a third person. Table 3 presents the rates of third-party presence and duration of stay in each of the nine countries. In most countries, a third party was present for part of the interview rather than the entire interview (the average rate across countries was 26%). On average, only 11% of the interviews had a third party present during *all* of the interview time. Partners were present in 19% of the interviews across all the countries. Again, partners were mostly present for some parts of the interview rather than the whole interview. Country-specific rates are presented in Table 4.

4.2. Effect of Third-Party Presence on Reporting: Results from Multilevel Logistic Regression Model

4.2.1. Suicidal Behavior

As hypothesized and shown in Table 5 (interaction model column), the respondent's need for social conformity moderated the relationship between third-party presence and

Table 2. Weighted outcome percent (s.e.)

	Suicide plan or attempt		High marital rating[a]		Physical chronic conditions[b]			
Country	*N*	Percent (s.e.)	*N*	Percent (s.e.)	*N*	High blood pressure Percent (s.e.)	Asthma percent (s.e.)	Arthritis percent (s.e.)
Bulgaria	2,233	1.3 (0.3)	—	—	2,206	20.6 (1.1)	1.9 (0.3)	9.3 (0.7)
Brazil	2,942	6.5 (0.5)	1,836	90.1 (1.0)	2,929	19.9 (1.1)	2.2 (0.3)	7.6 (0.7)
China	1,628	1.5 (0.2)	1,228	95.4 (0.9)	1,625	14.0 (1.1)	3.5 (0.7)	11.6 (1.1)
Japan	1,682	2.8 (0.4)	—	—	1,275	16.1 (1.3)	5.3 (0.9)	9.2 (0.9)
Lebanon	1,031	2.8 (0.4)	696	92.8 (1.1)	601	4.3 (0.9)	1.2 (0.5)	7.0 (1.2)
Mexico	2,362	4.1 (0.3)	—	—	2,362	9.7 (0.8)	2.2 (0.4)	7.5 (0.7)
Nigeria	2,143	1.0 (0.1)	1,411	95.1 (0.8)	2,138	3.0 (0.5)	0.6 (0.2)	16.9 (0.9)
Romania	2,356	1.0 (0.3)	—	—	2,350	17.2 (0.9)	3.0 (0.4)	33.3 (1.2)
USA	5,692	7.5 (0.3)	1,601	93.6 (0.8)	5,689	24.1 (0.8)	11.6 (0.5)	27.3 (0.8)

[a] Among married respondents. Dashes (—) refer to unavailable data

[b] Differences in sample size between Suicide Plan or Attempt column and this column are either due to subsampling into the chronic condition section or missing data

s.e. = standard error

Table 3. Percentage of any third-party presence (s.e.)

Country	*N*	No third party present during interview	Any third party present all interview time	Any third party present some of the interview time
Bulgaria	2,232	60.7 (1.0)	13.2 (0.7)	26.1 (0.9)
Brazil	2,942	41.0 (0.9)	19.7 (0.7)	39.3 (0.9)
China	1,628	63.9 (1.2)	12.8 (0.8)	23.3 (1.0)
Japan	1,346	87.5 (0.9)	2.7 (0.4)	9.8 (0.8)
Lebanon	1,031	33.6 (1.5)	21.5 (1.3)	44.9 (1.5)
Mexico	2,350	65.4 (1.0)	10.7 (0.6)	23.9 (0.9)
Nigeria	2,141	68.0 (1.0)	7.1 (0.6)	24.9 (0.9)
Romania	2,356	64.6 (1.0)	15.2 (0.7)	20.2 (0.8)
USA	5,304	70.0 (0.6)	5.0 (0.3)	25.0 (0.6)

Values are unweighted estimates of sample % (standard error)

reporting suicidal behavior. Ignoring this interaction would give a misleading picture of the direction of third-party presence effects.

Respondents who scored *high* on the social-conformity scale and who had a third person present during the interview were less likely to report suicidal behavior compared to those who had no one present. Among respondents who scored high on social conformity, the odds of reporting suicidal behavior were lower by a factor of .92 when a third party was present during the entire interview and by a factor of .56 when a third party was present part of the time relative to not having a third party present during the interview (see Table 5). The results were quite different among respondents who scored *low* on the CIDI social-conformity scale. Among those respondents, having a third person present all or part of the time increased reporting of suicidal behavior compared to being interviewed alone (OR = 1.20 and 1.40, respectively; see Table 5).

Contrary to what was hypothesized, however, neither the country's level of individualism nor its wealth significantly moderated the effect of third-party presence on reporting suicidal behavior.

Table 4. Percentage of partner presence (s.e.)

Country	*N*	No partner present during the interview	Partner present all interview time	Partner present some of the interview time
Bulgaria	2,232	73.9 (0.9)	8.5 (0.6)	17.6 (0.8)
Brazil	2,942	73.3 (0.8)	9.6 (0.5)	17.1 (0.7)
China	1,628	80.1 (1.0)	7.1 (0.6)	12.8 (0.8)
Japan	1,346	93.8 (0.7)	1.5 (0.3)	4.7 (0.6)
Lebanon	1,031	68.3 (1.5)	11.1 (1.0)	20.6 (1.3)
Mexico	2,350	88.6 (0.7)	4.0 (0.4)	7.4 (0.5)
Nigeria	2,141	91.6 (0.6)	1.7 (0.3)	6.6 (0.5)
Romania	2,356	79.3 (0.8)	8.3 (0.6)	12.4 (0.7)
USA	5,304	82.9 (0.5)	2.9 (0.2)	14.2 (0.5)

Values are unweighted estimates of sample % (standard error)

Table 5. Odds ratio and 95% confidence interval from multilevel logistic model predicting suicide attempt or plan (N = 21,329)[a]

	Main model	Interaction model
Presence of third party		
Third party present all of the time	1.20 (0.99–1.43)	1.20 (0.98–1.46)
Third party present some of the time	**1.31 (1.16–1.45)**	**1.40 (1.23–1.60)**
No third party present	1.00	1.00
Social-conformity score		
High score[b]	**0.69 (0.59–0.82)**	0.82 (0.67–1.01)
Low score	1.00	1.00
Individualism standardized score (IND)	**2.72 (2.26–3.28)**	**2.72 (2.26–3.28)**
Country's GNI per capita		
Low	**12.78 (6.54–24.97)**	**12.74 (6.52–24.88)**
Middle	**12.49 (6.96–22.41)**	**12.45 (6.94–22.34)**
High	1.00	1.00
Present all of the time × high social conformity	—	0.92 (0.55–1.54)
Present some of the time × high social conformity	—	**0.56 (0.38–0.82)**

[a] Significant odds ratios with $p < 0.05$ are presented in bold. Dashes (—) indicate variables not entered in the model. All models control for sex, age, marital status, education level, income level, employment status, household size, and the country's score on masculinity

[b] High score is greater or equal to one standard deviation above the mean

4.2.2. Marital Ratings

High ratings of marital relationships were positively associated with partner presence during the interview, controlling for other respondent-level characteristics, the country's level of wealth and masculinity (see Main Model, Table 6). Unlike suicidal behavior (and contrary to what was hypothesized), the association between partner presence and reporting a high marital rating was not significantly moderated by the respondent's need for social conformity.

The effect of partner presence on reporting marital rating was, however, significant among respondents interviewed in a country with middle GNI per capita (compared to high GNI per capita, specifically the United States), as hypothesized. Respondents who were interviewed in a middle-income country and in the presence of their partner during the whole interview had more than 1.5 times the odds of reporting a high marital rating score compared to those who had no partner present and were in a country with high GNI per capita (see Interaction Model, Table 6). Though the interaction effect among respondents interviewed in low-income countries was in the hypothesized direction (and similar to middle-income countries), it was not statistically significant.

4.2.3. Physical Chronic Conditions

Unlike reports of suicidal behavior and marital relationship rating, reporting high blood pressure, asthma, or arthritis was not significantly associated with third-party presence (see Table 7). Moreover, as hypothesized, no significant interaction effects were found between third-party presence and the respondent's need for social conformity or with either of the country-level factors for any of the three outcomes.

Table 6. Odds ratio and 95% confidence interval from multilevel logistic model predicting high marital rating score (N = 6,595)[a]

	Main model	Interaction model
Presence of partner		
Partner present all of the time	**1.59 (1.08–2.35)**	0.51 (0.24–1.10)
Partner present some of the time	**1.36 (1.07–1.73)**	1.51 (0.92–2.50)
No partner present	1.00	1.00
Social-conformity score		
High score[b]	**1.55 (1.17–2.05)**	**1.53 (1.16–2.02)**
Low score	1.00	1.00
GNI per capita		
Low	**1.63 (1.15–2.30)**	**1.56 (1.07–2.28)**
Middle	0.70 (0.47–1.03)	0.66 (0.43–1.01)
High	1.00	1.00
Partner present all of the time × low GNI per capita	—	1.35 (0.92–1.98)
Partner present some of the time × low GNI per capita	—	1.01 (0.77–1.31)
Partner present all of the time × middle GNI per capita	—	**1.64 (1.21–2.22)**
Partner present some of the time × middle GNI per capita	—	0.94 (0.78–1.14)

[a] Significant odds ratios with $p < 0.05$ are presented in bold. Dashes (—) indicate variables not entered in the model. All models control for sex, age, marital status, education level, income level, employment status, household size, and country's level of masculinity

[b] High score is greater or equal to one standard deviation above the mean

5. Discussion

This article is the first to investigate whether the effect of third-party presence on reporting sensitive information is moderated by the respondent's need for social conformity and the respondent's country of residence. Three types of outcomes were investigated: a sensitive behavioral outcome (suicidal behavior), a sensitive attitudinal measure (marital rating), and three relatively neutral health measures (having high blood pressure, arthritis, or asthma). Third-party presence effects are moderated by the respondent's need for social conformity and the respondent's country of residence. Though such moderating effects were hypothesized for both outcomes, they differed depending on whether the outcome is behavioral or attitudinal.

For sensitive behaviors, specifically suicidal behavior, having a third party present during the interview was associated with lower odds of reporting such behavior *only* among respondents who had *high* scores on the social-conformity scale. Having someone present during the interview might create a contextual stimulus that strengthens the respondent's already existing need for social conformity and increases their perceived likelihood that revealing such behavior in the presence of a third party will trigger negative consequences. To avoid such consequences, respondents might prefer not to disclose such information in the presence of the third party, a phenomenon sometimes referred to as

Table 7. Odds ratio and 95% confidence interval from multilevel logistic model predicting chronic conditions[a]

	High blood pressure $N = 20{,}482$	Asthma $N = 20{,}516$	Arthritis $N = 20{,}446$
Presence of third party			
Third party present all of the time	1.10 (0.97–1.25)	0.78 (0.61–1.01)	0.95 (0.82–1.09)
Third party present some of the time	1.07 (0.99–1.20)	1.02 (0.88–1.19)	1.07 (0.97–1.18)
No third party present	1.00	1.00	1.00
Social-conformity score			
High score[b]	0.93 (0.84–1.03)	1.03 (0.87–1.22)	**0.90 (0.81–1.00)**
Low score	1.00	1.00	1.00
Individualism standardized score	**2.08 (1.78–2.43)**	**1.74 (1.40–2.17)**	**1.28 (1.05–1.55)**
Country's GNI per capita			
Low	**4.06 (2.41–6.84)**	1.07 (0.49–2.33)	1.08 (0.56-2.05)
Middle	**6.35 (3.88–10.39)**	1.03 (0.51–2.10)	**0.47 (0.25–0.86)**
High	1.00	1.00	1.00

[a] Significant odds ratios with $p < 0.05$ are presented in bold. All models control for sex, age, marital status, education level, income level, employment status, household size, and the country's score on masculinity

[b] High score is greater or equal to one standard deviation above the mean

impression management. In fact, Paulhus (1984) shows that the Marlow Crowne Scale that is used in this paper as a measure of social conformity loads on impression management scales. The picture is quite different among respondents who do not have such needs and who scored low on social conformity. Among those respondents, the likelihood of reporting suicidal behavior was higher in the presence of a third party than in private. Respondents who scored low on social conformity might not have concerns about possible negative consequences from divulging such information, or they might have already confided in the third person present during the interview. Among such respondents, a third party who is present during the interview might act as a truth control increasing the reporting of such behavior. This has been also reported by two studies investigating sensitive undesirable outcomes in the United States (Aquilino 1997; Hoyt and Chaloupka 1994), where respondents who had a third party present were more likely to report illicit substance use than those interviewed in private. Future research on the effect of information already held by the third party on reporting sensitive information and how it interacts with respondent's need for social conformity is needed.

The interaction effect between the respondent's need for social conformity and third-person presence on reporting suicidal behavior was only statistically significant among interviews where a third person was present during "some" of the interview time. Though the direction of the interaction effect was the same when a third person was present "all" the interview time, it was not statistically significant. This difference between the two measures of third-party presence may reflect any of several factors: small cell size, psychological presence of third person and question location, and possible misclassification in the duration measure. First, the absence of a significant interaction effect between third person presence during "all" of the interview time and social conformity could be attributed to the small sample sizes, given that suicide behavior is rare (four percent in the overall sample) and that most of the nonprivate interviews had someone present "some" of the time rather than "all" of the time (70.4% vs. 29.6%). Second, though it is not possible to ascertain that the third party was physically present during the specific administration of the suicide questions in the interviews where the third person was present during "some" of the interview time, we suspect that people are typically present during the beginning of the interview and they might leave or come and go as the interview progresses. This could have primed the respondent to perceive the setting as nonprivate, even after the third person had physically left the interview setting. In fact, there is strong evidence in the literature that social influence is not only produced by the *actual* physical presence of family and friends but also by their *psychological* presence, through priming respondents with their name, words or questions about them, or just thinking about or imagining them (Shah 2003; Moretti and Higgins 1999; Berscheid and Reis 1998; Fitzsimons and Bargh 2003). It seems that the activation of the representation of others occurs outside of people's awareness and has a powerful automatic effect on people's perceptions and behaviors (Berscheid and Reis 1998). Thus, given that the suicide behavior questions are towards the first third to the middle of the interview (depending on the skip patterns), even if the third person was not physically present during the actual administration of the suicide questions, his or her physical presence during earlier questions might have activated a mental representation of the third party's presence during the administration of the suicide questions. However, how long the psychological presence could persist during the interview is a question for future investigation. It is possible that the psychological presence would only last briefly after the third person left and thus would manifest only in nearby questions but not those administered towards the end of a long interview (such as the marital rating question). Third, the difference in the statistical significance between the two measures of third person presence could be also attributed to misclassification of the duration of presence in either directions. This could happen especially in situations where multiple people are present and come and go at different points during the course of the interview. In fact, 27% of interviews conducted in the presence of third members are reported to have more than one third person present.

Besides the respondent's need for social conformity, we hypothesized that the respondent's country of residence would moderate the effect of third-party presence on reporting suicidal behavior. However, this interaction effect was not significant. Although third-party presence is more common in collectivist and lower-income countries (Mneimneh 2012), the effect of third-party presence on reporting suicidal behavior does not change across countries that differ on those characteristics. It seems likely that reporting suicidal behavior is highly sensitive across all societies, irrespective of their level of collectivism/individualism or wealth. Even if the general level of social desirability is higher among collectivist and lower-income societies, having a third person present during the interview does not differentially heighten the sensitivity of suicidal information across countries.

The second sensitive outcome investigated is respondents' attitudes toward their current marriages. Respondents interviewed in the presence of their partner might be more motivated to provide a positive characterization of their relationship than those interviewed in private. This finding was first documented by Aquilino (1997). Unlike suicide behavior, the effect of partner presence on reporting marital attitudes was not moderated by the respondents' need for social conformity. The absence of the hypothesized moderating effect of the respondent's high need for social conformity on partner presence and reporting a high marital rating could be explained by the nature of the marital rating measure. Unlike suicidal behavior, which is a factual measure, marital rating is subjective. Respondents who scored low on social conformity might have already confided in other household members about their suicide experiences. Thus low conformity respondents might be more inclined to report their "true" suicidal behavior when a third party was present. This might not be the case for marital ratings, because marital happiness is not factual and does not have a "true" value.

When investigating whether the effect of partner presence on reporting high marital rating varies by the country of residence, we found it was significant in countries with middle GNI per capita (compared to high GNI per capita). Respondents interviewed in middle-income countries might practice impression management in the presence of their partner and deliberately report higher ratings of their marital relationship compared to their counterparts in high-income countries so as to maintain a favorable image in front of their spouses. Such reporting behavior (impression management) has been found to be more present among collectivist cultures (compared to individualistic cultures), which are typically less wealthy (Lalwani et al. 2006). Though such an effect did not reach statistical significance in low-income countries, the same trend is observed in low-income countries, and upon further investigation it was not statistically different from the effect in middle-income countries. It is also important to note that the magnitude of the interaction effect was higher before controlling for country's level of masculinity (OR = 4.4 vs. OR = 1.6 after controlling for masculinity level). This could be attributed to the fact that the GNI-per-capita interaction effect is partially driven by the country's level of masculinity.

As discussed earlier, differences in the observed interaction effects between country-level variables and interview privacy in reporting suicidal behavior versus marital ratings could be attributed to the difference between the two in levels of sensitivity. It is possible that suicidal behavior is sensitive across country boundaries. Thus, controlling for the respondent's level of social conformity and irrespective where he or she lives, the third-party presence decreases reporting suicidal behavior similarly since the topic is considered very sensitive across countries. This mechanism could be different for outcomes such as marital rating, however, where a differential level of sensitivity might be observed across countries, reflected through interactions between contextual interview settings such as partner presence and country-level variables such as wealth. The third set of outcomes investigated is chronic physical conditions: high blood pressure, arthritis, and asthma. These outcomes are normally perceived as more neutral than suicidal behavior or marital happiness. They were chosen to compare the effect of third-party presence on reporting clearly sensitive information with its effect on reporting more neutral outcomes. As hypothesized, the presence of a bystander during the interview does not significantly affect reporting of such neutral outcomes across the different cultures studied.

In summary, in our analysis of the World Mental Health data collected in a sample of countries, the presence of a third party during the interview affects reporting of sensitive outcomes. The effect can go in either direction, depending on the respondent's need for social conformity, the respondent's cultural setting, and the type of question. Whether it increases or reduces reporting of potentially sensitive information, the presence of a third party during the interview adds some measurement variation among respondents in a sample. Such variation is another layer of error that needs to be minimized. This is even more important in cross-cultural research, where both the rates of third-person presence (Mneimneh 2012) and their effects vary by culture, jeopardizing comparability.

Researchers sometimes try to counter the effect of third-party presence on reporting by using self-administered modes. Though the fact that using self-administered modes reduces interviewer effects has been established in the literature (Tourangeau and Yan 2007), whether such modes reduce the reporting effects of third-party presence is still under debate. Only five studies used randomized mode experiments and investigated the effect of third-party presence among each randomized group (Aquilino 1997; Aquilino et al. 2000; Cahucahrd 2013; Couper et al. 2003; Moskowitz 2004). The results of these studies are mixed. The presence of a third party during the interview, especially if the third party is a household member, might prime the respondent and alter his or her frame of mind when answering sensitive questions. Even when using a self-administered mode, the interview might not feel as private when a third party is present compared to when the respondent is interviewed alone. In fact, Aquilino et al. (2000) reported that third-party effects are found even when the bystander did not interfere with the interview or communicate with the respondent. Still, due to the limited empirical evidence, further research is needed on the moderating effect of the interview mode on third-party presence and reporting sensitive information before any conclusions are made.

Given these findings, first, survey practitioners need to train interviewers better on how to achieve and maintain interview privacy, and the effect of such training interventions on interview privacy needs to be measured. Even when the mode of data collection uses self-administration, third-party effects are found (Aquilino et al. 2000). Second, researchers are encouraged to measure and train interviewers on collecting more specific data on the interview setting, such as the timing of the presence through section-specific measures, its dynamics, and information already known by the third person to better understand the effect of third-party presence. Third, if such effects are replicated, researchers need to take these measures into consideration when analyzing their data and control for them in their substantive investigations.

The findings presented here need to be interpreted with the following considerations in mind. First, the presence of a third party at respondents' homes during the interview is difficult to control by the researcher and was not randomized. Though we controlled for many factors known to be associated with third-party presence, it is possible that there are other unmeasured factors that we did not control for and which could have affected the relationship observed between third-party presence and the reporting of sensitive outcomes. Thus no causal interpretations of any of the findings can be made and only associations are reported.

Second, interview privacy measures are based on interviewer observations reported at the end of the interview. The duration of third-party presence during the interview is an overall measure for the whole interview and is not section specific. Thus there could be some misclassification in the duration of the stay, especially in situations where multiple people might have been present at different points during the interview. Moreover, in instances where the interviewer reported that someone was present during "some" of the interview time, it was not possible to determine whether the bystander was actually present when the target question was asked. This could explain some of the differences observed between the effect of the third-party presence on reporting the different sensitive outcomes. Information on suicidal behavior was collected toward the first half of the interview and might have been affected by the "psychological" presence of a third person, whereas marital ratings were collected toward the end of the second half and might have been less prone to such a psychological presence in a long interview like the CIDI 3.0. The duration-of-stay measure also reflects all of the bystanders that might have been present during the interview. Thus, when investigating partner presence, if the interviewer indicated that another person was present in addition to the partner, the duration of stay was assigned to all of the different bystanders present. However, this should not significantly affect the findings as in the large majority of interviews (83%), a partner was present but no other bystander was also present.

Third, the respondent's need for social conformity was measured using an adapted version of the Marlowe-Crowne Scale. Though this scale has been extensively used in the literature, and though the direction of the association between the respondent's need for social conformity and reporting outcomes is in the expected direction, these scales have their measurement problems. Self-reported scales that measure the need for social conformity are prone to misreporting, and their accuracy might vary depending on individual as well as cultural factors. Moreover, such scales assume that social conformity is stable from the time of its measurement; however, social conformity is contextual (Cialdini and Goldstein 2004) and could vary during the course of the interview.

Fourth, suicidal behavior and the rating of marital-relationship outcomes were chosen based on the authors' judgment of their undesirability/desirability across cultures (compared to chronic conditions). Unfortunately, there is no empirical evidence of the level of sensitivity of these outcomes (in general or across cultures). We believe that suicidal behaviors are generally seen as undesirable in all cultures; similarly, rating one's current marital relationship highly (especially in the presence of one's partner) is generally desirable across all cultures. Cultural differences in the level of sensitivity of these outcomes were investigated by testing for the interaction effects, where larger third-party presence reporting effects were hypothesized for the cultures with higher levels of social desirability according to the literature.

Fifth, the cultural-dimension indices used in these analyses come from data published in Hofstede et al. (2010), some of which were collected several years ago. One concern is the applicability of those indices to the WMH data. Nevertheless, a number of researchers have shown that while the values of many nations have been changing, the relative positioning of those nations has been maintained (Hofstede et al. 2010; Ingelhart and Baker 2000; Schwartz et al. 2000).

Finally, the lack of statistical significance of some hypothesized interactions could be attributed to the small-size interaction classes resulting from the low prevalence of both sensitive outcomes and the specific privacy interview setting, namely the presence of a third party during "all" the interview time.

6. Conclusion

Reporting sensitive information is affected by the respondent's personal characteristics and cultural values, the social context in which the topic is broached, and the players involved during an interview. For a given topic, such factors affect whether the respondents interpret the content as socially desirable or undesirable and whether they edit the information or not. This article demonstrates that the effect of the privacy of the interview setting on reporting is moderated by the need for social conformity and respondent's country of residence.

It is important for us to develop a better understanding of the dynamics surrounding interview privacy, how it is affected by respondents' and third parties' personal characteristics and cultural background, and how privacy affects different types of survey questions. To achieve that, future work that captures what information is already held by the third party, as well as more specific interviewer privacy observations are needed. Such improved privacy measures need to then guide both practical interventions on training interviewers to better achieve, maintain, and observe interview privacy, and empirical work on the possible moderating effects of personal, cultural, and other interview-setting factors on the response process for sensitive questions.

7. References

Anderson, B.A. and B.D. Silver. 1987. "The Validity of Survey Response: Insights from Interviews of Married Couples in a Survey of Soviet Emigrants." *Social Forces* 66: 537–554. Doi: http://dx.doi.org/10.1093/sf/66.2.537.

Aquilino, W. 1993. "Effects of Spouse Presence During the Interview on Survey Responses Concerning Marriage." *Public Opinion Quarterly* 57: 358–376.

Aquilino, W. 1997. "Privacy on Self-Reported Drug Use: Interactions with Survey Mode and Respondent Characteristics." In *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates (National Institute on Drug Abuse, Monograph 167)*, edited by L. Harrison and A. Hughes, pages 383–415. Rockville, MD: U.S. Department of Health and Human Services, National Institutes of Health, and National Institute on Drug Abuse, Division of Epidemiology and Prevention Research.

Aquilino, W.S., D.L. Wright, and A.J. Supple. 2000. "Response Effects Due to Bystander Presence in CASI and Paper-and-Pencil Surveys of Drug Use and Alcohol Use." *Substance Use and Misuse* 35: 845–867. Doi: http://dx.doi.org/10.3109/10826080009148424.

Bernardi, R.A. 2006. "Association Between Hofstede's Cultural Constructs and Social Desirability Response Bias." *Journal of Business Ethics* 65: 43–53. Doi: http://dx.doi.org/10.1007/s10551-005-5353-0.

Berscheid, E. and H.T. Reis. 1998. "Attractions and Close Relationships." In *The Handbook of Social Psychology*, edited by D.T. Gilbert, S.T. Fiske, and L. Gardner, pages 193–281. New York: McGraw-Hill.

Bond, R. and P.B. Smith. 1996. "Culture and Conformity: A Meta-Analysis of Studies Using Asch's (1952b, 1956) Line Judgment Task." *Psychological Bulletin* 119: 111–137. Doi: http://dx.doi.org/10.1037/0033-2909.119.1.111.

Cialdini, R.B. and N.J. Goldstein. 2004. "Social Influence: Compliance and Conformity." *Annual Review of Psychology* 55: 591–621.

Casterline, J. and V.C. Chidambaram. 1984. "The Presence of Others During the Interview and the Reporting of Contraceptive Knowledge and Use." In *Survey Analysis for the Guidance of Family Planning Programs*, edited by J. A. Ross and R. McNamara, 267–298. Liege: Ordina Editions.

Cahucahrd, S. 2013. "Using MP3 Players in Surveys: The Impact of a Low-tech Self-Administration Mode on Misreporting and Bystanders' Influence." *Public Opinion Quarterly* 77: 220–231. Doi: http://dx.doi.org/10.1093/poq/nfs060.

Couper, M.P., E. Singer, and R. Tourangeau. 2003. "Understanding the Effects of Audio-CASI on Self-Reports of Sensitive Behavior." *Public Opinion Quarterly* 67: 385–395.

Crowne, D.P. and D. Marlowe. 1960. "A New Scale of Social Desirability Independent of Psychopathology." *Journal of Consulting Psychology* 24: 349–354. Doi: http://dx.doi.org/10.1037/h0047358.

Fitzsimons, G.M. and J.A. Bargh. 2003. "Thinking of You: Nonconscious Pursuit of Interpersonal Goals Associated with Relationship Partners." *Journal of Personality and Social Psychology* 84: 148–164. Doi: http://dx.doi.org/10.1037/0022-3514.84.1.148.

Gfroerer, J. 1985. "Influence of Privacy on Self-Reported Drug Use by Youths." In *Self-Report Methods of Estimating Drug Use: Meeting Current Challenges to Validity*, edited by B.A. Rouse, N.J. Kozel, and L.G. Richards, pages 22–30. Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, and National Institute on Drug Abuse.

Harkness, J., B. Pennell, A. Villar, N. Gebler, S. Auilar-Gaxiola, and I. Bilgen. 2008. "Translation Procedures and Translation Assessment in the World Mental Health

Survey Initiative." In *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*, edited by R.C. Kessler and T.B. Üstün, pages 91–113. New York: Cambridge University Press.

Heeringa, S.G., J.E. Wells, F. Hubbard, Z.N. Mneimneh, W. Chiu, N.A. Sampson, and P.A. Berglund. 2008. "Sample Designs and Sampling Procedures." In *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*, edited by R.C. Kessler and T.B. Üstün, pages 14–32. New York, NY: Cambridge University Press.

Hofstede, G., G.J. Hofstede, and M. Minkov. 2010. *Culture and Organizations: Software of the Mind*, 3rd ed. New York, NY: McGraw-Hill.

Hoyt, G.M. and F.J. Chaloupka. 1994. "Effect of Survey Conditions on Self-Reported Substance Use." *Contemporary Economic Policy* 7: 109–121. Doi: http://dx.doi.org/10.1111/j.1465-7287.1994.tb00439.x.

Ingelhart, R. and W.E. Baker. 2000. "Modernization, Cultural Change, and the Persistence of Traditional Values." *American Sociological Review* 65: 19–51.

Johnson, T.P. and F.J.R. van de Vijver. 2003. "Social Desirability in Cross-Cultural Research." In *Cross-Cultural Survey Methods*, edited by J.A. Harkness, F.J.R. van de Vijer, and P.P. Mohler, pages 195–206. Hoboken, NJ: Wiley.

Kessler, R.C., J. Abelson, O. Demler, J.I. Escobar, M. Gibbon, M.E. Guyer, M.J. Howes, R. Jin, W.A., Vega, E.E. Walters, P. Wang, A. Zaslavsky, and H. Zheng. 2004. "Clinical Calibration of DSM-IV Diagnoses in the World Mental Health (WMH) Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI)." *International Journal of Methods in Psychiatric Research* 13: 122–139. Doi: http://dx.doi.org/10.1002/mpr.169.

Kessler, R.C. and T.B. Üstün. 2004. "The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI)." *International Journal of Methods in Psychiatric Research* 13: 93–121.

Lalwani, A.K., S. Shavitt, and T. Johnson. 2006. "What is the Relation Between Cultural Orientation and Socially Desirable Responding?" *Journal of Personality and Social Psychology* 90: 165–178. Doi: http://dx.doi.org/10.1037/0022-3514.90.1.165.

Moretti, M.M. and E.T. Higgins. 1999. "Internal Representations of Others in Self-Regulations: A New Look at a Classic Issue." *Social Cognition* 17: 186–208. Doi: http://dx.doi.org/10.1521/soco.1999.17.2.186.

Mneimneh, Z.N. 2012. "Interview Privacy and Social Conformity Effects on Socially Desirable Reporting Behavior: Importance of Cultural, Individual, Question Design and Implementation Factors." Available at: http://deepblue.lib.umich.edu/handle/2027.42/96051

Moskowitz, J.M. 2004. "Assessment of Cigarette Smoking and Smoking Susceptibility Among Youth: Telephone Computer-Assisted Self-Interviews Versus Computer-Assisted Telephone Interviews." *Public Opinion Quarterly* 68: 565–587. Doi: http://dx.doi.org/10.1093/poq/nfh040.

Paulhus, D.L. 1984. "Two-Component Models for Socially Desirable Responding." *Journal of Personality and Social Psychology* 46: 598–609. Doi: http://dx.doi.org/10.1037/0022-3514.46.3.598.

Pennell, B., Z.N. Mneimneh, A. Bowers, S. Chardoul, J.E. Wells, M.C. Viana, K. Dinkelmann, N. Gebler, S. Florescu, Y. He, Y. Huang, T. Toma, and G.V. Saiz. 2008. "Implementation of the World Mental Health Surveys." In *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*, edited by R.C. Kessler and T.B. Üstün, pages 33–57. New York: Cambridge University Press.

Pennell, B., J.A. Harkness, R. Levenstein, and M. Quaglia. 2010. "Challenges in Cross-National Data Collection." In *Survey Methods in Multinational, Multiregional, Multicultural Contexts*, edited by J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P. Mohler, B. Pennell, and T.W. Smith, pages 269–298. Hoboken, NJ: John Wiley & Sons.

Podmore, D., D. Chaney, and P. Golder. 1975. "Third Parties in the Interview Situation: Evidence from Hong Kong." *The Journal of Social Psychology* 95: 227–231. Doi: http://dx.doi.org/10.1080/00224545.1975.9918708.

Pollner, M. and R.E. Adams. 1994. "The Interpersonal Context of Mental Health Interviews." *Journal of Health and Social Behavior* 35: 283–290.

Pollner, M. and R.E. Adams. 1997. "The Effect of Spouse Presence on Appraisals of Emotional Support and Household Strain." *Public Opinion Quarterly* 61: 615–626. Doi: http://dx.doi.org/10.1086/297820.

Schwartz, S.H., A. Bardi, and G. Bianchi. 2000. "Value Adaptation to the Imposition and Collapse of Communist Regimes in East-central Europe." In *Political Psychology: Cultural and Cross-cultural Foundations*, edited by S.A. Renshon and J. Duckitt, pages 217–237. London: Macmillan.

Shah, J. 2003. "Automatic for the People: How Representations of Significant Others Implicitly Affect Goal Pursuit." *Journal of Personality and Social Psychology* 84: 661–681. Doi: http://dx.doi.org/10.1037/0022-3514.84.4.661.

Smith, P.B., M.H. Bond, and Ç. Kağıtçıbaşı. 2006. *Understanding Social Psychology Across Cultures: Living and Working in a Changing World*. London: Sage Publications.

Smith, T.W. 1997. "The Impact of the Presence of Others on a Respondent's Answers to Questions." *International Journal of Public Opinion Research* 9: 33–47. Doi: http://dx.doi.org/10.1093/ijpor/9.1.33.

The World Bank, "GNI per Capita, Atlas Method (current US$)" Accessed July 26, 2012. Available at: http://data.worldbank.org/indicator/NY.GNP.PCAP.CD.

Ting-Toomey, S. 1999. *Communicating Across Cultures*. New York: Guilford Press.

Tourangeau, R. and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133: 859–883. Doi: http://dx.doi.org/10.1037/0033-2909.133.5.859.

Triandis, H.C. 1989. "The Self and Social Behaviour in Differing Cultural Contexts." *Psychological Review* 96: 506–520. Doi: http://dx.doi.org/10.1037/0033-295X.96.3.506.

Triandis, H.C. 1995. *Individualism and Collectivism*. Boulder, CO: Westview Press.

Van de Vijver, F.J.R. 2003. "Bias and Equivalence: Cross-Cultural Perspectives." In *Cross-cultural Survey Methods*, edited by J.A. Harkness, F.J. R. van de Vijver, and P.P. Mohler, pages 143–156. Hoboken, NJ: Wiley.

Van Hemert, D.A., F.J.R. van de Vijver, Y.H. Poortinga, and J. Georgas. 2002. "Structural and Functional Equivalence of the Eysenck Personality Questionnaire Within and

Between Countries." *Personality and Individual Differences* 33: 1229–1249. Doi: http://dx.doi.org/10.1016/S0191-8869(02)00007-7.

Welkenhuysen-Gybels, J. and J. Billiet. 2001. "The Impact of Third Party Presence in Survey Interviews on the Measurement of Political Knowledge." *Acta Politica* 36: 287–306.

Received June 2013
Revised May 2015
Accepted June 2015

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 699–722, http://dx.doi.org/10.1515/JOS-2015-0041

Frameworks for Guiding the Development and Improvement of Population Statistics in the United Kingdom

James Raymer[1], *Phil Rees*[2], *and Ann Blake*[3]

The article presents central frameworks for guiding the development and improvement of population statistics. A shared understanding between producers and users of statistics is needed with regard to the concepts, data, processes, and outputs produced. In the United Kingdom, population estimates are produced by conducting decennial censuses and by estimating intercensus populations through the addition and subtraction of the demographic components of change derived from registers of vital events and from a combination of administrative data and surveys for internal and international migration. In addition, data cleaning, imputation, and modelling may be required to produce the desired population statistics. The frameworks presented in this paper are useful for aligning the required concepts of population statistics with the various sources of available data. Taken together, they provide a general 'recipe' for the continued improvement and expansion of official statistics on population and demographic change.

Key words: Population statistics; frameworks; data sets.

1. Introduction

Population statistics are used by national governments to distribute money across local governments for managing services, such as healthcare and education. They are used to establish the boundaries for political constituencies and to provide denominators for other measures, such as fertility or unemployment rates. They are used to manage and plan for future water, power, and sanitation needs. They can also be used to decide the number and types of homes to be built and design centres for shopping and leisure activities. Companies use population statistics to target their goods and services at specific groups of people (Boyle and Dorling 2004) and for workforce planning purposes. In other words,

[1] Australian National University, School of Demography, ANU College of Arts and Social Sciences, 9 Fellows Road, Acton Act 2601, Australia. Email: james.raymer@anu.edu.au
[2] University of Leeds, School of Geography, University Road, Leeds, West Yorkshire LS2 9JZ, UK. Email: p.h.rees@leeds.ac.uk
[3] Office for National Statistics, Segensworth Road, Fareham, Hampshire PO15 5RR, UK. Email: ann.blake@ons.gsi.gov.uk

Acknowledgments: An earlier version of this article was prepared for the Office for National Statistics under a Contract for the Provision of Research Services in Statistical Methodology, Contract Ref PU-10/0141 and under the direction of Ann Blake, available at http://www.ons.gov.uk/ons/guide-method/method-quality/imps/latest-news/conceptual-framework/index.html. The University of Southampton was supported in the delivery of the report through an Academic Collaboration Agreement with the University of Leeds. The authors would like to thank Peter Boden, James Brown, George Disney, Nikolas Lomax, Paul Norman, John Stillwell and everyone else who commented and provided feedback on the report.

population statistics are central for understanding society and societal change and are widely used in comparisons with other countries and areas within countries (Stone 1971; United Nations 1975).

Published outputs of population statistics are the result of matching available data to particular needs or concepts. The concepts are driven by user demand. In some cases, the data are processed or combined with other information to make statistics. Furthermore, the published statistics may not meet all of the needs of various user groups. As this article highlights, there are many requirements and types of population statistics. However, there are rarely single sources of information that cover a wide variety of needs. Instead, population statistics bring together data from many sources, all with their relative strengths and weaknesses.

The production of population statistics is further complicated because populations are both dynamic and heterogeneous. They change continuously according to the addition of births, subtraction of deaths and addition or subtraction of migrants. These processes are influenced by the social and cultural environments, economic environments and natural and built environments in which the populations live, as well the intersections between them (Bycroft 2011). In order to understand population statistics, one must first realise that they only represent a 'snapshot' of a population at a particular time or a flow between two fixed time points.

With such a demand and need for information on populations and their demographic behaviours, trusted, independent, and robust information about the size, structure and characteristics of a population is seen to be an essential underpinning of a modern society (Statistics New Zealand 2011). Such information is essential for improving the well-being, prosperity and legitimacy of modern democratic institutions and society alike. It is therefore vital not only that the statistics are reliable and robust, but also that users understand how the different statistics are compiled, how they relate to each other and what each variable actually represents. To achieve this, population information needs to be publicly available, transparent and understandable.

To help with the task of assembling population statistics for a country, national statistics offices require frameworks or agreed sets of concepts and methods, aligned to those available through current collection instruments and databases. Shared understanding of the frameworks within an organisation and across users facilitates the production and improvement of population statistics. In this article, we show how frameworks for population statistics can provide the main basis for achieving this. We use the United Kingdom (UK) as the illustration, but the frameworks may be generalised or adapted to other countries and data systems.

2. Frameworks for Population Statistics

In 1971, Richard Stone defined a system of social and demographic statistics, which he then refined for the United Nations in 1975. The system covered the whole range of government statistics, starting with demographic stocks and flows, moving on to families and households, social class and stratification, income and wealth, housing and the built environment, the use of time, social security and welfare, learning and educational services, employment, health, and public order. Stone's system also introduced the concept of the life course, which is so important in contemporary social science research. Stone's system represents an important and basic foundation for how we think about official statistics today. Our frameworks presented in this article focus on demographic stocks and flows, which form the basic elements of the wider design set out by Stone.

In this section, we first present a framework for official statistics and then illustrate how this can be applied to produce statistics on the usual resident population. We also present a framework for the underlying mechanism of population change. The frameworks are designed to be general across the many different ways of producing population statistics. For illustration, we describe how they are being used by the United Kingdom's Office for National Statistics (ONS) in their Beyond 2011 Programme, which is "taking a fresh look at options for the production of population and small area socio-demographic statistics for England and Wales" (ONS 2013a). This includes exploring options based on administrative data (ONS 2013b).

2.1. *Official Statistics*

In designing a framework for population statistics, it is useful to first think about the activities and stages involved in producing official statistics (Laux 2002). In Figure 1, the functions of Society, Concepts, Data, Processing, Outputs, and Validation are presented. *Society* represents the economic and social conditions of the country producing the statistics. It determines the type of information required. *Concepts* refer to particular statistics of interest, such as the usual resident population, rates of unemployment, welfare provision or persons present without citizenship. *Data* are any information gathered concerning the required statistic, usually obtained from censuses, surveys or administrative registers. *Processing* represents data cleaning, imputation, combining two or more information sources through matching or proportioning, and statistical modelling to ensure that the data more closely match the required concept. *Outputs* are the published statistics or estimates. *Validation* is the procedure of assessing the quality of the published statistics in relation to the concepts required, often resulting in periodic revisions or improvements in data collection or processing.

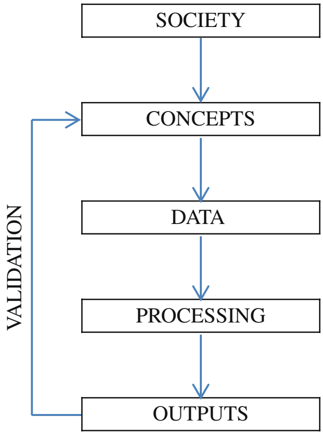


Fig. 1. A framework for producing official statistics

2.2. Population Statistics

The general framework for official statistics presented in Figure 1 may be applied to the production of population statistics. Many types of population statistics can be produced, such as the population present at the time of a census or survey (*de facto* population), the population considered permanent or usually resident (*de jure* population), the population considered temporary (tourists, business travellers, short-term migrants), the population born in the country (native born) and the population born abroad (foreign born). We focus on the statistics that form the basic requirement for most national statistical offices, namely the usual resident population by age and sex, living in area *i* at time *t*. Adopting the United Nations (2007) definition, usual residents are all persons who reside, or intend to reside, in a place continuously for either most of the last twelve months or for twelve months or more. This includes nationals, foreigners, undocumented persons, applicants for asylum and refugees. Counts of usual residents provide ". . . the best indication of where people will demand and consume services, and . . . is therefore most relevant for planning and policy purposes" (United Nations 2007, 132).

To illustrate how the framework for official statistics can be applied in the production of statistics of usual residents by age, sex and local authority in the United Kingdom, consider the expanded framework presented in Figure 2. Here, more details are provided under the main headings of society, concept, data, processing, and output. Validation is not included in the diagram, but of course it is necessary for the continued improvement of the population statistics over time. We also include two columns covering background and other considerations, which may be interpreted as the context and specific factors for a particular society, respectively.

There are many issues to consider in the production of population statistics, which are largely driven by the need for distributing national resources, planning, and social welfare. What makes the UK different from other countries is its economy and culture, and this varies over time depending on, for example, the political climate and available resources. The UK does not have a population register or common identification number or code system (Poulain et al. 2006, 112–113), and has historically relied on decennial censuses to produce population statistics, with demographic estimation (described in Subsection 2.3) used to produce statistics for years between censuses. Note that care must be taken, when comparing censuses, to allow for coverage differences as well as conceptual differences in defining the population of interest. Very often, to aid these comparisons, several different population bases will be provided by national statistical offices for the basic population counts.

As mentioned previously, at any time, there are always several populations that may be measured or conceptualised. In the UK, a midyear (30 June/1 July) 'usual resident' population estimate is produced. This statistic is recommended by the United Nations for international comparability, although the practicalities of identifying usual residents may differ from country to country. Furthermore, the size and characteristics of the population may vary greatly, depending on both the time of day and day of the year measured. The usual resident population represents a 'night-time' measure, which captures the population where it sleeps. A 'day-time' measure captures the population where it goes to school, to work, to shop or to pursue leisure activities. Finally, it is important to be clear about the

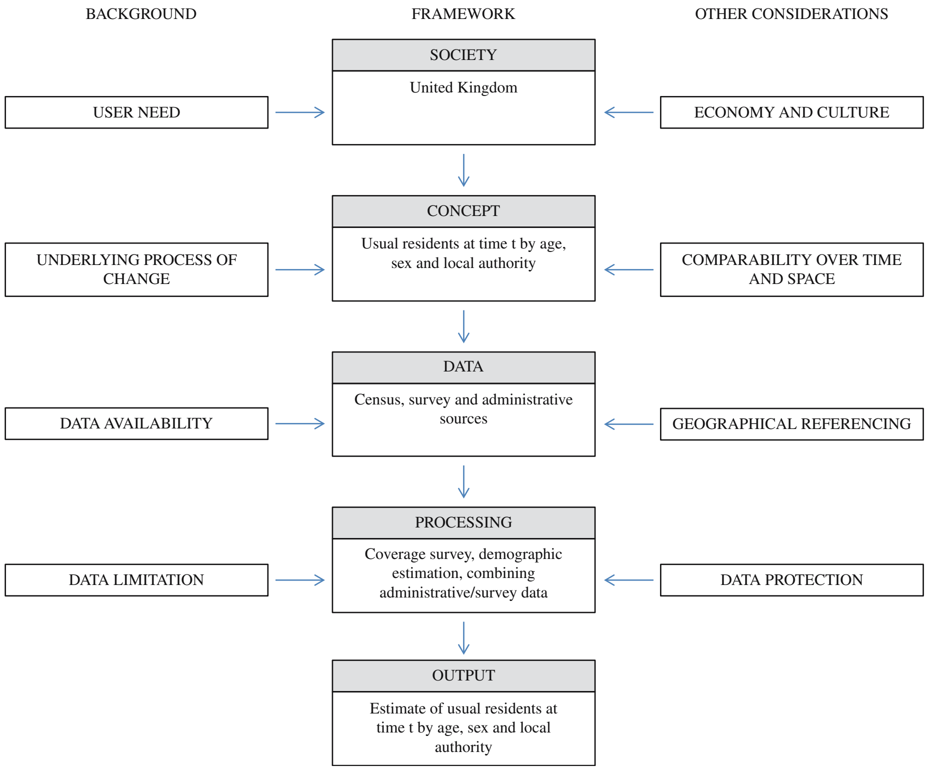

BACKGROUND
FRAMEWORK
OTHER CONSIDERATIONS
SOCIETY
United Kingdom
USER NEED
ECONOMY AND CULTURE
CONCEPT
Usual residents at time t by age, sex and local authority
UNDERLYING PROCESS OF CHANGE
COMPARABILITY OVER TIME AND SPACE
DATA
Census, survey and administrative sources
DATA AVAILABILITY
GEOGRAPHICAL REFERENCING
PROCESSING
Coverage survey, demographic estimation, combining administrative/survey data
DATA LIMITATION
DATA PROTECTION
OUTPUT
Estimate of usual residents at time t by age, sex and local authority

Fig. 2. A framework for producing population statistics in the United Kingdom

criteria for the inclusion or exclusion of individuals. Business travellers and visitors are usually excluded from official and international statistics on population and migration. Temporary workers may be included in official estimates of short-term migrants, but not in the usual resident population.

Regardless of the population concept being measured, it can only ever represent a snapshot of the continuous process of population change. This is further illustrated in Figure 3, which shows how populations present in an area may change over time. If the interest is tracking the changes of usual residents, each of the entry and exit components needs to be defined accordingly. Here, for example, only those who change their country of usual residence would be included as international entries or exits. All other international entries and exits would be excluded from the official statistics. Likewise, the

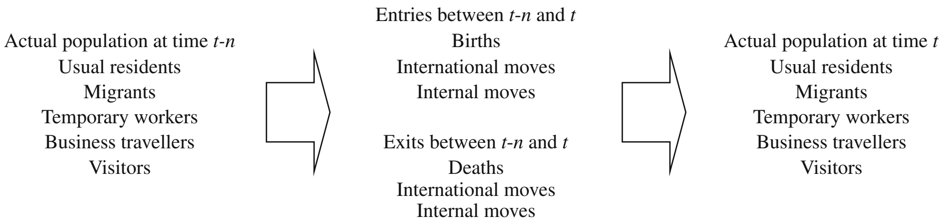


Fig. 3. The dynamics of actual population change for a geographic location within a country

births, deaths and domestic movements of temporary workers, business travellers and visitors would also be excluded.

There are two ways in which estimates of the usual resident population can be obtained: (1) using a source or combination of sources that can be used to count the population at a particular point in time and (2) combining sources of information on population stocks (e.g., recent census) with information on demographic events over time (see Subsection 2.3). The sources of demographic data may be censuses or surveys, or may arise from administrative processes including the registration of births and deaths or the need to access healthcare or state benefits. Each data source has limitations with regard to collection and storage. This usually concerns the amount of detail included and whether it contains individual information or an aggregation of individual information. It also concerns the amount of characteristic detail included in the data. Typically, censuses and surveys include much characteristic information for understanding societal differences and change, whereas administrative data contain only basic information necessary for operational purposes (see Subsection 2.4).

It is difficult to design data sources to measure particular population concepts and to capture all of the population required. Moreover, in the case of administrative data, population measurement is not the primary purpose. Thus, national statistical offices often use additional methods to improve the data, aligning the information with the required concept. In the UK, for example, coverage surveys and imputation are used to produce the census estimates (ONS 2013c). For estimates between census years, demographic event data are combined with statistical estimation used to augment the data on internal and international migration. Finally, there is an increasing need to provide measures of uncertainty – or, conversely expressed, of accuracy – with the estimates so that they may be interpreted correctly (see Section 3). The outputs or published statistics are estimates, since they are only able to come close to the ‘true’ conceptual measure but do not match it exactly. These statistics are also published with supporting information about how they have been produced, allowing the user community to make informed judgements on the current state of the population and its likely future.

To summarise, it is important to understand the various factors involved in producing population statistics. The framework for population statistics starts with the societal context and extends to the required published outputs. Next, we describe how the usual resident population can be aligned with the underlying mechanism for population change, involving the demographic accounting equation.

2.3. *Underlying Mechanism of Demographic Change*

There are two ways that population statistics can be produced: enumeration and demographic accounting. In practice, both are often used to understand the population change occurring and to verify the quality of the estimates being produced. In the UK, censuses are used to provide accurate estimates of the population every ten years. In between censuses, midyear estimates are produced by rolling forward the age- and sex-specific census estimates based on the number of births, deaths, domestic migrations and international migrations that occur within each year utilising the demographic accounting equation (see Figure 4 and equations below). To maintain an accurate picture of population

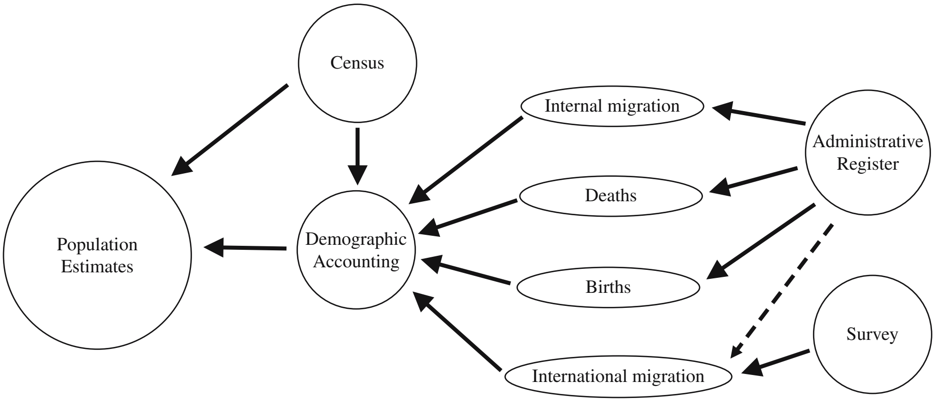


Fig. 4. Overview of population estimation in the UK

change over time, care must be taken to match the model used for demographic estimation with the nature of the data available. Note that for the production of international migration statistics in the UK, both survey (major source) and administrative (minor source) data are used (represented by solid line and dashed line arrows, respectively).

When producing estimates based on demographic accounting, it is important to take into consideration the close dependence between the migration definition and the demographic accounting model. There are two types of migration data: events and transitions. Events refer to the number of moves that occur within a particular time period, whereas transitions refer to data collected on places of residence at two points in time. The demographic accounting model essentially rolls forward the population estimate from the last reliable estimate (e.g., census) to successive years until the next reliable estimate, where the error of the process can be assessed. The same demographic accounting process is also used for estimating future population totals (i.e., population projections), except that observations on fertility, mortality and migration are replaced with estimated future (projected) figures.

Demographic change accounts based on events are presented in Table 1. These accounts represent an extension of those originally developed by Stone (see United Nations 1975). The variable M_{ij} represents internal migration events from one area i to another area j. The n subscript denotes the number of areas. Note that we ignore internal migration events when $i = j$. Instead, we enter terms R_i, which are accounting balances, the result of subtracting from the start population all possible exit events. Total (internal) outmigrations from each area are denoted by M_{i+} and total (internal) inmigrations to each area are denoted by M_{+i} (see the Table 1 notes for explicit definitions). Here we use the subscript i as a general index for a region. The I_i variable signifies the number of immigration events from outside the system of interest and E_i tabulates the corresponding emigration events. The vital events of death, D_i, and birth, B_i, complete the flows in the table. The sum of the numbers in the rows adds up to the populations at the beginning of the time interval, $P_i(t)$. The balancing term is obtained by subtracting the total number of outmigrations, emigrations and deaths from the population at the beginning of the time interval, that is,

$$R_i = P_i(t) - M_{i+} - E_i - D_i. \qquad (1)$$

Table 1. A demographic accounting framework for population statistics

	To destination:								
From origin:	Area *1*	Area *2*	. . .	Area *i*	. . .	Area *n*	Emigrations	Deaths	Total
Area *1*	R_1	M_{12}	. . .	M_{1i}	. . .	M_{1n}	E_1	D_1	$P_1(t)$
Area *2*	M_{21}	R_2		M_{2i}		M_{2n}	E_2	D_2	$P_2(t)$
:	:	:		:		:	:	:	:
Area *i*	M_{i1}	M_{i2}	. . .	R_i	. . .	M_{in}	E_i	D_i	$P_i(t)$
:	:	:		:		:	:	:	:
Area *n*	M_{n1}	M_{n2}	. . .	M_{ni}	. . .	R_n	E_n	D_n	$P_n(t)$
Immigrations	I_1	I_2	. . .	I_i	. . .	I_n	*0*	*0*	I_+
Births	B_1	B_2	. . .	B_i	. . .	B_n	*0*	*0*	B_+
Total	$P_1(t+1)$	$P_2(t+1)$	. . .	$P_i(t+1)$	. . .	$P_n(t+1)$	E_+	D_+	

Definitions of variables and subscripts: $P =$ population, $R =$ balancing terms, $M =$ internal migrations (within a country), $E =$ emigrations, $I =$ immigrations, $B =$ births, $D =$ deaths, $0 =$ structural zeroes, $t =$ time, $i =$ subscript for area, $n =$ number of areas and $+ =$ summation over areas. Total internal outmigrations from area $i = \Sigma_{j=1 \text{ to } n, j \neq i} M_{ij} = M_{i+}$. Total internal inmigrations to area $i = \Sigma_{j=1 \text{ to } n, j \neq i} M_{ji} = M_{+i}$.

The variables in the columns of Table 1 add up to the populations at the end of the time interval. We can compute these by adding to the balancing term the total inmigrations, immigrations and births, that is,

$$P_i(t+1) = R_i + M_{+i} + I_i + B_i. \tag{2}$$

If we combine these two equations, the balancing term cancels out, and we obtain the familiar components of the population change equation:

$$P_i(t+1) = P_i(t) - M_{i+} - E_i - D_i + M_{+i} + I_i + B_i. \tag{3}$$

Although the demographic accounting model is applied similarly, it is important to understand how the movement and transition concepts of migration differ. To illustrate, consider Figure 5, where the vertical axis represents space and is divided into two regions and the horizontal axis represents time (one time interval). The lines on the graph (A, B) plot the location of two people. Person A starts in Region 1 and migrates to Region 2 at time $t + 0.7$ and then remains there to be recorded in Region 2 at time $t + 1$. Person B starts in Region 2 and migrates to Region 1 at time $t + 0.2$ but then migrates back to Region 2 at time $t + 0.4$. These two persons make one move from Region 2 to Region 1 and two moves from Region 1 to Region 2. Person A makes one transition, from Region 1 at time t to Region 2 at time $t + 1$. Person B is recorded in Region 1 at time t and at time $t + 1$ and so fails to make a transition. Note that the net migration between regions is $+1$ for Region 2, whether migration is measured as a move or as transition.

The demographic accounts in Table 1 are specified for the whole population (all ages). When the population is rolled forward from one year to the next, the accounts and associated population change equations can be specified for each age, using period cohorts. The only difference is that the birth terms are omitted from the accounts and instead are used as the starting population in the accounts for the new-born period cohort. Furthermore, Table 1 is specified for one country with several subnational regions and one

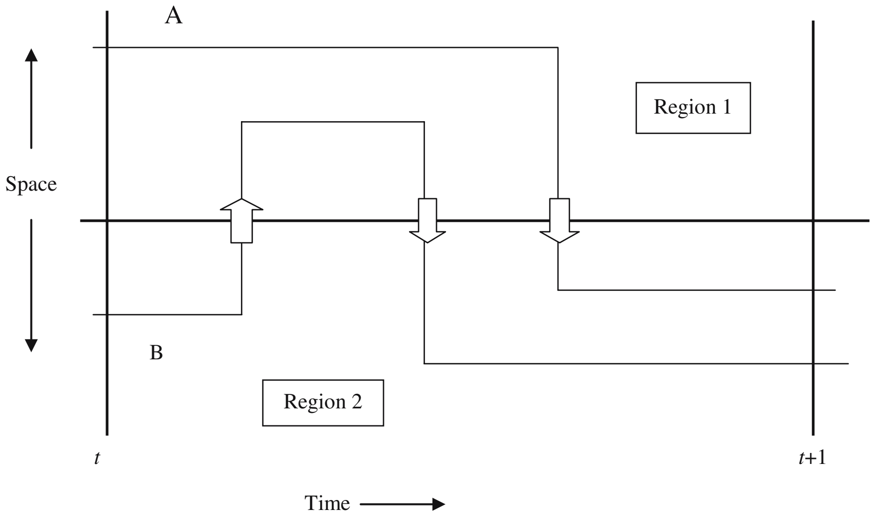


Fig. 5. A time-space diagram illustrating different migration measurement concepts

external region, the rest of the world. However, this spatial system can be expanded to include a large number of areas. The system can also be collapsed to estimate the national population change. Furthermore, the demographic accounting framework could also designate internal regions as separate countries (e.g., the 28 members of the European Union) or a combination of both national and subnational units can be employed (see, e.g., Kupiszewski and Kupiszewska 2008).

If an end of interval population, $P_i(t+1)$, is independently available from a population register or an end-of-decade census, then the error of estimation may be calculated. To reconcile population change through the interval with the start and end population stocks, some corrections must be made to the component terms using assessments of the likely errors involved. To estimate the population at time $t+1$, the demographic accounting model therefore requires a population stock at time t and information about births and deaths (natural increase measures) and domestic and international migration between time t and $t+1$.

Perhaps the most accurate information we have is on the number of live births and deaths for locations in the UK over time. This is because all births and deaths have to be registered by law. Births are published according to the sex and birthplace of the child and the age and residential location of the mother. Deaths are recorded for all persons by age, sex and residential location. Migration data, on the other hand, are obtained from general-purpose censuses, surveys or administrative registers. Unlike fertility or mortality, the practical measures of migration obtained from these sources often do not coincide with theoretical or contextual definitions of migration (Bell et al. 2002). For example, Raymer and Smith (2010, 703) describe migration as “a loosely defined process that represents the relocation of people during a period of time that causes them to relinquish the ties with their previous locality.” Migration can involve people moving within a country and within localities, as well as across international borders. The factors that separate migration from other forms of mobility (e.g., daily commuting, weekday/weekend commuting, holiday visits or seasonal moves) are generally the distance travelled and the length of time spent in the destination (or away from the origin). In practical terms, migration can be defined as relocations between administrative areas and mobility as relocations within areas. Intra-area migration is not required for population change accounts (Table 1). ONS does not restrict the spatial scale of residential migration, as any such classification would be arbitrary. Therefore, a residential migration within the same suburb is still a migration that may have relevance for the estimation of the population of very small areas, such as output areas or postal sectors.

2.4. Population Characteristics

Users of population statistics are interested in the characteristics of the population and how those characteristics change over time. Age and sex are considered the baseline characteristics required for population statistics because many other attributes have a close relationship with them. For understanding change or differences between population groups, it is useful to have more detailed attribute information, depending on the need or users. For example, for those interested in the integration of immigrants, information on the foreign population, their levels of education and their occupations are useful. For those wishing to set migration policy, understanding the reasons or drivers for migration is important. For provision of services, information on population health, number of children and economic activity are useful. The availability of data on characteristics will depend on the type of information required and the data available to meet the need.

2.5. Geographic Classifications

Geographic classifications are fundamental for understanding society and population change. There are many different ways of representing geography, depending on the location information available in the data source. Statistics are usually produced for local-authority districts, counties and regions in England and Wales, and electoral wards and council areas in Scotland. Sometimes, the actual geography is not of interest but rather the area type, such as urban, rural or coastal. Geographic information is important for planning schools, hospitals, workforces, as well as for comparing different spatial patterns of residence according to ethnicity or density.

3. Output Uncertainty

Information about uncertainty in the published statistics can help users to gain a better understanding and use of the statistics. If the sources of uncertainty are known, they can also be used to inform data providers where improvements can be made to their data collection or estimation methodologies and to evaluate the scale of improvement achieved. For example, in 2009, the UK Statistics Authority report "Migration Statistics — The Way Ahead?" recommended that "ONS should flag the level of reliability of individual local authority population estimates." As part of the Migration Statistics Improvement Programme, ONS responded by developing methods for measuring error and defining confidence intervals for the England and Wales local authority midyear population estimates (ONS 2012b). They also developed a table of key indicators, which can be used to identify local authorities with characteristics associated with greater likelihoods of uncertainty in their midyear estimates, such as the proportion of students present.

Uncertainty can come into the population estimation model in many ways. At the onset, the base population taken from, for example, a recent census may contain error. The components of change may contain errors. For example, in the UK, internal migration and international migration are considered the most problematic in terms of population error. Registrations of births and deaths, on the other hand, are considered highly reliable since they have a very clear legal framework and long history of data collection. The degree of uncertainty in the time series of population estimates is most often revealed when the results of the next census become available and often, as a result, the annual population estimates are revised to coincide with both the most recent and the previous census. To achieve a fully consistent time series of population change, it may also be necessary to revise the components of change. The vital statistics (births and deaths) normally do not need revision, though the associated demographic rates are often revised as a result of changes to the population at risk.

Within the area of survey methodology, the framework for survey error provided by Groves et al. (2009) provides a useful basis for considering the errors underlying data collected by surveys. The framework is also useful when considering data generated by a

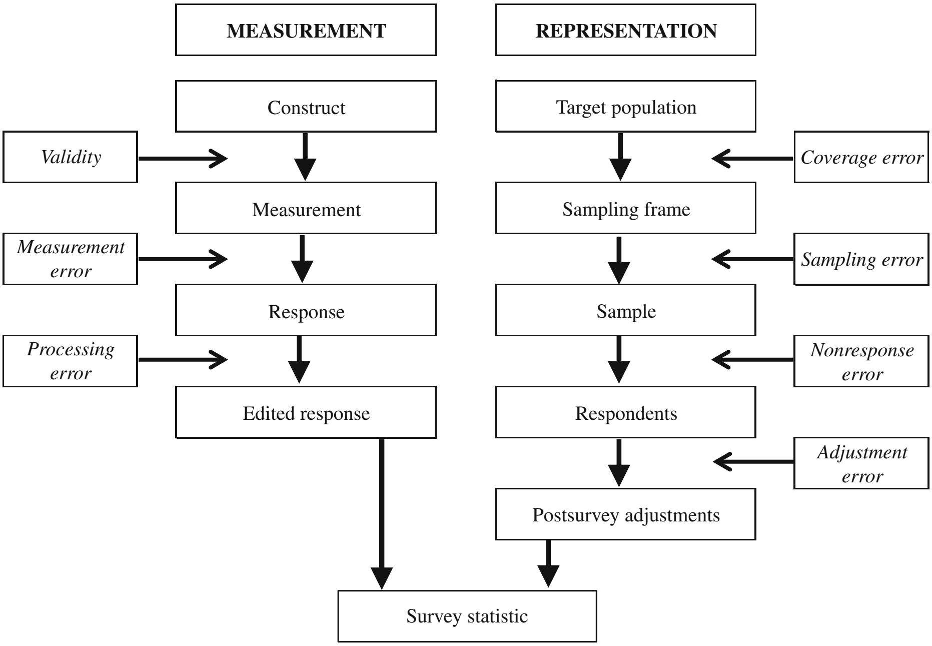


Fig. 6. A framework for survey error. Source: Adapted from Figure 2.5 in Groves et al. 2009, 48

census or administrative system. Figure 6 sets out the framework, showing the processes used in gathering population data and where errors arise. Errors may be systematic or random, and they affect the uncertainty of the statistics in two different ways: systematically or randomly. An important purpose of including uncertainty measures in official statistics is to prevent users from making inferences that are not supported by the data.

3.1. Accuracy

Accuracy refers to the closeness of the estimates to the true values. When producing population estimates, we know that accuracy is greater for larger populations as it takes more to influence their population change. We also know that accuracy decreases the further away in time the estimates are from a census. For example, in the UK, we would expect the 2012 population estimates to be considerably more accurate than the 2020 population estimates will be when they are produced, because the most recent (available) census occurred in 2011.

Bias, which is caused by systematic errors, refers to whether the estimates have higher or lower values on average in comparison to the true values. The sources of bias include both those that are known beforehand and those that appear unexpectedly. In general, the known biases are those that involve areas that are difficult to count, such as those with highly mobile populations, including students, migrants, homeless persons or armed-forces personnel, and those with high deprivation, unemployment or crime. These areas contain populations that are less likely to fill in questionnaires or register in an administrative source. For example, the age profile of internal migration in England and

Wales, as measured using data from the National Health Service Central Register (NHSCR), exhibits a markedly lower level for males than females aged 15–29 years (Raymer et al. 2011, 80–81). In 2007, there were 55 per cent more females in the 20–24 year old age group changing their General Practitioner than males. If we believe the two levels should be similar for these ages, then overall, it can be said there is a 15 per cent undercount of males or about a seven per cent undercount for all migrants in the NHSCR data. Overcounting may also be an issue in some administrative sources as a result of lags or failures in deregistration. Finally, unexpected biases may result from a particular methodology used to estimate the population totals or to a change in the way data are collected. For example, when net migration rates are used in demographic accounting models instead of origin-destination-specific rates, there is a tendency to overestimate areas of growth and underestimate areas of decline (Rogers 1990).

Information on the potential sources of inaccuracy (due to systematic or random errors, or both) can help users understand the range of plausible totals for a particular area and the reasons why their area or group is considered to be estimated with more or less uncertainty. These two aspects of uncertainty – size and causes – can also point to areas where methodology could be improved or where additional data should be gathered. They also provide a more realistic picture of population change. Furthermore, as national statistics offices continue to improve their methodology for estimating populations, the availability of measures of uncertainty could inform the extent of the improvement. The actual measures of uncertainty may vary from probabilistic predictive intervals to summary statistics, for example, Mean Algebraic Percentage Error for systematic deviation (bias) or Mean Absolute Percentage Error as a measure of precision, or spread, due to random variation. Finally, difficulties may arise in calculating the uncertainty measures, especially for population estimates where no information exists on the true values (e.g., postcensus population estimates and projections). Here, statistical models may be used to estimate the uncertainty.

3.2. The Challenge of Measuring Migration

Measuring migration is particularly difficult, and deserves special treatment here. First, long-distance migration is a relatively rare event in the lives of most people. Second, many data sources are not designed to capture migration specifically; rather, migration measures are a by-product obtained by additional analyses of administrative or general purpose databases. This has implications for measurement (duration, coverage), accuracy and detail. Thus, when considering appropriate data sources to measure migration, there are inevitably a number of obstacles to be overcome:

1. Difference between what the available data measure and what is required by the population estimation model (Rees 1985).
2. Small sample size of most survey data sources. Because most people do not migrate within a given time period, small sample size leads to estimates with very high standard errors and, consequently, the inability to use sample data for subnational areas without auxiliary information.
3. Coverage differences between the populations targeted in survey and administrative sources and the coverage needed in the model for generating population estimates or projections.

4. Underreporting of migrations (as events), which also manifests itself as excessive lags in the reporting of address changes. Underreporting may occur as there is no legal requirement to report a migration event in the UK, and individuals may be slow to advise administrations of changes of address where either (a) there is no incentive to do so or (b) there is no need for them to access a particular service.
5. Misreporting of destinations or origins of migration, depending on the data set involved. Misreporting occurs for several reasons, and is primarily related to survey or census data. Respondents arriving in the UK may be unsure or vague about their destination. Respondents to household surveys or the census may inaccurately recall where they were living some time ago, particularly if they have experienced high mobility in between.

What consequences does a migration measure have for population estimation? Rees (1985) points out that you need to match the concept used to measure migration with the concept used in the population estimation equation. The standard estimation equation assumes that migration is measured as a count of relocation events alongside the event counts of births and deaths. It is possible to develop a population estimate equation that uses migration measured as transitions, but the model is complicated for the UK for two reasons. First, there is no population register from which transitions can be counted for the population as a whole. In the UK, a census is administered only once in ten years and the main migration question captures transitions only for the year prior to the census. Second, the system of population statistics used by ONS is based on the movement concept using data on events. This is why ONS applies adjustment factors to the Patient Register Data System (PRDS)-derived migration measures which are based on the transition concept. The adjustment ratios are computed by comparing counts of moves (migrations as relocation events) between Former Health Authorities in a data set based on the NHS Central Register against the counts of transitions aggregated appropriately from the PRDS system (ONS 2012c).

3.3. The Challenge of Measuring Special Populations: The Armed Forces and Prisoners

The population estimation methods described above depend on being able to measure all demographic components of change for all population groups. This is often true for births and deaths, where the registration has a legal basis. For the internal and international migration components, some populations may be left out. For instance, the NHS register does not report the movements of the Armed Forces or the prison population. The current approach to estimating populations is to subtract, at the start-of-time interval, the Armed Forces population and the prisoner population from the start population stock and then add fresh stock estimates to the end population. The data for these populations are supplied by the Ministry of Defence and the Ministry of Justice, respectively.

4. Applications of the Population-Statistics Framework

In this section, we illustrate how the frameworks presented in Section 2 can be applied to local area immigration statistics and to combine various administrative data sources to measure the usual resident population.

4.1. Immigration to Local Authorities in the United Kingdom

As an illustration of how the population statistics framework may be applied, consider the user requirement for the population within a particular local authority in 2010. The need could be conceptualised as "the usual resident population within local authority *j*, 2010". Here, data are required that match the concept of "all migrants arriving in the UK and residing for at least a year in local authority *j*, mid-2009 to mid-2010" to allow estimation of the population. Next, assume that there are only two sources of information available: migrants identified in the International Passenger Survey (IPS) and National Insurance Number (NINo) registrations to non-UK nationals. As shown in Figure 7, the IPS identified 536 (95% confidence interval: 506 to 566) thousand persons coming from abroad between mid-2009 to mid-2010 who intended to stay at least twelve months (ONS 2013g) (Note that this estimate of immigration was revised upwards after the 2011 Census to 579 thousand). They were surveyed at the point of arrival and were asked about their destination in the UK. However, because of the relatively small sample size, this information is only reliable at the national level. For the same period, there were 668 thousand new NINo applications recorded for foreign citizens aged 16 years or over. The information was collected at the time of registration with reliable address information but no information on length of stay.

The main advantage of the IPS is that it measures "all migrants arriving in the UK intending to stay for at least a year." The main advantage of the NINo registration is that it measures those "residing in local authority *j*." Neither of these sources alone can provide data that meets the concept above. Additional information about the number of visitor and migrant switchers (i.e., those changing their intentions) for the IPS data and the lag between arrival and registration and the duration of stay for the NINo data would be required. The two data sources can be aligned by separating the foreign citizens aged 16+ years from the rest of immigrants in the IPS. It is then possible to produce estimates of "foreign citizens aged 16+ years arriving in the UK for at least a year residing in local authority *j*, 2010." This illustration shows how the framework may be used to take a user requirement and align the available data to the concept needed; it also shows the obstacles and limitations that often occur with such a process.

Figure 7 represents an illustrative example based on the more complex estimation process described in ONS (2011) based on work by Boden and Rees (2010), which has improved estimates of immigration to local authorities in England and Wales. The method divides the national estimate of immigration into different flows by purpose of immigration – for work, for study, as returning citizens or residents, and as dependants. Different administrative sources are used to distribute each "purpose group." For example, higher-education records, which cover all newly registered foreign students, are used to estimate the local-authority distribution of foreign student arrivals. This method replaces a less accurate process that hierarchically distributed yearly IPS immigration totals: (1) into regions according to three-year averaged Labour Force Survey estimates; (2) into 'intermediate' geographies based on IPS estimates; and (3) into local authorities based on the previous Census (2001) counts.

4.2. Combining Administrative Data to Estimate Usual Residents by Age and Sex

In this subsection, we describe how ONS is exploring the quality and application of administrative data sources, including combinations of them, to estimate usual resident

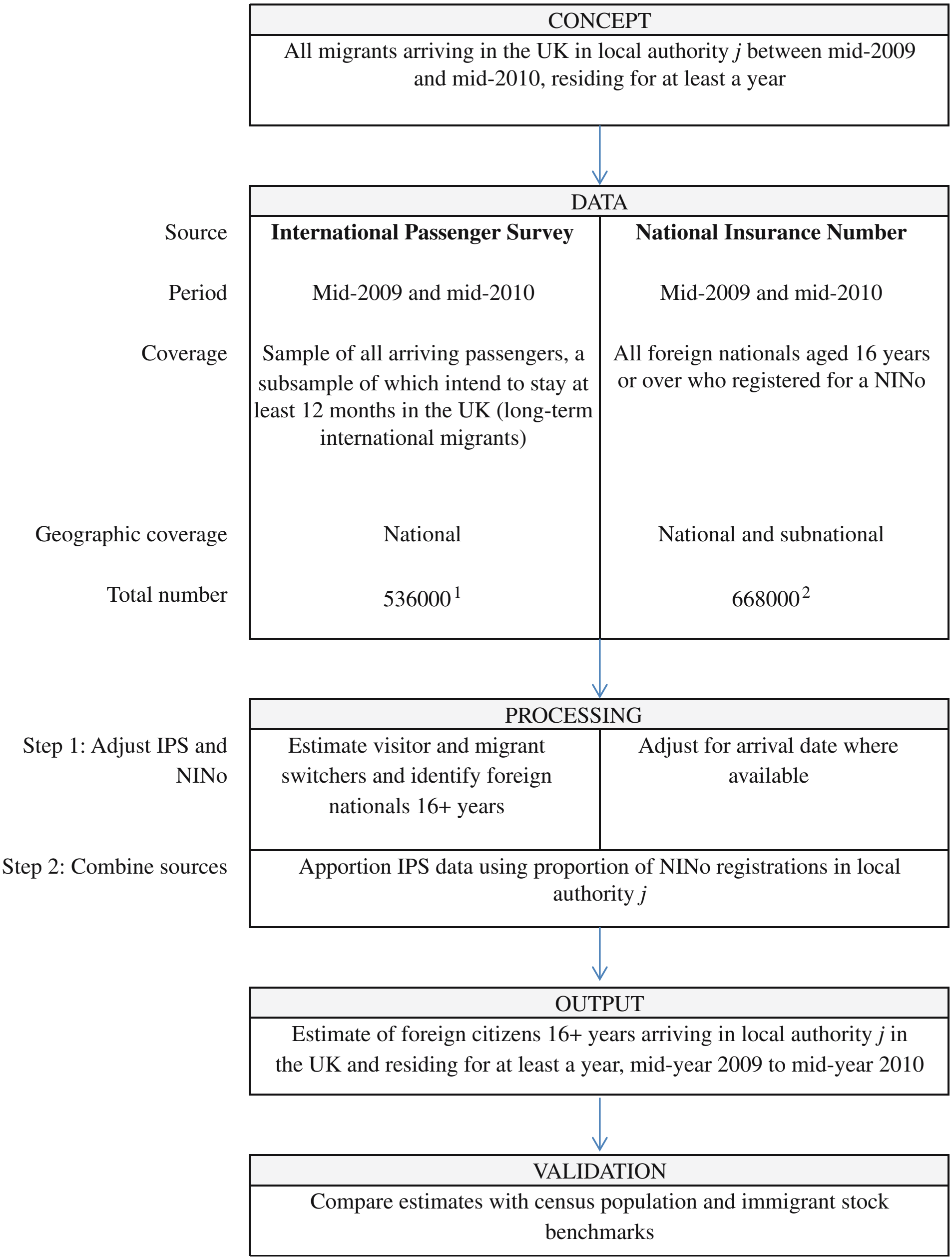

CONCEPT
All migrants arriving in the UK in local authority j between mid-2009 and mid-2010, residing for at least a year
DATA
Source
International Passenger Survey
National Insurance Number
Period
Mid-2009 and mid-2010
Mid-2009 and mid-2010
Coverage
Sample of all arriving passengers, a subsample of which intend to stay at least 12 months in the UK (long-term international migrants)
All foreign nationals aged 16 years or over who registered for a NINo
Geographic coverage
National
National and subnational
Total number
536000 1
668000 2
PROCESSING
Step 1: Adjust IPS and NINo
Estimate visitor and migrant switchers and identify foreign nationals 16+ years
Adjust for arrival date where available
Step 2: Combine sources
Apportion IPS data using proportion of NINo registrations in local authority j
OUTPUT
Estimate of foreign citizens 16+ years arriving in local authority j in the UK and residing for at least a year, mid-year 2009 to mid-year 2010
VALIDATION
Compare estimates with census population and immigrant stock benchmarks

Fig. 7. Application of the population statistics framework to combine International Passenger Survey (IPS) and National Insurance Number (NINo) registrations to measure international migration in the UK. Note: (1) Total immigration estimated by factoring up sample count; (2) a full count of new NINo registrations by foreign nationals.

populations as part of their ‘Beyond 2011’ programme (ONS 2012a). Administrative data may be used to directly or indirectly inform annual population estimates by age, sex and geography. Here, it is useful to refer to the frameworks presented in Figure 1, Figure 2 and Figure 6, as well as those presented in Zhang (2012). These frameworks can be used to assess how the data sources relate to the concepts we are concerned with, including how they capture information about location, births, deaths and migration, and to determine where potential sources of error may arise, respectively. The assessment may be conveyed visually using Venn diagrams as outlined in Figure 8 below. The aim of these diagrams is to illustrate how the concept of the usual resident population relates to the population covered by the administrative source (or sources) in a specific area *j* at a specific point in time *t*.

To understand the relationship between a particular administrative data source and the usual resident population, consider the UK National Health Service's patient register. General Practitioners (GPs) are the first point of contact for nearly all UK National Health Service (NHS) patients (NHS 2011). Most individuals and households are registered with a GP near their home. The NHS patient registers are used to maintain an accurate list of all persons registered with a GP, allowing the timely transfer of medical records and correct payments to doctors. ONS receives a list of everyone who is registered with a GP in England and Wales. This source of individual-level data has a specific administrative purpose and is not designed to specifically measure the populations.

The patient register contains the address details of patients; however, depending on the extract, the complete address may not always be available for analysis. Location information derived from just a post code may differ from that derived from the full address where these lie very close to the boundary of a local authority, for example. The quality of the address information is also dependent on individuals keeping their GP up to date with changes, or registering with a new GP if they have made a longer-distance move. Any failure or lag in updating address information results in measurement error

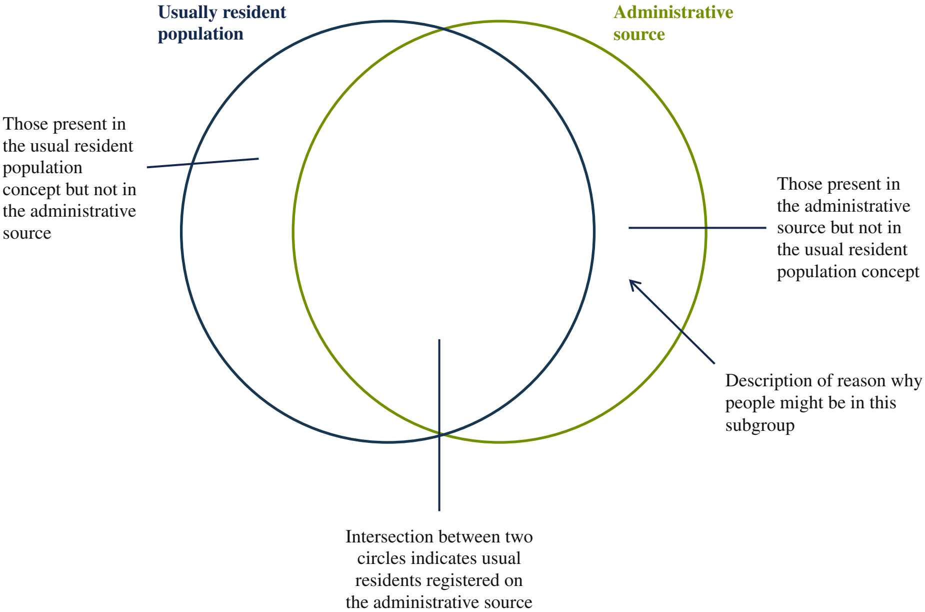

Usually resident population
Administrative source
Those present in the usual resident population concept but not in the administrative source
Those present in the administrative source but not in the usual resident population concept
Description of reason why people might be in this subgroup
Intersection between two circles indicates usual residents registered on the administrative source

Fig. 8. Understanding the relationship between an administrative data source and the usual resident population concept

arising in the location information on the patient register at a particular point in time. For persons who “live” at more than one address, the situation is more complex. For example, some children at boarding school register with a GP using their boarding school address, others using their parental address.

The coverage of the patient register is all persons registered with a GP in England and Wales. The population covered by the patient register includes the following subgroups not present in the usual resident population:

- Persons staying in the country for fewer than twelve months (*short-term migrants*);
- Patients not removed from the register because of unregistered deaths or when they have moved out of an area and not deregistered from the NHS (*list inflation or registration lag for outmigrants or emigrants*);
- Persons issued with new duplicate NHS numbers, for example when away from home (*multiple NHS numbers*).

Conversely, the usual resident population has subpopulations not recorded in the patient register:

- ○ Persons who have failed to register at their destination after an inmigration or immigration, more likely among young adult males (*nonregistration or registration lag for inmigrants or immigrants*);
- ○ Patients removed from the patient register when they have not moved (*erroneous list cleaning*);
- ○ Military personnel treated by their own health services (*armed forces*); and
- ○ Inmates of prisons treated by the prison medical service (*prisoners*).

In terms of timing, the Patient Register extract used by ONS is taken one month after the date at which the usual resident population is estimated to counteract some of the lag in changing GP registrations following moves. Note that the NHS Patient Register is not used directly as a population estimate. Instead, patients are matched between two registers one year apart to count patient transitions. These are converted to movements before being used in the demographic accounting equation that produces the new end of internal population estimate.

Another important administrative data source is the specific social security and revenue information held within the Department of Work and Pensions (DWP), referred to as the Customer Information System (CIS) (ONS 2013d). The CIS contains a record for all individuals that have been issued with a NINo. For the DWP, the primary purpose of the CIS is to store basic identifying information, such as name, address and date of birth. The extract of the CIS provided to ONS on the 2011 Census day (27 March) contained approximately 96 million records. In Figure 9, a Venn diagram is presented showing the differences between the usual resident population with a NINo and the CIS records.

The NHS patient register provides good coverage of the population and represents the timeliest source of administrative data currently available. The CIS also provides good coverage of the population but includes a large number of people who are no longer resident in the UK. Additional administrative sources include the English and Welsh School Censuses and Higher Education Statistics Agency data (ONS 2013e), which

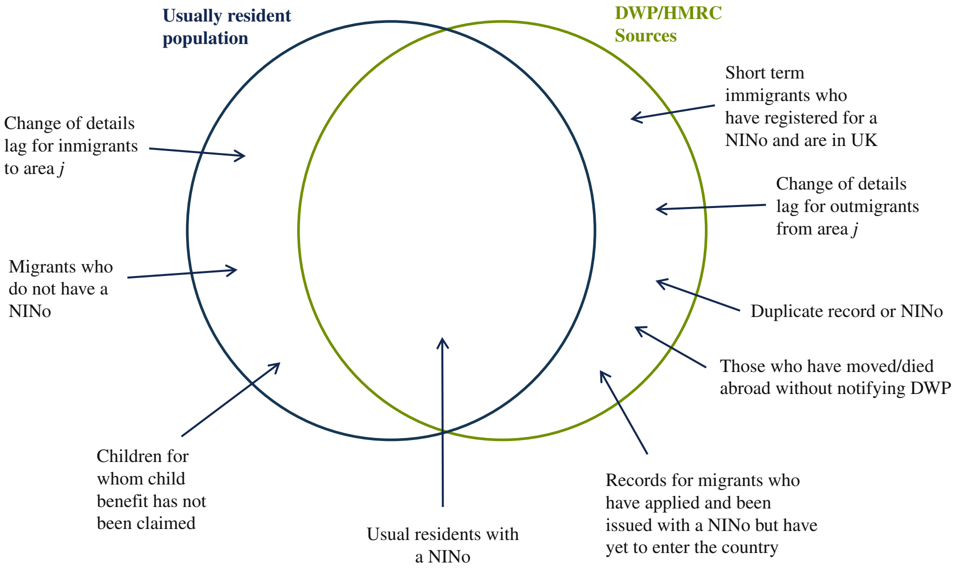
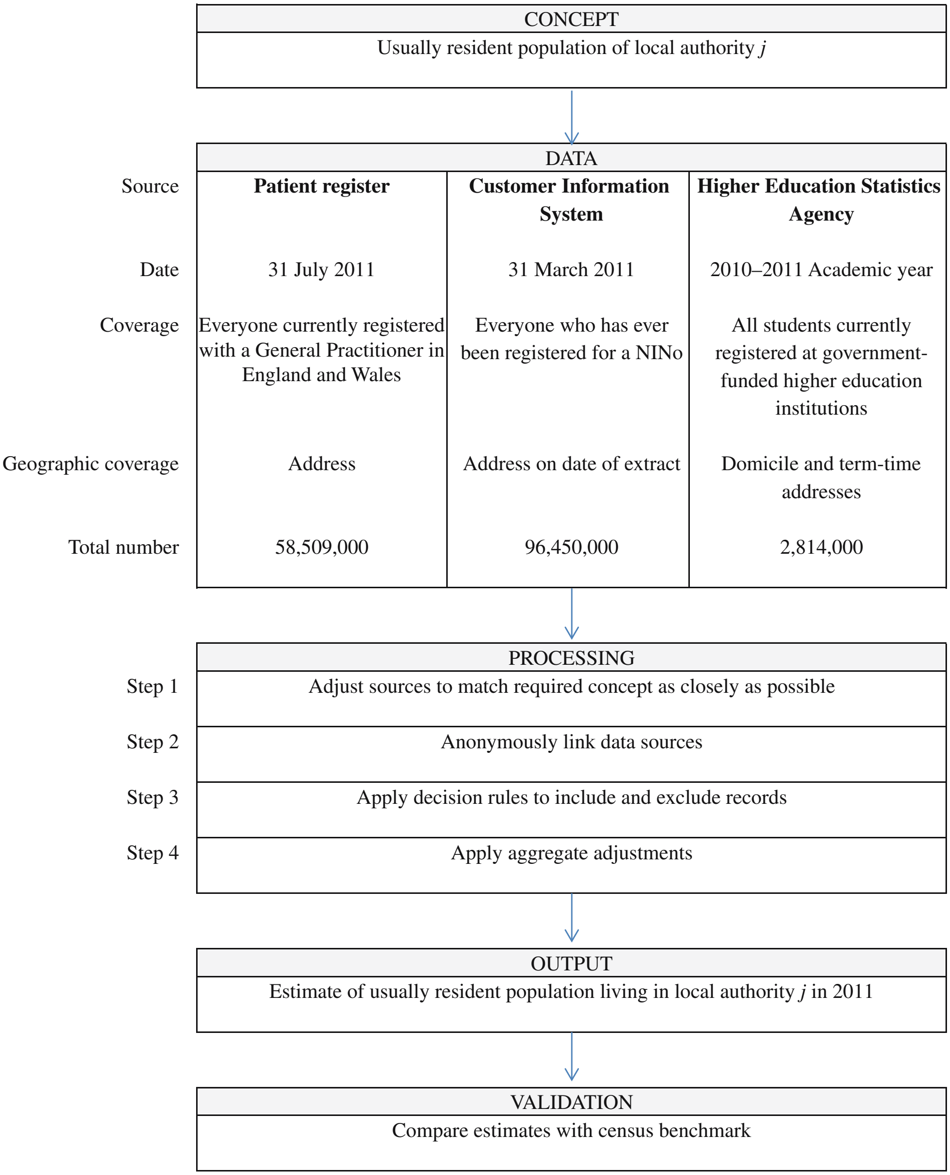

Usually resident population
DWP/HMRC Sources
Short term immigrants who have registered for a NINo and are in UK
Change of details lag for inmigrants to area *j*
Change of details lag for outmigrants from area *j*
Migrants who do not have a NINo
Duplicate record or NINo
Those who have moved/died abroad without notifying DWP
Children for whom child benefit has not been claimed
Usual residents with a NINo
Records for migrants who have applied and been issued with a NINo but have yet to enter the country

Fig. 9. Relating the CIS to the usual resident population in area j at time t

provide comprehensive information but only for subsections of the population. In combination, all these sources and others have the potential to provide comprehensive and timely information about population stocks and distributions in the UK. In order to produce these estimates, ONS needs to incorporate knowledge of the coverage and measurement error associated with each of the sources to process them individually, and to combine them in a way that captures the concept of usual residence. Work to achieve these combined estimates has begun (ONS 2013a). One illustration of this approach is summarised in Figure 10 below within the context of the population statistics framework. For each data set an attempt to remove those individuals not resident needs to be made. It is also necessary to adjust for those who may be recorded in the wrong place. The coverage, or representation error, within all the sources is the product of a complex web of issues that needs to be disentangled. For areas where populations are relatively stable the issues have a less significant impact than for those areas experiencing high levels of population turnover. Note that national statistics agencies, including ONS, are also interested in other populations, such as present population, short-term population and visitors. Which concept is used depends on the purpose to which the population statistics will be put.

ONS is undertaking the Beyond 2011 programme to explore options for producing population and sociodemographic statistics in future. Similar work is being undertaken by National Records for Scotland (NRS) and the Northern Ireland Statistics and Research Agency (NISRA). Some of the options involve using administrative registers to estimate the full population while large household surveys are used to assess coverage and to add the attribute detail delivered by the traditional census (ONS 2013f). These new methods have the potential to deliver population statistics more frequently, though the precision may be lower for small-area population statistics than for large-area population statistics.



CONCEPT

Usually resident population of local authority *j*

DATA

	Patient register	Customer Information System	Higher Education Statistics Agency
Source	**Patient register**	**Customer Information System**	**Higher Education Statistics Agency**
Date	31 July 2011	31 March 2011	2010–2011 Academic year
Coverage	Everyone currently registered with a General Practitioner in England and Wales	Everyone who has ever been registered for a NINo	All students currently registered at government-funded higher education institutions
Geographic coverage	Address	Address on date of extract	Domicile and term-time addresses
Total number	58,509,000	96,450,000	2,814,000

PROCESSING

Step 1 Adjust sources to match required concept as closely as possible

Step 2 Anonymously link data sources

Step 3 Apply decision rules to include and exclude records

Step 4 Apply aggregate adjustments

OUTPUT

Estimate of usually resident population living in local authority *j* in 2011

VALIDATION

Compare estimates with census benchmark

Fig. 10. Application of the population statistics framework to combine administrative data sources to measure the usual resident population in the UK (ONS 2012d, ONS2013a, ONS2013d, ONS 2013e)

5. Conclusion

The production of population statistics is complicated because populations are dynamic, heterogeneous, and influenced by the social and cultural environments, economic environments and natural and built environments in which the populations reside. In this article we have presented central frameworks for developing and improving population estimates. The key elements are user requirement, concepts, data, processing and outputs. User requirements determine the population statistics of interest. Concepts are required to understand and define the population statistics required and how the demographic components of change should be related to them. Concepts are also needed to match the measurements in the available data, which may come from censuses, surveys or administrative sources, to the statistic of need. As the available data are unlikely to match the concept exactly, and are very likely to contain error or miss certain groups of the population, processing is required in order to produce the final statistical output.

The frameworks for official statistics, population statistics, demographic accounting and survey error presented in this article are useful for facilitating communication with users of population and migration statistics, and for ensuring that everyone understands the underlying concepts, the nature of the data available and the methods used to derive estimates of key statistics. One important message that we have focused on is the tension between concepts and outputs. There are many requirements and types of population statistics but there are rarely single sources of information that cover all of these and match the concepts exactly. The published outputs, therefore, are only able to address some of the needs and are often produced by bringing together data from multiple sources, all with their relative strengths and weaknesses. Future work should consider extending the framework to include the relationships between different user needs and statistical releases.

In conclusion, we hope this article provides a better understanding of the relationships between the population concepts required by users and the data available to measure them. The frameworks presented in this article should also provide a guide for any country considering alternative data or methods for producing population and migration statistics. We focused on the UK and the work being carried out in the Office for National Statistics because of the transition that is occurring to incorporate and make better use of administrative data to estimate local populations. In particular, these frameworks have been used to guide the Beyond 2011 project that assessed administrative data quality, including coverage from a variety of sources, and have provided a single reference point for both users and providers.

6. References

Bell, M., M. Blake, P. Boyle, O. Duke-Williams, P. Rees, J. Stillwell, and G. Hugo. 2002. “Cross-National Comparison of Internal Migration: Issues and Measures.” *Journal of the Royal Statistical Society Series A* 165: 435–464. Doi: http://dx.doi.org/10.1111/1467-985X.t01-1-00247.

Boden, P. and P. Rees. 2009. “Using Administrative Data to Improve the Estimation of Immigration to Local Areas in England.” *Journal of the Royal Statistical Society Series A* 173: 707–731. Doi: http://dx.doi.org/10.1111/j.1467-985X.2009.00637.x.

Boyle, P. and D. Dorling. 2004. “The 2001 UK Census: Remarkable Resource or Bygone Legacy of the ‘Pencil and Paper Era’?” *Area* 36: 101–110. Doi: http://dx.doi.org/10.1111/j.0004-0894.2004.00207.x.

Bycroft, C. 2011. *Social and Population Statistics Architecture for New Zealand*. Wellington: Statistics New Zealand.

Groves, R., F. Fowler, M. Couper, J. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*, 2nd ed. New York: Wiley.

Kupiszewski, M. and D. Kupiszewska. 2008. "International Migration Component in Population Dynamics Models." In *International Migration in Europe: Data, Models and Estimates*, edited by J. Raymerand and F. Willekens, 309–327. Chichester: Wiley.

Laux, R. 2002. "Review of the Framework for Labour Market Statistics." *Labour Market Trends*: 485–492. Available at: http://www.ons.gov.uk/ons/rel/lms/labour-market-trends–discontinued-/volume-110–no–9/review-of-the-framework-for-labour-market-statistics.pdf

NHS 2011. *GP Choice: Choosing a GP*. Available at: http://www.nhs.uk/choiceintheNHS/Yourchoices/GPchoice/Pages/ChoosingaGP.aspx (accessed 7 August 2012).

ONS. 2011. *Improved Methodology for Estimating Immigration to Local Authorities (LAs) in England and Wales. Office for National Statistics*. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/imps/improvements-to-local-authority-immigration-estimates/index.html (accessed 8 October 2015).

ONS. 2012a. *Beyond 2011. Office for National Statistics*. Available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes—projects/beyond-2011/index.html (accessed 7 August 2012).

ONS. 2012b. *Uncertainty in Local Authority Mid year Population Estimates*. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/imps/latest-news/uncertainty-in-la-mypes/index.html (accessed 7 August 2012).

ONS. 2012c. *Estimating Internal Migration: Customer Guidance Notes*. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/internal-migration-methodology/estimating-internal-migration-customer-guidance-notes—november-2012.doc (accessed 23 August 2013).

ONS. 2012d. *Beyond 2011: Administrative Data Sources Report: NHS Patient Register*. Available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes—projects/beyond-2011/news/reports-and-publications/sources/beyond-2011–administrative-data-sources-report–nhs-patient-register.pdf (accessed 23 August 2013).

ONS. 2013a. *Beyond 2011: Producing Population Estimates Using Administrative Data: In Practice*. Available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes—projects/beyond-2011/news/reports-and-publications/beyond-2011-producing-population-estimates-m7.pdf (accessed 23 August 2013).

ONS. 2013b. *Beyond 2011: Reports and Publications*. Available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes—projects/beyond-2011/news/reports-and-publications/index.html (accessed 23 August 2013).

ONS. 2013c. *Census 2011: Coverage Assessment and Adjustment Methods*. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/methods/coverage-assessment-and-adjustment-methods/index.html (accessed 23 August 2013).

ONS. 2013d. *Beyond 2011: S5 Administrative Data Sources Report: Department for Work and Pensions (DWP) and Her Majesty's Revenue and Customs (HMRC) Benefit and Revenue Information (CIS) and Lifetime Labour Market Database (L2)*. Available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes—projects/beyond-2011/news/reports-and-publications/sources/beyond-2011-administrative-data-sources-report—dwp-and-hmrc-cis-and-l2-combined–s5-.pdf (accessed 23 August 2013).

ONS. 2013e. *Beyond 2011: Administrative Data Sources Report: Higher Education Statistics Agency: Student Record*. Available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes—projects/beyond-2011/news/reports-and-publications/sources/beyond-2011–administrative-data-sources-report–higher-education-statistics-agency-student-record–s4-.pdf (accessed 23 August 2013).

ONS. 2013f. *Beyond 2011: Options Explained 2*. Available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes—projects/beyond-2011/news/reports-and-publications/methods-and-policies/beyond-2011-options-explained-2.pdf (accessed 23 August 2013).

ONS. 2013g. *1.02 IPS Margins of Error, 1975–2012. Table 1.02 Long-Term International Migration, Estimates from International Passenger Survey, Annual Data 2010, Contacts, Estimates, 95% Confidence Intervals and Ranges for the 95% Confidence Intervals for Actual Length of Stay*. Available at: http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-346438 (accessed 8 October 2015).

Poulain, M., N. Perrin, and A. Singleton. 2006. *THESIM: Towards Harmonised European Statistics on International Migration*. Louvaine: UCL Presses Universitaires.

Raymer, J. and P.W.F. Smith. 2010. "Editorial: Modelling Migration Flows." *Journal of the Royal Statistical Society, Series A* 173: 703–705. Doi: http://dx.doi.org/10.1111/j.1467-985X.2010.00660.x.

Raymer, J., P.W.F. Smith, and C. Giulietti. 2011. "Combining Census and Registration Data to Analyse Ethnic Migration Patterns in England from 1991 to 2007." *Population, Space and Place* 17: 73–88. Doi: http://dx.doi.org/10.1002/psp. 565.

Rees, P. 1985. "Does It Really Matter Which Migration Data You Use in a Population Model?" In *Contemporary Migration Studies*, edited by P. White and G. van der Knaap, 55–77. Norwich: GeoBooks.

Rogers, A. 1990. "Requiem for the Net Migrant." *Geographical Analysis* 22: 283–300. Doi: http://dx.doi.org/10.1111/j.1538-4632.1990.tb00212.x.

Stone, R. 1971. "An Integrated System of Demographic, Manpower and Social Statistics and its Links with the System of National Economic Accounts." *Sankhyā: The Indian Journal of Statistics* 33: 1–184.

Statistics New Zealand. 2011. *Population Domain Plan 2012. Draft for Public Consultation*. Wellington: Statistics New Zealand. Available at: http://www.stats.govt.nz/~/media/Statistics/browse-categories/population/population-domain-plan/pop-dom-plan-draft-2012.pdf (accessed 8 October 2015).

United Nations. 1975. *Towards a System of Social and Demographic Statistics*. Department of Economic and Social Affairs, Studies in Methods, Series F, No. 18. Publication ST/ESA/STAT/SER.F/18. New York: United Nations. Available at: http://unstats.un.org/unsd/publication/SeriesF/SeriesF_18E.pdf (accessed 8 October 2015).

United Nations. 2007. *Principles and Recommendations for Population and Housing Censuses, Revision 2*, Statistical papers Series M. No. 67/Rev.2, Statistics Division, Department of Economic and Social Affairs, United Nations. New York: United Nations. Available at: http://unstats.un.org/unsd/demographic/sources/census/docs/P&R_%20Rev2.pdf (accessed 8 October 2015).

Zhang, L.-C. 2011. "Topics of Statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Received August 2012
Revised March 2015
Accepted March 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 737–761, http://dx.doi.org/10.1515/JOS-2015-0043

Quality Indicators for Statistical Disclosure Methods: A Case Study on the Structure of Earnings Survey

Matthias Templ[1]

Scientific- or public-use files are typically produced by applying anonymisation methods to the original data. Anonymised data should have both low disclosure risk and high data utility.

Data utility is often measured by comparing well-known estimates from original data and anonymised data, such as comparing their means, covariances or eigenvalues.

However, it is a fact that not every estimate can be preserved. Therefore the aim is to preserve the most important estimates, that is, instead of calculating generally defined utility measures, evaluation on context/data dependent indicators is proposed.

In this article we define such indicators and utility measures for the Structure of Earnings Survey (SES) microdata and proper guidelines for selecting indicators and models, and for evaluating the resulting estimates are given. For this purpose, hundreds of publications in journals and from national statistical agencies were reviewed to gain insight into how the SES data are used for research and which indicators are relevant for policy making.

Besides the mathematical description of the indicators and a brief description of the most common models applied to SES, four different anonymisation procedures are applied and the resulting indicators and models are compared to those obtained from the unmodified data. The disclosure risk is reported and the data utility is evaluated for each of the anonymised data sets based on the most important indicators and a model which is often used in practice.

Key words: Statistical disclosure control; data utility; quality indicators; R.

1. Introduction

Anonymisation methods are applied to microdata to reduce their disclosure risk. By applying too much or overly heavy anonymisation, the data utility is reduced and the information loss is increased. However, users who analyse anonymised microdata want to have as precise parameter estimates as possible. It is therefore of great interest to measure the data and user context utility of a microdata set after disclosure limitation methods have been applied.

1.1. *General Methods for Measuring Data Utility*

Anonymised data should have the same structure as the original data and should allow for analysis with high precision.

[1] Statistics Austria, Dept. of Methodology, Guglgasse 13, Vienna 1110, Austria. Email: matthias.templ@gmail.com

Acknowledgment: I wish to thank three anonymous referees for their constructive reviews. My special thanks goes to the corresponding associate editor who improved the quality of the article.

To evaluate the precision, the estimation of different classical estimates such as means and covariances are often focused upon. By using the R-package **sdcMicro** (Templ and Meindl 2010; Templ 2008; Templ et al. 2015), it is possible to calculate 26 different measures on continuous scaled variables that are based on classical (most of these measures are described in Hundepool et al. 2012) or robust distances. These measures are computed for the original data and the perturbed data and then compared. To evaluate the multivariate structure of perturbed data, comparisons based on eigenvalues and robust eigenvalues may also be made. The comparison of means and covariances by mean squared errors, mean absolute errors, and mean variations is also proposed in Domingo-Ferrer et al. (2001). A generalisation is given by Domingo-Ferrer and Torra (2001) by averaging the mean variations and mean absolute errors. They also define information loss measures for categorical variables: direct comparison of categorical values, comparison of contingency tables, and entropy-based measures. For the direct comparison, a distance is defined over the range of categories. When the range of categories of a variable is of ordinal scale, the distance between two categories is proportional to the number of categories between them. For nominal scale, the Hamming distance (zero when equal, otherwise one) is chosen. The comparison of contingency tables considers the number of differences between the two contingency tables, normalised by dividing by the number of cells of a table. The entropy-based measure is suitable for the PRAM method, where the logarithm of the transition probabilities of one category to another is used.

Shlomo (2008) uses some methods to evaluate data utility based on a contribution from Gomatam and Karr (2003) and extends them by measures on impact of association and a measure based on the between variance of a proportion fitted by regression models.

Woo (2009) proposes the use of propensity-score methods. The idea is to merge or join the original and the perturbed data sets and then create a new index variable with ones for the original data and zeros for observations from anonymised data. A logistic regression model is then fitted using the new index variable as the response variable. Predictions from this model are then compared with the proportion of observations of the perturbed data to the original data (usually 1/2). Woo also describes two other measures, one based on cluster analysis (evaluating the cluster sizes) and another which compares the empirical cumulative distribution function. They concentrate only on data utility measures and do not account for disclosure risk. Karr et al. (2006) propose measures based on differences between inferences on original and perturbed data that are tailored to normally distributed data, and they also use the propensity score method in Oganian and Karr (2006).

Reiter (2012) mentions, without presenting numerical results, that the comparison of measures based on specific models is often done informally. If the regression coefficients obtained from original and perturbed data are considered close, for example if the confidence intervals obtained from the models largely overlap, the released data have high utility for that particular analysis (see also Karr et al. 2006).

1.2. Trade-Off Between Data Utility and Disclosure Risk

The goal of statistical disclosure control is always to release a safe microdata set with high data utility and a low risk of linking confidential information to individual respondents.

Disclosure risk can be measured in different ways. Several methods have been suggested, such as the individual risk approach (Franconi and Polettini 2004) that is used in this contribution, methods based on log-linear models (Rinott 1990; Carlson 2002) or the SUDA concept (Manning et al. 2008). So, firstly, a decision on which method, or methods, for measuring disclosure risk will be used is necessary. Secondly, the data holder has to decide on the level of disclosure risk that is acceptable and sufficient for distributing the data. For example, in the case of the SES, anonymised microdata is sent to Eurostat. However, many countries do not agree with the proposed rules for anonymisation communicated by Eurostat, nor can they allow the use of remote access systems such as the PiEP Lissy project (Marsden 2010) because of restrictions in national legislation. Therefore, almost every country applies different anonymisation methods to their data (the anonymisations and the disclosure risk are therefore somehow fixed in advance), but Eurostat wants to ensure that the most important statistics can be estimated with high precision.

In this study, the focus is not specifically on disclosure risk, however, and hence only one disclosure risk measure, the individual risk approach, was used. Several anonymisation procedures were however applied to the data and the data utility for each case is reported. It is up to the data holders to decide whether a particular anonymisation procedure is sufficient. In this study we have simply assumed that the chosen anonymisations are sufficient from a risk point of view and devote our attention to data utility.

1.3. Outline of the Article

In Section 2 we describe the basic ideas of the proposed approach for utility assessment. Section 3 introduces the Structural Earnings Survey (SES). In addition, the usage of this particular survey is analysed and the most important projects which have their main focus on this data set are mentioned. Based on this analysis, the most important indicators are discussed in Section 4 and the three most important indicators and one model are described in detail in Section 5. Confidentiality aspects are briefly discussed in Section 6. Results from the analysis using the selected data utility measures are presented in Section 7. Section 8 concludes the article.

2. Data and Context-Driven Utility Measures

In practice it is not possible to create an anonymised file that has exactly the same structure as the original file. Contrary to general methods described previously, we propose that the differences between estimates based on anonymised and original data need to be small, or even zero, only for the most important statistics. This approach measures the data utility based on quality indicators (Ichim and Franconi 2010; Franconi et al. 2011; Templ 2011a) and is another more user-driven approach than applying general tools, since for the users it might not be relevant to estimate all popular statistics with high precision, but just those that are relevant for their analysis.

The first step in quality assessment is to decide on a set of quality indicators. To do so, one has to evaluate the user needs, that is, what is analysed by the users, and report on the most important estimates. These estimators are often named *benchmarking indicators* (see, e.g., Templ 2011a,b) and referred to here as quality indicators.

The general procedure is quite simple – although much work is necessary. It can be described in the following steps:

i) Analysis of the user needs of researchers, policy makers, and society regarding a specific data set. Analysis of the aim for which the underlying data have been used.
ii) Selection of a set of quality indicators after the detailed analysis in (i).
iii) Estimation of all quality indicators on the original, unmodified microdata set.
iv) Estimation of the quality indicators on the protected microdata set.
v) Comparison of statistical properties such as point estimates, variances or overlaps in confidence intervals for each quality indicator.
vi) Assessment of the data utility of the protected microdata set.

If the quality of the data is reasonable, the anonymised microdata set may be published. Note that the anonymisation procedure chosen has to lead to a reasonably low disclosure risk of the anonymised data.

If the deviations of the main indicators calculated from the original and the protected data are too large, the anonymisation procedure should be revised by modifying selected parameters used for the applied disclosure methods or by a complete revision of the anonymisation process.

Usually the evaluation is focused on the properties of numeric variables given unmodified and modified microdata. However, it is of course also possible to look at the impact of local suppression or recoding that has been conducted to reduce individual reidentification risks.

Another possibility to evaluate the data utility is to define a model that is fitted to both, the original, unmodified microdata and the anonymised data. The main idea is to look at differences in the regression coefficients. If the deviations are small enough, one may go on to publish the safe and protected microdata set. Otherwise adjustments in the protection procedure need to be carried out.

It may also be of interest to evaluate the set of quality indicators not only for the entire data set but also for some domains. The evaluation of quality indicators is then performed for each of the h groups by looking at differences between indicators for original and modified data in each group.

3. The Structural Earnings Statistics Survey

The Structural Earnings Statistics Survey (SES) is conducted in almost all European countries, and the most important figures are reported to Eurostat.

3.1. Sampling Design, Data Preparation Issues, and Data Sources

SES is a complex survey of enterprises and establishments with more than ten employees (e.g., 11,600 enterprises in Austria), NACE C-O, including a large sample of employees (e.g., in Austria: 199,909). In many countries, a two-stage design is used where in the first stage a stratified sample of enterprises and establishments on the NACE one-digit level, NUTS 1 and employment size range is used – large companies have higher inclusion probabilities. In stage two, systematic sampling is applied within each enterprise using unequal inclusion probabilities with regard to employment size-range categories.

In the Austrian case, for example, the sample has only 2.4% nonresponse. Regression imputation is applied by using tax data to replace these missing values. If information on imputed values is available, variance estimation procedures should account for this extra variability.

Calibration is applied to reflect certain population characteristics corresponding to NUTS 2 and NACE one-digit level, but also for gender (number of men and women in the population).

SES compromises information from different perspectives and sources:

Information on the enterprise level: Enterprises are asked question batteries, such as whether the enterprise is private or public or whether it has a collective bargaining agreement (both binary variables). As a multinomial variable, the type of collective agreement is included in the questionnaire.

Information on the individual employment level: The following questions to employees come with the standard questionnaire: social identity number, date of employment, weekly work hours, kind of work agreement, occupation, amount of annual leave, place of work, gross earnings, earnings for overtime, and amount of overtime.

Information from registers: All other information may come from registers, such as information about age, size of enterprise, occupation, education, amount of employees, NACE, and NUTS classifications.

3.2. Standard Publications and Use of the Microdata

The standard publication from national statistical offices is issued every four years after the survey is conducted. In addition, a special publication about low incomes and non-common occupation employment is published by some member states, such as Statistics Austria's report on low incomes (see Geissberger and Knittler 2010). In Austria, a special report has been written for the Austrian women's report focused on the gender pay gap and socioeconomic studies (Geissberger 2010). Many other national publications by statistical agencies or researchers are available in almost every country (for some summaries about publications until 1999, see Belfield 1999; Nolan and Russel 2001; Dupray et al. 1999; Frick and Winkelmann 1999; Dell'Aringa et al. 2000).

However, social scientists have mostly carried out qualitative analysis or rough quantitative interpretations of a few official figures, mainly because of lack of access to micro data for researchers. One exception are publications made with direct or follow-up data connection and using the PiEP Lissy project and its remote access system (Marsden 2010) to various SES data. Actually, 10-15 projects are running within Eurostat's Safe Center and anonymised CD-ROM (see the next section).

3.3. Access to SES Microdata and European Projects

Access to Data Provided by Eurostat: Anonymised SES 2002 and 2006 data from 23 countries can be accessed for research purposes by means of research contracts through the safe center or anonymised CD-ROM at the premises of Eurostat. The output will be checked by Eurostat for confidentiality and quality. Further plans include automatic output checking of data to reduce the workload of the statistical institutes. More technical details on the safe center can be found in Reuter (2010); and Reuter and Museux (2011). To obtain the data, see Eurostat's website: http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/ses.

Access to Data Through PiEP Lissy: The *Pay Inequalities and Economic Performance Project* (PiEP) studied wage differentials based on SES data (Marsden 2010) in depth. SES microdata from the Czech Republic, Hungary, Ireland, Italy, Latvia, Lithuania, the Netherlands, Norway, Portugal, Slovakia, and Spain can also be analysed via the PiEP Lissy remote-access system. The user can run Stata code on the PiEP Lissy server, for example, although some commands (twelve in total) are blocked by the system to prevent listing of individuals.

Synthetic SES Population Data: A synthetic population is simulated in Templ and Filzmoser (2014) and a sample of this population is included in the R-package laeken (Alfons and Templ 2013).

The LEED Project: Within the EU project on *Linked Employer-Employee Data* (LEED), studies assessing the potential of linked employer-employee and panel data sets for the analysis of European labour-market policy are carried out. They concentrate on SES data and use the PiEP Lissy remote access system to gain access to the data of twelve different countries, see http://cep.lse.ac.uk/leed/.

The Dynamic Wage Network: The dynamic wage network was founded by the European Central Bank and it consists of four research groups. The microdata group pursues three directions of research one of which is on wage differentials and modelling of earnings. The SES data is one of the main data sources for this group, used by many authors (see, e.g., Caju et al. 2010, 2009a,b; Messina et al. 2010; Dybczak and Galuscak 2010; Simón 2010; Pointner and Stiglbauer 2010).

4. Important Indicators Estimated from SES Data

4.1. Research Potential of SES Microdata

Statistical agencies usually provides, amongst other things, tables on average hourly earnings on domain level (Geissberger 2009), for country comparisons (see, e.g., Mittag 2005) and for special groups like low incomes (Geissberger and Knittler 2010; Casali and Alvarez 2010).

SES data includes information on enterprise and employment level. Generally such linked employer-employee data are used to identify determinants/differentials of earnings, some indicators are also directly derived from hourly earnings, such as the gender pay gap or the Gini index (Gini 1912). The most classical example is the income inequality between genders as discussed in for example, Groshen (1991).

A correct identification of factors influencing earnings could lead to relevant evidence-based policy decisions. Research studies are usually focused on examining the determinants of disparities in earnings. Earnings comparisons between different industries or regions are frequently performed (see, e.g., Stephan and Gerlach 2005; Caju et al. 2010, 2009b,a; Messina et al. 2010; Dybczak and Galuscak 2010; Simón 2010; Pointner and Stiglbauer 2010). Sometimes socioeducational factors are investigated as possible explanatory variables of income, for example in Bowles et al. (2001). The overview of the analyses performed using SES data highlighted that, generally, the log hourly earnings are modelled. The explanatory variables correspond to employer activity (related to the enterprise), his or her experience (education, length of stay in service, qualification, etc.) and working hours. It was also observed that linear models are extensively used. ANOVA analysis, linear mixed-effects models, and multi-level models are other examples of statistical tools that have been applied. However, a lot of similar models are applied in the literature to model the log hourly earnings.

It should also be noted that the distribution of errors is always assumed to be normal. The estimates are generally computed by means of ordinary least squares by ignoring the sampling design and corresponding weights which is not good practice.

4.2. Summary of the Most Important Analyses from SES Data

In summary, the most important analyses using SES data are related to

Gender pay/wage gap: The gender wage gap is currently one of the most important indicators obtained from SES in many European countries (Research Center for Education and the Labour Market at the Maastricht University 2009) and intensively discussed in the European Union (Dupré 2010). In Austria, for example, many publications about the gender wage gap are published by Statistics Austria and the national authorities (Stockinger 2010). The topic *Women and Equality* is of central interest not only for the Federal Minister for Women and the Civil Service, and socioeconomic studies are carried out with support from the state (one example is Geissberger 2010) or European institutions where regression models are also applied to estimate the adjusted gender pay gap (Research Center for Education and the Labour Market at the Maastricht University 2009).

Wage differentials and interindustry wage differentials: Differences in earnings for workers employed in different industries and occupations has long been recognised as an important issue for the labour market and several studies have been carried out (Caju et al. 2010, 2009a,b; Messina et al. 2010; Dybczak and Galuscak 2010; Simón 2010; Pointner and Stiglbauer 2010). Pointner and Stiglbauer (2010) use several workplace-specific dummy variables for the employee's occupation (ISCO 1) within the firm, the sector (NACE-2 digits) of the employer, for firm size and location (NUTS-1 digits), and a control for private ownership of the firm as predictors. Caju et al. (2010, 2009b) modelled the gross hourly wages with sex, education, age class, number of years of employment, type of employment contract, part/full-time, bonus for shift work, night and/or weekend work, a dummy for paid overtime and occupation sector effect. Messina et al. (2010) used a model to predict the log hourly wages with firm size, firm size squared, age class, female employment proportion and proportion of high- and low-skilled workers as predictors. Caju et al. (2009a) used age, capital-labour ratio, profit elasticity and the percentage of blue-collar workers covered by single-employer collective agreements to model the log hourly earnings.

Low-pay dynamics: In some countries, great changes in the distribution of earnings are observed (see, e.g., Dell'Aringa et al. 2000; Geissberger 2009) with a widening of inequality and an increase in dispersion. The Gini index and the quintile share ratio are two of the main indicators to estimate the inequality (Graf et al. 2011; Kolb et al. 2011).

Enterprise characteristics that affect earnings or profit: The differential that describes the profit of an enterprise is an interesting aspect, that is how enterprises integrate a combination of systems to provide greater flexibility in pay, and how information sharing and the size of the enterprise influences the profitability of an enterprise. On the other hand, it is of interest to investigate the prediction of pay flexibility using the size of the enterprise, level of competition, training, job rotation, time flexibility, and so on (see, e.g., Marsden 2010).

Collective bargaining: Due to the unions importance in determining wages, to measure the extent of the union-nonunion wage gap is of interest (for an example from the US, see Edwards 2010; also see Fitzenberger et al. 2006).

Average Earnings: Average earnings in enterprises as an indicator for productivity or performance (Winter-Ebmer and Zweimüller 1999; Marsden 2010). The idea is that in a competitive market environment, employees' pay corresponds to the value of their output, that is deviations from this position would lead to difficulties in recruitment and retention. In branches with high output, earnings would therefore be higher compared to enterprises in low economic branches with low production.

Occupation and length of employment: Another interesting analysis includes the difference in income for different occupation levels or by the length of employment.

Comparative studies between countries play an increasingly important role. However, our purpose is to study how estimates of a defined set of indicators from protected microdata perform compared to estimates based on the original, unmodified data. Therefore, such comparative studies are not directly within the scope of this work, since good estimates on a country level should ensure that comparisons between countries are possible.

5. Two Indicators and One Model for Quality

In the following, three measures that we have identified as the most important and have selected as quality indicators are described in full detail. Note that in a real-life setting, one would include any number of measures deemed important enough and not just the three we have chosen. However, in order to avoid this article becoming overly long, we limit the investigation to only these three quality statistics.

First, the (unadjusted) gender pay gap (GPG) is described, since it is one of the most important indicators obtained from SES data; thereafter the Gini index is described. The GPG and the Gini index (for hourly earnings) are extremely sensitive to changes in the upper and lower tail of the distribution (see e.g., Alfons et al. 2013). If these estimators are not affected by anonymisation, one can be quite sure that the corresponding variables have high data utility, since it is most difficult to preserve the structure of the data in the upper tail of the distribution.

Lastly, a model-based estimation on employment level is described, representative for all model-based estimations. Note that our choice of indicators and model is subjective; even so, the choice is based on our review of dozens of contributions (see Subsection 4.1). However, it can be expected that differences in estimations between anonymised and original data according to this model will be comparable in similar models.

5.1. The Gender Pay Gap

As already noted, the GPG is probably the most important indicator derived from the SES data.

The calculation of the GPG is based on each person's hourly earnings. The hourly earnings equals to the gross monthly earnings from labour divided by the number of hours usually worked per week during 4,33 weeks, (see EU-SILC 2009; Beblot et al. 2003).

5.1.1. Definition Gender Pay Gap

The GPG in unadjusted form is defined on population level as the difference between average gross earnings of male paid employees and of female paid employees divided by the earnings of male paid employees (EU-SILC 2009).

5.1.2. Estimation of the Gender Pay Gap

Since the GPG is usually estimated by survey information, the estimation has to consider sampling weights in order to ensure sample representativity. Therefore, all our estimations consider sampling weights.

We let $\boldsymbol{x}: = (x_1, \ldots, x_n)'$ denote the hourly earnings where $x_1 \leq \ldots \leq x_n$ and $\boldsymbol{w}: = (w_i, \ldots, w_n)'$ denote the corresponding personal sample weights, where n denotes the number of observations.

We define the index set

$$J^{(M)} := \{j \in \{1, \ldots, n\} \mid \text{worked as least 1 hour per week } \wedge (16 \leq \text{age} \leq 65)$$

$$\wedge \text{ person is male}\},$$

and let $J^{(F)}$ be the corresponding index set for female employees.

With these index sets, the GPG in its unadjusted form is estimated by

$$GPG_{(mean)} = \frac{\frac{\sum_{i \in J^{(M)}} w_i x_i}{\sum_{i \in J^{(M)}} w_i} - \frac{\sum_{i \in J^{(F)}} w_i x_i}{\sum_{i \in J^{(F)}} w_i}}{\frac{\sum_{i \in J^{(M)}} w_i x_i}{\sum_{i \in J^{(M)}} w_i}}. \tag{1}$$

The definition from EU-SILC (2009) differs from the definition used by the Bureau of Labour Statistics of the United States (see, e.g., Weinberg 2007), where weighted medians are used instead of arithmetic means.

The GPG is usually estimated at domain level such as economic branch, education and age groups (Geissberger 2009).

In addition, it is important to estimate the variances of the estimations.

5.2. The Gini Index for the Estimation of Inequality

The Gini index (Gini 1912) is a well-known measures of inequality of a distribution and is widely applied in many fields of research.

The Gini index according to EU-SILC (2004, 2009) is estimated by

$$\widehat{Gini} := 100\left[\frac{2\sum_{i=1}^{n}\left(w_i x_i \sum_{j=1}^{i} w_j\right) - \sum_{i=1}^{n} w_i^2 x_i}{\left(\sum_{i=1}^{n} w_i\right)\sum_{i=1}^{n}(w_i x_i)} - 1\right]. \tag{2}$$

The Gini index is closely related to the Lorenz curve (Lorenz 1905), which plots the cumulative proportion of the total income against the corresponding proportion of the population.

The Gini index and the GPG are typically – among other domains – estimated with breakdowns by age and gender, or age, gender, and region, or by education level. The latter domain is used in the following.

5.3. *Model-Based Predictions on Employment Level*

As representative of all model-based estimations at employment level, we choose a model described in Marsden (2010) applied within the PiEP Lissy project and also used in Dybczak and Galuscak (2010). They fit OLS regression models where they modelled the gross hourly earnings of workers in enterprises using age, age^2, sex, education, and occupation as predictors.

The data from the Lissy system is also used for the LEED project (see Subsection 3.3) where similar studies and modelling have been carried out (see, e.g., Simón 2010). Similar models are also fitted within the *wage dynamics network* of the European Central Bank (Caju et al. 2010; Pointner and Stiglbauer 2010).

In the following estimations, the following model is used:

$$\begin{aligned} log(\text{hourly earnings}) \sim\ & \text{sex}\,(2) + \text{age}\,(6) + \text{education}\,(6) + \text{occupation}\,(23) \\ & + \text{location}\,(5) + \text{economic activity}\,(12) + \text{error term} \end{aligned}$$

The numbers in brackets correspond to the respective number of categories for each of the categorical variables in the original SES data.

It seems that the sampling weights are mostly ignored in the literature on fitting models to SES data. However, in our study the weights are taken into account by using weighted least squares regression.

5.4. *Variance Estimation*

A calibrated bootstrap to estimate the variances (Bruch et al. 2011; Templ and Alfons 2011) for the GPG and the Gini index is applied.

Let X denote a survey sample with n observations and p variables. Then the *calibrated bootstrap algorithm* for estimating the variance and confidence interval of an indicator can be summarised as follows:

1. Draw R independent bootstrap samples $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_R^*$ from $\boldsymbol{X}$.

2. Calibrate the sample weights for each bootstrap sample $\boldsymbol{X}^*_r$, $r = 1, \ldots, R$. Generalised procedures are then used for calibration: a multiplicative method known as *raking*, an additive method or a logit method (see Deville and Särndal 1992; Deville et al. 1993).
3. Compute the bootstrap replicate estimates $\hat{\theta}^*_r := \hat{\theta}(X^*_r)$ for each bootstrap sample X^*_r, $r = 1, \ldots, R$, where $\hat{\theta}$ denotes an estimator for a certain indicator of interest. The sample weights need to be considered in the computation of the bootstrap replicate estimates.
4. Estimate the variance $V(\hat{\theta})$ by the variance of the R bootstrap replicate estimates:

$$\hat{V}(\hat{\theta}) := \frac{1}{R-1} \sum_{r=1}^{R} \left(\hat{\theta}^*_r - \frac{1}{R} \sum_{s=1}^{R} \hat{\theta}^*_s \right)^2 \tag{3}$$

5. Estimate the confidence interval at confidence level $1 - \alpha$ by the percentile method: $\left[\hat{\theta}^*_{((R+1)\frac{\alpha}{2})}, \hat{\theta}^*_{((R+1)(1-\frac{\alpha}{2}))}\right]$, as suggested by Efron and Tibshirani (1993), where $\hat{\theta}^*_{(1)} \leq \ldots \leq \hat{\theta}^*_{(R)}$ denote the order statistics of the bootstrap replicate estimates.

6. Confidentiality Issues and Perturbation of SES

6.1. *Disclosure Scenario*

In principle, two reidentification scenarios are related to the SES data. The identification of an enterprise may lead to information about their employees. Key variables at enterprise level might be *location* (3), NACE one-digit level codes (*economic activity*) (12), *size* of the enterprise (5), and distinction between *public or private* enterprises (2); the bracketed numbers are the respective number of categories. However, here we only focus on reidentification scenarios on employment level since the fraction of employees asked in each company, is rather high (lower for large enterprises, larger to all employees in smaller companies). Furthermore, to limit the scope of the paper, more serious disclosure situations on employment level will not be considered.

Categorical key variables at employment level might be *location* (3), *age class* (6), *education* (7), *economic activity* (12), and *size* (5). This leads to 7,560 strata. Of course, the choice of key variables for disclosure scenarios is a somewhat subjective decision and might vary across countries. For example, Ichim and Franconi (2007) proposed to use only *location*, *economic activity*, *size* and *age class* as categorical key variables. Continuous key variables at employment level might be the *hourly earnings* and *overtime earnings*. This choice of scenario is also a subjective decision.

Remark: Anonymised SES 2002 and 2006 data from 23 countries can be accessed for research purposes through the safe center at the premises of Eurostat. Anonymisation is done by recoding NACE, NUTS, and size, removing citizenship and building six age classes, microaggregation (individual ranking) for absence days and earnings and removing the sampling weights.

6.2. *Anonymisation of SES*

Various methods exists to anonymise microdata (see, e.g., Hundepool et al. 2012; Templ and Meindl 2010). Two possibilities (amongst others) for anonymisation are the following:

a) To provide *k*-anonymity (Sweeney 2002) for categorical key variables (for enterprises, for employees), and to apply microaggregation or adding (correlated) noise (Brand 2004) for continuous key variables.
b) Synthetic data generation of all variables (Alfons et al. 2011; Templ and Filzmoser 2014), that is, simulation of all variables by drawing from predictive distributions. Note that by simulating only a part of variables (e.g., gross earnings) and leaving other variables (such as the categorical variables) unchanged, intruders might be able to identify persons based on the unchanged variables and this might not be in scope with specific legislations on data privacy.

Fixed rules to protect the microdata may not always be accepted by all data providers (e.g., member states of EU); some freedom to choose protection methods must be given. However, some minimal quality requirements must be fulfilled by the applied protection methods (Ichim and Franconi 2010).

We do not go into detail about the anonymisation methods *per se* since the main focus of this paper is on evaluating the data utility of anonymised data.

Nevertheless, three possible perturbations to make the data confidential are outlined and applied. First, variables *size*, *age*, *sex*, *location*, *education* and *economic activity* are selected as categorical key variables and *hourly earnings* and *overtime earnings* as continuous key variables. Then the following anonymisation procedures are applied (note that this choice of anonymisation methods is subjective and many other disclosure scenarios and perturbation methods can be applied):

1. Recoding from 53 categories to twelve categories for the variable *economic activity*: local suppression to achieve three-anonymity (optimal local suppression following Templ et al. 2015); microaggregation (individual ranking method for fast computations) applied on each strata defined by *economic activity* of *hourly earnings* and *overtime earnings* with aggregation level 4.
2. Same recoding and local suppression as in 1: adding correlated noise (Brand 2004) to *hourly earnings* and *overtime earnings* with noise parameter 150 (for details, see Templ et al. 2015).
3. Swapping *location* and *economic activity* using the (invariant) postrandomization method (PRAM, see Gouweleeuw et al. 1998) with default parameters (see Templ et al. 2013); microaggregation as in 1.
4. Experimentally, shuffling (Muralidhar and Sarathy 2006) with a rather small model is applied (earnings hour + earnings overtime ~ sex + age + education); the anonymisation of categorical key variables are done as in 1 (and 2).

The amount of local suppression (to achieve three-anonymity) for Procedures 1, 2 and 4 is 0.001% (one value out of 199,909) for *size*, 0.115% (230 values) for *economic activity* and 0.005% (nine values) for *age*.

Table 1. Comparison of different anonymisation methods using the absolute relative bias (arb) for the GPG and the Gini indices. Overall indicates the estimation of bias without taking domains into account; for GPG, arb over domains education and age is calculated; for the Gini index, arb over domain age × sex is calculated. Global disclosure risk and the number of expected reidentifications are reported in the last two columns, specifically the amount of observations violating three-anonymity (three-anon), the sum of individual risks (grisk) and the expected numbers of observations disclosed (for the latter two, see Franconi and Polettini 2004). Asterisks (for PRAM) indicate that the estimates might not be reasonable

	Gender pay gap			Gini index		Global risk		
	Overall	Education	Age	Overall	Age-sex	3-anon	Grisk	Ident.
original	–	–	–	–	–	4414	0.010	2024
rec + ls + ma	0.176	0.671	0.861	0.081	0.191	0	0.002	426
rec + ls + noise	0.669	0.524	2.478	0.646	1.484	0	0.002	426
pram + ma	0.010	0.185	0.314	0.001	0.100	1011*	0.003*	539*
rec + ls + shuffle	11.918	18.677	152.381	0.009	18.856	0	0.002	426

For the application of PRAM in Procedure 3, 18,151 values changed their category in *location* and 18,867 values in *economic activity*.

7. Results

The utility measures chosen – based on the quality indicators that have been defined in Section 5 – are the following:

- The difference in the estimation of the GPG and the Gini from the original and perturbed data defined for h domains given by the (well-known) absolute relative bias:

$$arb = \frac{1}{h}\sum_{i=1}^{h}\frac{\left|\hat{\theta}_i - \tilde{\theta}_i\right|}{\hat{\theta}_i}, \tag{4}$$

- where $\hat{\theta}$ and $\tilde{\theta}$ denote the estimates from the original and the anonymised data set respectively. Note that the $\hat{\theta}$ have to be nonzero, which is practically always the case.
- The variances are estimated and the overlap of the confidence interval of the perturbed and original data is evaluated and reported as percentages.
- The model defined in Subsection 5.3 is fitted using weighted least squares regression on original and perturbed data. To stay comparable, the categories of *economic activitiy* are equal, that is, the NACE one-digit level is chosen.

7.1. *Absolute Relative Bias*

Table 1 shows the absolute relative bias (arb) for the GPG and the Gini index; both the overall estimate and the mean over the domains is shown. Here, the domain *education* and *age* is chosen for the GPG and for the Gini index, the domain (sex × age class) is used since this is reported to be one of the most interesting domains (see, e.g., Geissberger 2009, EU-SILC 2009 and Section 5).

The global measure of individual risk and the expected number of reidentifications are reported in the last two columns of Table 1. Note that the sum over individual risks gives the number of expected reidentifications. The number of reidentifications is not high in the original data set (2,024 of 199,909 observations) and it is reduced by applying the

Table 2. Lower (l) and upper (u) limits of the confidence intervals for the GPG for each category of education

Data	ISCED 0–1	ISCED 2	ISCED 3–4	ISCED 5A	ISCED 5B
original (l)	0.15938	0.12102	0.22572	0.29568	0.21744
original (u)	0.26525	0.15023	0.23944	0.35010	0.25835
rec + ls + ma (l)	0.16123	0.12144	0.22624	0.28891	0.21290
rec + ls + ma (u)	0.27062	0.15211	0.23970	0.34381	0.25904
rec + ls + noise (l)	0.17012	0.12106	0.22399	0.29135	0.21152
rec + ls + noise (u)	0.27011	0.15172	0.23776	0.34551	0.25805
pram + ma (l)	0.17682	0.12200	0.22554	0.29064	0.21946
pram + ma (u)	0.27230	0.15065	0.24197	0.33822	0.26172
rec + ls + shuffle (l)	−0.01865	0.09365	0.18510	0.19294	0.19859
rec + ls + shuffle (u)	0.24584	0.12496	0.20950	0.25071	0.26183

Table 3. Coverage rates for confidence intervals of the gender pay gap in each educational sector between the original and perturbed data

Data	ISCED 0 and 1	ISCED 2	ISCED 3 and 4	ISCED 5A	ISCED 5B
rec + ls + ma	98.25	98.55	96.21	88.45	88.65
rec + ls + noise	89.85	99.86	87.81	91.58	99.26
pram + ma	83.52	96.63	83.45	78.18	95.08
rec + ls + shuffle	81.67	13.51	0.00	0.00	64.67

anonymisation methods. For those anonymisation methods that use (optimal) local suppression, three-anonymity is achieved. 4,414 observations violate three-anonymity in the original data set. The PRAM method performs best in terms of data utility since none of the variables that are used in these estimations are altered. The second best is recoding + local suppression + microaggregation. Recoding + local suppression + shuffling performs worst. The reasons for this could be that continuous variables are shuffled and also shuffled between gender, which is the most important variable when estimating the GPG and that the prediction quality of the model used for the shuffling procedure is low.

In general, recoding + local suppression + microaggregation and pram + microaggregation reports very low bias and clearly outperform shuffling and adding noise.

7.2. *Overlap of Confidence Intervals*

As an example, the upper and lower confidence intervals for the GPG in the domain *education* are given in Table 2. It is easy to see that the length of the confidence intervals is shorter for category ISCED 3-4 and largest for ISCED 0-1.

Again, the shuffling method does not seem to be able to give approximately the same confidence intervals.

A clearer picture is supported by Table 3, where the overlap of the confidence intervals for the GPG – estimated from the perturbed and the original data – is reported.

The coverage rates are relatively high for all methods except recoding + local suppression + shuffling. Differences in some categories are visible when comparing the other methods, whereas no clear ranking of them in terms of quality can be made.

The coverage rates for the gender pay gap in domain *age* (Table 4) are similar. Mostly the recoding + local suppression + microaggregation methods performs slightly better than recoding + local suppression + adding noise and pram + microaggregation.

However, a completely different picture is seen for the absolute relative bias of the Gini index in Table 5. Recoding + local suppression + microaggregation outperforms all other

Table 4. Coverage rates for confidence intervals of the GPG in each age class between the original and perturbed data

Data	(0,19)	(19,29)	(29,39)	(39,49)	(49,59)	(59,120)
rec + ls + ma	98.81	76.40	99.28	82.41	95.82	91.45
rec + ls + noise	94.90	80.27	94.31	89.60	89.70	96.76
pram + ma	84.26	88.92	95.02	88.55	92.58	86.94
rec + ls + shuffle	0.00	32.75	0.00	0.00	0.00	0.00

Table 5. Coverage rates for confidence intervals of the Gini indices in each age × gender domain between the original and perturbed data

Data	(0,19):f	(0,19):m	(19,29):f	(19,29):m	(29,39):f	(29,39):m
rec + ls + ma	93.64	81.66	96.83	94.93	89.24	95.63
rec + ls + noise	52.71	0.00	22.12	37.18	63.09	87.92
pram + ma	88.29	82.49	88.39	93.05	85.36	94.50
rec + ls + shuffle	82.61	0.00	0.00	0.00	20.38	0.00
	(39,49):f	(39,49):m	(49,59):f	(49,59):m	(59,120):f	(59,120):m
rec + ls + ma	84.69	75.33	99.21	94.59	95.40	92.22
rec + ls + noise	88.49	83.07	80.52	89.70	93.94	96.03
pram + ma	97.89	85.00	96.78	82.25	88.31	94.94
rec + ls + shuffle	12.55	55.93	0.00	0.00	0.00	0.00

Table 6. Regression coefficients

	Original	rec + ls + ma	rec + ls + noise	pram.ma	rec + ls + shuffle
(Intercept)	1.50454	1.52627	1.51374	1.40474	1.63726
Sexmale	0.20478	0.20484	0.20433	0.20970	0.19733
age(19,29]	0.57210	0.57190	0.58560	0.57659	0.76536
age(29,39]	0.73750	0.73745	0.75186	0.74388	0.91469
age(39,49]	0.81758	0.81746	0.83260	0.82634	0.96199
age(49,59]	0.85660	0.85597	0.87072	0.86754	0.89338
age(59,120]	0.81553	0.81067	0.82604	0.82169	0.49264
educationISCED 2	0.03692	0.02006	0.01011	0.03834	−0.25102
educationISCED 3 and 4	0.28314	0.26646	0.25737	0.28667	−0.16874
educationISCED 5A	0.73406	0.71508	0.70647	0.74198	0.09813
educationISCED 5B	0.44484	0.42802	0.41959	0.45337	−0.01251
LocationAT2	−0.07516	−0.07523	−0.07528	−0.06368	−0.00673
LocationAT3	−0.01230	−0.01207	−0.01132	−0.00900	−0.00098
NACE1D-Manufactoring	−0.05542	−0.06029	−0.05441	0.01740	−0.01600
NACE1E-Electricity	0.09709	0.09018	0.09264	0.12244	−0.02588
NACE1F-Construction	−0.12280	−0.12891	−0.12260	−0.03775	−0.01806
NACE1G-Trade	−0.18916	−0.19422	−0.18848	−0.09872	−0.02576
NACE1H-Hotels	−0.37478	−0.37962	−0.37589	−0.24398	−0.02269
NACE1I-Transport	−0.17130	−0.17632	−0.17061	−0.07943	−0.00939
NACE1J-FinancInt	0.14921	0.14532	0.15055	0.19273	−0.01993
NACE1K-RealEstate	−0.13433	−0.13901	−0.13517	−0.05156	−0.02072
NACE1M-Education	−0.16289	−0.16650	−0.16300	−0.07845	−0.02505
NACE1N-Health	−0.11299	−0.11734	−0.11360	−0.02939	−0.01838
NACE1O-Other	−0.19113	−0.19585	−0.19353	−0.10283	−0.01054

methods. PRAM + microaggregation also gives acceptable results but recoding + local suppression + adding noise gives low coverage rates for age classes below 29 years. Shuffling results in the estimates with the highest bias.

7.3. Differences in Regression Coefficients

As already mentioned, to compare the regression coefficients of original and anonymised data sets, the same categories in the explanatory variables of the model must be present. Thus the recoded twelve categories of *economic activity* are used also for the original data set, keeping in mind that this means a certain kind of information loss.

In Table 6 the regression coefficients for the original and the anonymised data sets are shown.

The regression coefficients and their confidence intervals are visualised in Figure 1, whereas the original estimates (in black) are compared with the estimates from anonymised data (in grey).

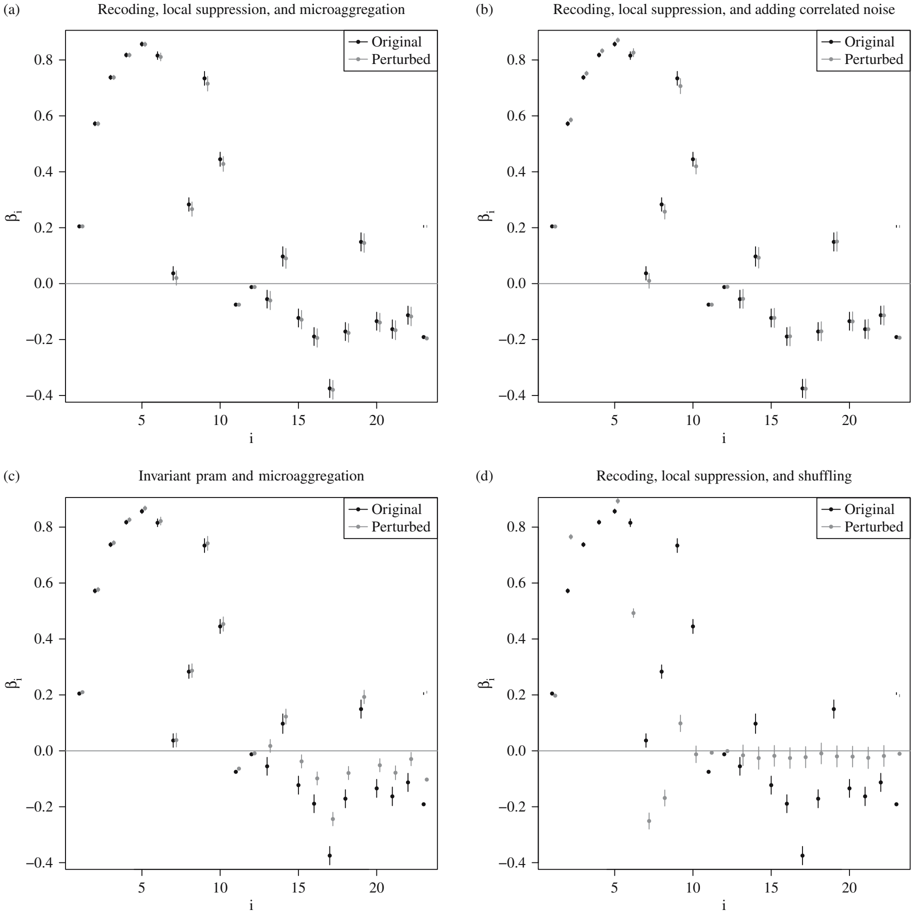


Fig. 1. Confidence intervals for the regression coefficients for the original data (black lines) and the perturbed data (grey dotted lines).

Recoding + local suppression + microaggregation again performs best and the confidence intervals obtained from the anonymised data almost always cover the confidence intervals obtained from the original data completely. Almost as good is the quality of data anonymised by recoding + local suppression + adding correlated noise. The results from invariant pram + microaggregation are good for all coefficients except those related to *economic activity*. This is not surprising, since this variable was one of the variables which was changed using PRAM. Some few coefficients are well preserved from the recoding + local suppression + shuffling anonymised data, but others are not. The reason is that even if the distribution of the continuous shuffled variables is well preserved, the relation to other variables that are not included in the shuffling model might be not preserved. A better model would probably lead to better results.

8. Conclusions

This article focuses upon the use of the most important measures of a particular survey as quality indicators of utility to evaluate anonymised data sets.

As a case study, the use of the Structure of Earnings Survey is analysed in detail in order to identify the most important variables, indicators and models applied to this data set. Based on the knowledge gained, the most important indicators are selected and the data utility of the anonymised data is evaluated; the disclosure risk is briefly reported. The evaluation is done on point and variance estimates from the selected indicators as well as on inferences on regression coefficients of a selected model. The evaluation of the regression coefficients in particular shows various problems with data utility. Thus such a comparison of model estimates should always be focused upon especially because a model reflects the multiple relationships between variables. Out of hundreds of different possible models, those models that are most often applied in practice should be chosen and an analysis of the literature is therefore necessary. The aim is to preserve the estimates from the most-used indicators and models and those anonymisations should be chosen that achieve both the minimum requirements in terms of disclosure risk and high precision on the chosen quality indicators.

The aim of this investigation was not to find the best anonymisation procedure from a risk perspective, but how to evaluate data utility. Nevertheless, four different possible anonymisations were applied and evaluated. The best results are obtained by the anonymisation: recoding + local suppression to achieve three-anonymity + microaggregation in each stratum defined by economic activity. For the invariant pram method, some problems become visible for those variables that have been ‘pramed’. The shuffling method did not perform well, but this may depend on the shuffling model used (in our study several models were tested and the best was chosen); good results on other data sets may perform better as the method seems very promising (see, e.g., Muralidhar and Sarathy 2006).

This case study is only focused on one particular survey, the Structural Earnings Statistics survey, but we have demonstrated a general concept of how to identify the most important indicators and models and how to evaluate the quality of the protected data based on estimates of these indicators. Although this key idea is not new in principle, it is demonstrated practically in a large case study in a larger setting.

The used (and other) indicators have been implemented in the R package **laeken** (Alfons and Templ 2013), which makes the application of the methods to complex data, such as the SES, easy.

9. References

Alfons, A. and M. Templ. 2013. "Estimation of Social Exclusion Indicators from Complex Surveys: The R package laeken." *Journal of Statistical Software* 54: 1–25.

Alfons, A., S. Kraft, M. Templ, and P. Filzmoser. 2011. "Simulation of Close-to-Reality Population Data for Household Surveys with Application to EU-SILC." *Statistical Methods & Applications* 20: 383–407. doi:10.1007/s10260-011-0163-2.

Alfons, A., M. Templ, and P. Filzmoser. 2013. "Robust Estimation of Economic Indicators from Survey Samples Based on Pareto Tail Modeling." *Journal of the Royal Statistical Society Series C* 62: 271–286.

Beblot, M., D. Beniger, A. Heinze, and F. Laisney. 2003. *Methodological Issues Related to the Analysis of Gender Gaps in Employment, Earnings and Career Progression*. Final Project Report, European Commission Employment and Social Affairs DG.

Belfield, R. 1999. *Pay Inequalities and Economic Performance: A Review of the UK Literature*. Technical Report PiEP Report, Centre for Economic Performance, London School of Economics.

Bowles, S., H. Gintis, and M. Osborne. 2001. "The Determinants of Earnings: a Behavioral Approach." *Journal of Economic Literature* 39: 1137–1176.

Brand, R. 2004. "Microdata Protection through Noise Addition." In *Privacy in Statistical Databases. Lecture Notes in Computer Science*, edited by J. Domingo-Ferrer. 347–359. New York: Springer.

Bruch, C., R. Münnich, and S. Zins. 2011. *Variance Estimation For Complex Surveys*. Research Project Report WP3–D3.1, FP7-SSH-2007-217322 AMELI. Available at: http://ameli.surveystatistics.net (accessed December 2013)

Caju, P., C. Fuss, and L. Wintr. 2009a. "Understanding Sectoral Differences in Downward Real Wage Rigidity: Workforce Composition, Institutions, Technology and Competition." Working Paper Series no. 1006, European Central Bank. Available at: http://www.ecb.int/pub/pdf/scpwps/ecbwp1006.pdf (accessed December 2013)

Caju, P., F. Rycx, and I. Tojerow. 2009b. "Inter-industry Wage Differentials: How Much Does Rent Sharing Matter?" *Journal of the European Economic Association* 79: 691–717.

Caju, P., F. Rycx, and I. Tojerow. 2010. "Wage Structure Effects of International Trade: Evidence From a Small Open Economy." Working Paper Series no. 1325, European Central Bank. Available at: http://www.ecb.int/pub/pdf/scpwps/ecbwp1325.pdf (accessed December 2013)

Carlson, M. 2002. "Assessing Microdata Disclosure Risk Using the Poisson-inverse Gaussian Distribution." *Statistics in Transition* 5: 901–925.

Casali, S. and V. Alvarez. 2010. *17% of Full-time Employees In the EU Are Low-wage Earners. Statistics in focus*. Research Report. KS-SF-10-003-EN-N, Eurostat/European Commission. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-SF-10-003/EN/KS-SF-10-003-EN.PDF (accessed December 2013)

Dell'Aringa, C., P. Ghinetti, and C. Lucifora. 2000. "Pay Inequality and Economic Performance in Italy: a Review of the Applied Literature." In Proceedings of the LSE conference, November 3–4, 2000. 1–28. London.

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., C.-E. Särndal, and O. Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88: 1013–1020.

Domingo-Ferrer, J. and V. Torra. 2001. "A Quantitative Comparison of Disclosure Control Methods for Microdata." *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, edited by P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz. 111–134, Eurostat.

Domingo-Ferrer, J., J.M. Mateo-Sanz, and T. Torra. 2001. "Comparing sdc Methods for Microdata on the Basis of Information Loss and Disclosure." Proceedings of ETK-NTTS 2001: Eurostat, Luxembourg June 18–20, 2001. 807–826. Luxembourg: Eurostat.

Dupray, D., H. Nohara, and P. Béret. 1999. *Pay Inequality and Economic Performance: a Review of the French Literature*. Technical Report PiEP Report, Centre for Economic Performance, London School of Economics

Dupré, D. 2010. "The Unadjusted Gender Pay Gap in the European Union." In Joint UNECE/Eurostat Work Session on Gender Statistics, Geneva April 14–16, 2010. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.30/2010/1.e.pdf (accessed November 2015)

Dybczak, K., and K. Galuscak. 2010. "Changes in the Czech Wage Structure: Does Immigration Matter?" Working Paper Series no. 1242, European Central Bank. Available at: http://www.ecb.int/pub/pdf/scpwps/ecbwp1242.pdf (accessed December 2013)

Edwards, C. 2010. "Public Sector Unions and the Rising Costs of Employee Compensation," *Cato Journal* 30: 87–115.

Efron, B. and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

EU-SILC. 2004. *Common Cross-sectional EU Indicators Based on EU-SILC: the Gender Pay Gap*. EU-SILC 131-rev/04, Working Group on Statistics on Income and Living Conditions (EU-SILC). Luxembourg: Eurostat.

EU-SILC. 2009. Algorithms to Compute Social Inclusion Indicators Based On EU-SILC and Adopted under the Open Method of Coordination (OMC). EU-SILC LC-ILC/39/09/ENrev.1, Directorate F: Social and Information Society Statistics Unit F-3: Living Conditions and Social Protection, European Commission. Luxembourg: Eurostat.

Fitzenberger, B., K. Kohn, and A. Lembcke. 2006. *Union Wage Effects in Germany: Union Density Or Collective Bargaining Coverage?* Research Report FSP 1169, DFG research programme, The London School of Economics and Political Sciences, London.

Franconi, L. and S. Polettini. 2004. "Individual Risk Estimation in μ-Argus: a Review." In *Privacy in Statistical Databases: Lecture Notes in Computer Science*, edited by J. Domingo-Ferrer. 262–272. New York: Springer.

Franconi, L., D. Ichim, and M. Templ. 2011. *First Steps to Define a Framework For Comparable Dissemination of the European Structure of Earning Survey. Deliverable d1.1-a. Task 1: Harmonisation of Microdata Release in Multiple Countries*. Essnet Project on Common Tools and Harmonised Methodologies for SDC in the ESS. Available at: http://neon.vb.cbs.nl/casc/..%5Ccas%5CESSNet2%5Cdeliverable%201%20full%20august2012.pdf (accessed November 2015)

Frick, B., and K. Winkelmann. 1999. *Pay Inequalities and Economic Performance: A Review in Literature*, Technical Report Research Report HPSE-CT-1999-00040, Ernst-Moritz-Arndt-Universität Greifswald.

Geissberger, T. 2009. *Verdienststrukturerhebung 2006, Struktur und Verteilung der Verdienste in Oösterreich*. Vienna: Statistik Austria.

Geissberger, T. 2010. *Frauenbericht. Teil 4: Sozioökonomische Studien*, Technical Report 4, Federal Ministry for Women and the Civil Service of Austria.

Geissberger, T. and K. Knittler. 2010. "Niedriglöhne und Atypische Beschäftigung in Österreich." *Statistische Nachrichten* 6: 448–461.

Gini, C. 2012. "Variabilità e Mutabilità: Contributo Allo Studio delle Distribuzioni e delle Relazioni Statistiche." *Studi Economico-Giuridici della R. Università di Cagliari* 3: 3–159.

Gomatam, S. and A. Karr. 2003. *Distortion Measures for Categorical Data Swapping*. Report no. 131, National Institute of Statistical Sciences (NISS).

Gouweleeuw, J., P. Kooiman, L. Willenborg, and P-P. De Wolf. 1998. "Post Randomisation for Statistical Disclosure Control: Theory and Implementation." *Journal of Official Statistics* 14; 463–478.

Graf, M., A. Alfons, C. Bruch, P. Filzmoser, B. Hulliger, R. Lehtonen, B. Meindl, R. Münnich, T. Schoch, M. Templ, M. Valaste, A. Wenger, and S. Zins. 2011. *State-of-the-art of laeken Indicators*. Research Project Report WP1 – D1.1, FP7-SSH-2007-217322 AMELI. Available at: http://ameli.surveystatistics.net (accessed December 2013)

Groshen, E. 1991. "The Structure of the Female/Male Wage Differential." *Journal of Human Resources* 26: 455–472.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer, and P.-P. de Wolf. 2012. *Statistical Disclosure Control*. New York: Wiley.

Ichim, D. and L. Franconi. 2007. "Disclosure Scenario and Risk Assessment: Structure of Earnings Survey." In Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Manchester, December 17–19, 2007. Doi: 10.2901/Eurostat.C2007.004

Ichim, D. and L. Franconi. 2010. "Strategies to Achieve sdc Harmonisation at European Level: Multiple Countries, Multiple Files, Multiple Surveys." *Privacy in Statistical Databases '10*, edited by J. Domingo-Ferrer and E. Kajkos, Springer, New York. 284–296.

Karr, A.F., C.N. Kohnen, A. Oganian, J.P. Reiter, and A.P. Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." *The American Statistician* 60: 224–232. Doi: 10.1198/000313006X124640.

Kolb, J.-P., R. Münnich, S. Beil, A. Chatziparadeisis, and J. Seger. 2011. *Policy Use of Indicators on Poverty and Social Exclusion*. Research Project Report WP9–D9.1, FP7-SSH-2007-217322 AMELI, 2011. Available at: http://ameli.surveystatistics.net (accessed December 2013)

Lorenz, M.O. 1905. "Methods of Measuring the Concentration of Wealth." *Publications of the American Statistical Association* 9: 209–219.

Manning, A.M., D.J. Haglin, and J.A. Keane. 2008. "A Recursive Search Algorithm For Statistical Disclosure Assessment." *Data Mining and Knowledge Discovery* 16: 165–196. Doi: 10.1007/s10618-007-0078-6.

Marsden, D. 2010. *Pay Inequalities and Economic Performance*, Technical Report PiEP Final Report V4, Centre for Economic Performance, London School of Economics. London: London School of Economics. Available at: http://www.ist-world.org/ProjectDetails.aspx?ProjectID=fa5bb4adfff74d60aeca90b56441a601&SourceDatabaseID=9cd97ac2e51045e39c2ad6b86dcelac2.

Messina, J., M. Izquierdo, P. Caju, C.F. Duarte, and N.L. Hanson. 2010. "The Incidence of Nominal and Real Wage Rigidity: an Individual-based Sectoral Approach." *Journal of the European Economic Association* 8: 487–496.

Mittag, J. 2005. *Gross Earnings In Europe. Main Results of the Structure of Earnings Survey 2002*. Statistics in Focus. Research Report. KS-NK-05-012-EN-N, European Communities. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-NK-05-012/EN/KS-NK-05-012-EN.PDF (accessed December 2013)

Muralidhar, K. and R. Sarathy. 2006. "Data Shuffling – a New Masking Approach for Numerical Data." *Management Science* 52: 658–670.

Nolan, B. and H. Russel. 2001. *Pay Inequality and Economic Performance In Ireland: a Review of the Applied Literature*. Technical Report PiEP Report, The Economic and Social Research Institute, Dublin.

Oganian, A. and A.F. Karr. 2006. "Combinations of sdc Methods for Microdata Protection." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and L. Franconi. 102–113. Berlin: Springer. Doi: 10.1007/11930242_10.

Pointner, W., and A. Stiglbauer. 2010. "Changes In the Austrian Structure of Wages." Working Paper Series no. 1268, European Central Bank. Available at: http://www.ecb.int/pub/pdf/scpwps/ecbwp1268.pdf (accessed December 2013)

Reiter, J.P. 2012. "Statistical Approaches to Protecting Confidentiality For Microdata and their Effects on the Quality of Statistical Inferences." *Public Opinion Quarterly* 76: 163–181. Doi: 10.1093/poq/nfr058.

Research Center for Education and the Labour Market at Maastricht University. 2009. "Development of Econometric Methods to Evaluate the Gender Pay Gap Using Structure of Earnings Survey Data." Research paper no. ks-ra-09-011-en-n, European Commission. Available at: http://www.ecb.int/pub/pdf/scpwps/ecbwp1006.pdf (accessed December 2013)

Reuter, W. 2010. *Establishing an Infrastructure for Remote Access to Microdata at Eurostat*. Bachelor's thesis., Vienna Univesity of Economics.

Reuter, W. and J-M. Museux 2010. "Establishing an Infrastructure for Remote Access to Microdata at Eurostat." In *Privacy in Statistical Databases: Lecture Notes in Computer Science*, edited by J. Domingo-Ferrer. 249–257. New York: Springer.

Rinott, Y. 2003. “On Models for Statistical Disclosure Risk Estimation.” In Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. April 7–9, 2003. 275–285, United Nations Statistical Commission, Geneva.

Shlomo, N. 2008. “Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Data Utility.” In Section on Survey Research Methods, JSM. August 3–7, 2008, Denver, Colorado, USA. 229–240. Available at: https://www.amstat.org/sections/srms/proceedings/y2008/Files/300242.pdf (accessed November 2015)

Simón, H. 2010. “International Differences in Wage Inequality: A New Glance with European Matched Employer-Employee Data.” *British Journal of Industrial Relations* 48: 310–346.

Stephan, G. and K. Gerlach. 2005. “Wage Settlements and Wage Settings: Evidence from a Multilevel Model.” *Applied Economics* 37: 2297–2306.

Stockinger, S. 2010. *Frauenbericht 2010*. Technical report, Federal Ministry for Women and the Civil Service of Austria. Vienna: Available at: http://www.bka.gv.at/site/6811/default.aspx (accessed December 2013)

Sweeney, L. 2002. “k-Anonymity: a Model for Protecting Privacy.” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10: 557–570.

Templ, M. 2008. “Statistical Disclosure Control for Microdata Using the R-package sdcMicro.” *Transactions on Data Privacy* 1: 67–85.

Templ, M. 2011a. *Estimators and Model Predictions from the Structural Earnings Survey for Benchmarking Statistical Disclosure Methods*. Research Report CS-2011-4, Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria.

Templ, M. 2011b. “Comparison of Perturbation Methods Based on Pre-defined Quality Indicators.” In Joint UNECE/Eurostat work session on statistical data confidentiality, 26–28 October, 2011, Tarragona, Spain, 1–10. Unece, Geneva, Italy.

Templ, M. and A. Alfons. 2011. *Variance Estimation of Social Inclusion Indicators Using the R Package laeken*. Research Report CS-2011-3, Department of Statistics and Probability Theory, Vienna University of Technology. Available at: http://www.statistik.tuwien.ac.at/forschung/CS/CS-2011-3complete.pdf (accessed December 2013)

Templ, M. and P. Filzmoser. 2014. “Simulation and Quality of a Synthetic Close-to-Reality Employer-Employee Population.” *Journal of Applied Statistics*, 41: 1053–1072.

Templ, M. and B. Meindl. 2010. “Practical Applications in Statistical Disclosure Control Using R.” In *Privacy and Anonymity in Information Management Systems: Advanced Information and Knowledge Processing*, edited by J. Nin and J. Herranz. 31–62. London: Springer.

Templ, M. A. Kowarik, and B. Meindl. 2015. “Statistical Disclosure Control for Micro-Data Using R Package sdcMicro.” *Journal of Statistical Software*. 67: 1–36.

Weinberg, D.H. 2007. “Earnings by Gender: Evidence from Census 2000.” *Monthly Labor Review Online* 130: 26–34.

Winter-Ebmer, R. and J. Zweimüller. 1999. “Firm Size Wage Differentials in Switzerland: Evidence from Job Changers.” *American Economic Review* 89: 89–93.

Woo, M., J.P. Reiter, A. Oganian, and A.F. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality* 1: 111–124.

Received October 2012
Revised December 2013
Accepted January 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 723–736, http://dx.doi.org/10.1515/JOS-2015-0042

B-Graph Sampling to Estimate the Size of a Hidden Population

Marinus Spreen[1] *and Stefan Bogaerts*[2]

Link-tracing designs are often used to estimate the size of hidden populations by utilizing the relational links between their members. A major problem in studies of hidden populations is the lack of a convenient sampling frame. The most frequently applied design in studies of hidden populations is respondent-driven sampling in which no sampling frame is used. However, in some studies multiple but incomplete sampling frames are available. In this article, we introduce the B-graph design that can be used in such situations. In this design, all available incomplete sampling frames are joined and turned into one sampling frame, from which a random sample is drawn and selected respondents are asked to mention their contacts. By considering the population as a bipartite graph of a two-mode network (those from the sampling frame and those who are not on the frame), the number of respondents who are directly linked to the sampling frame members can be estimated using Chao's and Zelterman's estimators for sparse data. The B-graph sampling design is illustrated using the data of a social network study from Utrecht, the Netherlands.

Key words: Network sampling; capture recapture; hidden populations.

1. Introduction

Estimating the sizes of hidden populations is important in the field of official statistics in order to provide local or national institutions with insights into the nature and extent of a social problem. Hidden populations are characterized by the lack of well-defined complete sampling frames due to the privacy-threatening nature of the variable that defines the study population (Spreen 1992; Heckathorn 1997). Privacy-threatening traits are often illegal activities and/or activities that are not socially accepted. Examples of illegal activities are drug trafficking, human trafficking, sexual abuse, child abuse, domestic violence, terrorist activities, criminal acts, and so on (e.g., Brugal et al. 1999; Holland et al. 2006; Surjadi et al. 2010; Kunst et al. 2010; Palusci et al. 2010). Depending on the culture and/or legal system of a nation, privacy-threatening traits can also be activities that are not socially accepted, like drug use, selling sex, buying sex, undeclared work, or tax evasion (Bogaerts and Daalder 2011). Requesting privacy-threatening information from members of hidden populations will lead to high rates of uncooperative individuals or unreliable answers (Heckathorn 1997). Two different data collection procedures can be distinguished for estimating the size of a hidden population. In capture-recapture procedures, official

[1] Stenden University of Applied Sciences-School of Social Work and Art Therapies, Rengerslaan 8 Leeuwarden 8917 DD, The Netherlands. Email: Marinus.Spreen@Stenden.com
[2] Tillburg University, Department of Developmental Psychology, Warandelaan 2, 5037 AB Tilburg, The Netherlands. Email: s.bogaerts@uvt.nl

registration sources are used as sampling frames to estimate hidden population sizes. In link-tracing or network sample procedures, social links between hidden population members are used as sampling frames for estimation purposes. The difference between these two procedures lies in the way data are collected. In capture-recapture procedures, hidden population members themselves are not sampled and interviewed, only registered. In link-tracing procedures, hidden population members are sampled and interviewed about their social links with other members of that hidden population.

In this article, a practical sampling design called the B-graph sampling design is introduced and illustrated. This design has been elaborated for research contexts in which one or multiple registration sources are available but each source on its own is considered too small to produce valid capture-recapture estimations. However, if pooling all available registration sources results in a substantial coverage of the unknown population according to local experts, this pooled source can be considered a plausible sampling frame to start a link-tracing data collection procedure. For example, all neighbourhood youth workers in a city agree that the number of names on the pooled list cover a substantial part of the total unknown population. For estimation purposes, the population of interest can be divided into two subpopulations, namely registered and unregistered persons. Drawing a probability sample from the (pooled) registered part of the hidden population and employing a link-tracing procedure by asking each sampled person to disclose his contacts with other hidden population members, the size of the unregistered subpopulation directly linked to registered persons can be estimated. Furthermore, if the assumption that each unregistered person of the study population has at least one direct link to a registered person is held to be plausible, each unregistered person has a positive probability of being included in the link-tracing sample. Thus the resulting estimate gives an indication of the total population size. The estimation problem of the number of persons directly linked to a known subset of persons is of interest in a variety of (forensic) social network studies. For instance, if the known set of persons is hooligans or gang members, the number of directly related unregistered hooligans or gang members can be estimated. If the known subset of persons is arrested problem youths in some city, their number of contacts with other youths may provide valuable information about the size of the problem.

In this article, we discuss three capture-recapture estimators by considering the hidden population as a bipartite graph of a two-mode network (registered and unregistered persons); for example, we focus on the social links between the two subpopulations. This approach is illustrated by data obtained from a social network study conducted among the population of opiate users in the city of Utrecht, the Netherlands (Ten Den et al. 1995). In the original study, three sampling procedures were applied: a random sample from the files of three drug-assistance organisations, a convenience fieldwork sample and a snowball sample to find unregistered opiate users. To illustrate the B-graph design, the three client lists are pooled into one sampling frame (excluding the respondents from the convenience and snowball sample), from which a random sample is drawn and a link-tracing procedure applied to sample unregistered opiate users. The statistical problem is to estimate the number of unregistered opiate users directly related to the clients of the aid agencies. The outline of the article is as follows. In Section 2, a brief review of capture-recapture techniques for hidden population size estimation using administrative sources and estimation techniques for research contexts in which sampling frames are lacking is given.

Section 3 introduces the proposed B-graph sampling design. Because newly mentioned users will be rather sparse in most contexts, we focus on size estimators based on multiple-capture techniques for sparse data in Section 4 (Chao 1987; 1988; 1989; Zelterman 1988; Böhning 2010). Finally, Section 5 is concerned with the illustration, and the article ends with some concluding remarks.

2. Review of Literature on Estimating Hidden Populations

According to Böhning and van der Heijden (2009), capture-recapture methods are conventionally used to estimate the size of a hidden population when only (multiple) registration sources are available. In particular, the so-called Petersen-Lincoln (PL) estimator has been widely applied in animal studies, but nowadays this estimation technique is also employed in social studies where two registration sources are available (McCullough and Hirth 1998; Chao et al. 2008). The PL estimator is based on the number of n_1 units captured in Source 1, the number of n_2 units captured in Source 2, and the number of m_2 units captured in both sources. By assuming that the two sources are independent of each other, the units not captured in one of the sources can be estimated because the odds ratio is close to unity (Brittain and Böhning 2009).

$$\hat{N}_{PL} = \frac{n_1 n_2}{m_2} \tag{1}$$

The standard procedure for estimating the size of an animal population in a two-sample capture-recapture study is to capture a first sample, mark the captured animals and release them. Subsequently, a second sample is captured, and the number of animals captured in the first, the second and both samples is used to estimate the size of the population with the PL estimator (1). The standard procedure for estimating the size of a human population where two registration sources are available mirrors the animal population procedure by considering persons on the lists to be “marked”. Like the trapping samples in animal studies, the number of persons on the first, the second, and both lists are used to employ the PL estimator (1). Examples of registration sources are hospitals, treatment centres, pharmacies, police registers, birth registers, and so on. The assumptions for producing valid estimates by capture-recapture methods are more or less identical in animal and in human population studies. According to Chao (2001), the validity of a capture-recapture estimator for animal populations depends on:

1. Demographic closure assumption: there is no birth, death, or migration, so that the population size is stable over trapping times;
2. Equal catchability assumption: all animals have the same capture probability in each sample, although the probability can be allowed to vary among samples.

To fulfil the first assumption, in animal studies data are collected during a relatively short time period. The second assumption refers to the independence of the samples. Dependence between samples can occur through local list dependence and unequal catchabilities (Chao et al. 2008). Local list dependence occurs whenever captured animals are easier or more difficult to capture by next samples as a consequence of their trapping history. Unequal catchability refers to the process that samples are dependent because their capture probabilities are heterogeneous (Chao 2001).

To produce valid estimators in human populations, the following assumptions must be met (Brittain and Böhning 2009):

1. Independence between registration sources or lists,
2. The population must be closed,
3. Independence between individuals.

In most empirical situations, these assumptions are violated. For instance, in drug abuse studies the registration sources of addiction centres and police registers are often combined to estimate the size of the number of drug users who are not registered. However, both data sources may have administration flaws. If arrested drug users are structurally assigned to certain addiction centres, Assumption 1 is violated. If there is also a high death or removal rate, Assumption 2 is violated. If certain ethnic groups of drug users are treated by the same institution, Assumption 3 is violated. There is a growing amount of literature on how to deal with these types of dependencies (see the special issues of the Biometric Journal, 2008, volume 50, the AStA Advances in Statistical Analysis, 2009, volume 93 and Journal of Official Statistics, volume 31).

In some empirical research contexts, registration sources are simply lacking or of such poor quality (for example, incomplete registration systems) that valid capture-recapture estimation is debatable. In such situations, link-tracing sampling procedures can be applied (Spreen 1992). Link-tracing designs use existing relational structures within the study population for sampling purposes. Up-to-date respondent-driven sampling (RDS) is the link-tracing procedure applied most frequently to estimate hidden populations sizes when (proper) sampling frames are lacking (Heckathorn 1997; Salganik and Heckathorn 2004; Volz and Heckathorn 2008). In RDS, the hidden trait to be estimated is viewed as a network phenomenon because it is assumed "that those best able to access members of hidden populations are their own peers" (Heckathorn 1997, 178). The sampling procedure starts with the recruitment of individuals (called "seeds") from the target population. This recruitment is nonrandom. The recruited individuals are offered dual incentives: they are financially rewarded for completing the interview and for recruiting other individuals (typically 3-5 persons) into the study. Subsequently, the newly recruited persons are asked to become recruiters themselves and are also rewarded financially. To estimate the size $\hat{y}$ of a hidden population, Volz and Heckathorn (2008) defined the RDS estimator (Formula 7, p. 85) as:

$$\hat{y} = \frac{1}{\sum_{i \in S} \frac{1}{d_i}} \sum_{i \in S} \frac{y_i}{d_i}, \tag{2}$$

where S is the set of all sampled persons and d_i the number of persons mentioned by i (degree).

The RDS estimator takes account of the network structure within the hidden population by weighing each interviewed respondent with the number of persons he or she is linked to in the network. These individual degree weights are assumed to be arbitrary positive inclusion probabilities which can be expanded to reach the level of the whole population (Särndal et al. 1992). According to Volz and Heckathorn (2008), it is usually prudent to exclude the initial recruits of the sample because they are not randomly found, although the estimator will be asymptotically unbiased.

Other link-tracing design-based estimators for hidden population sizes are the Frank and Snijders estimators (1994). Like RDS, their sampling design is based on the assumption that the population of interest can be viewed as a social network. In their theoretical (one-wave snowball) design, a random sample of n persons (vertices) is drawn from an unknown network and the selected persons are asked to mention other persons (their degree) they know in the network. Frank and Snijders propose the following estimator:

$$\hat{v}_{F-S} = \frac{(n-1)T_{01}}{T_{00}} + n, \tag{3}$$

where n is the size of the initial sample, T_{00} the number of times initial respondents mentioned each other, and T_{01} the number of times newly mentioned fellow hidden population members are mentioned by initial respondents. Estimator (3) can be understood in terms of capture-recapture, where capture is interpreted as drawn in the initial sample and recapture as mentioned by initial respondents. Frank and Snijders (1994) considered the initial sample to be a Bernoulli sample, which is not feasible in practical research. To relax this assumption, they recommend using some variant of targeted sampling (Watters and Biernacki 1989). To approximate a Bernoulli initial sample to a reasonable extent, Frank and Snijders (1994) recommend using several unrelated sources of well-defined social meeting places during the sampling phase. There are other examples of link-tracing designs in literature, such as multiple-wave snowball designs (Goodman 1961; Frank 1979), random-walk designs (Klovdahl 1989), and adaptive sampling designs (Thompson and Frank 2000), which we will not discuss.

RDS and the Frank-Snijders estimators are both elaborated for situations in which sampling frames are lacking. For situations in which various scattered sampling frames are available, B-graph sampling can be used.

3. B-Graph Sampling

Consider a hidden population in some well-defined geographic area for which it is assumed that its members know each other because of the hidden activity. For instance, a group of hooligans know each other because they operate as group against other groups of hooligans, drug users know each other for economic reasons (e.g., procuring drugs, knowing the market), terrorists know each other for political reasons, homeless people know each other from the street, and so forth.

Hidden populations are often registered by multiple administrative sources. For instance, a population of drug users may be registered as clients of a local drug-assistance institution but also as detainees by the police. In this situation, the Petersen-Lincoln estimator (1) for two sample closed experiments can be employed using both registration resources to estimate the number of unregistered drug users. Obviously, the quality of the estimate is dependent on different issues. For instance, administration flaws may render the accuracy of the registration systems too questionable to be valid for capture-recapture estimation. In such situations, one may consider a B-graph sampling procedure. A B-graph sampling design consists of the following steps. In Step 1, it is decided whether the hidden activity to be estimated leads to relations and/or administrative records by different institutions. Step 2 consists of collecting all available administrative records of all relevant institutions; all collected individual records are turned into one sampling frame and a local team of fieldworkers evaluate whether the persons on the list cover a substantial part of the population. Most of the time, local field workers have a good overview of their caseloads and neighbourhood (Heckathorn 1997). If the constructed sampling frame is considered to cover a substantial part of the population, the unknown total population can be considered as a bipartite graph (Figure 1).

For argument's sake, the four uncoloured vertices represent registered hidden population members pooled into one sampling frame from different sources, that is, sampling frame $\alpha = \{1, 2, 3, 4\}$. The unknown hidden populations members are coloured vertices, that is, subset $\beta = \{5, 6, 7\}$. Note that all coloured vertices have at least one link to an uncoloured vertex, that is, all unregistered hidden population members have a positive probability of being included in a sample when the registered hidden population members are asked to give their relations with unregistered hidden network members. In this article, the problem of estimating the number of unknown hidden population members (coloured vertices) is considered.

In Step 3 of a B-graph sample, a simple random sample S of s vertices from sampling frame α is drawn. Each sampled $i \in S$ is asked to mention his or her relations with other hidden population members according to a predefined inclusion criteria. As a result, a sample of subset β is observed. Throughout this article, we assume that this observation is without measurement error (each respondent completely discloses his contacts in the hidden network). The total number of observed distinct unregistered hidden population members (coloured vertices) in the final sample is denoted $m(S)$. The number of unregistered $u \in \beta$ mentioned exactly t times by the s selected registered hidden population members is denoted f_t, that is, $\sum_{t=1}^{s} f_t = m(S)$. As an illustration, consider Figure 2, in which a sample S of $s = 2$ uncoloured vertices from α is drawn from the bipartite graph of Figure 1. The selected uncoloured vertices are vertices 2 and 4.

In Figure 2, the total number of distinct vertices $u \in \beta$ observed is $m(S) = 2$, that is, vertices 6 and 7. Vertex 7 is involved two times with a vertex $i \in S$, while vertex 6 is involved one time, that is $f_2 = 1$ and $f_1 = 1$, respectively. Using this sample information, multiple capture-recapture estimators for the size of vertex set β can be employed.

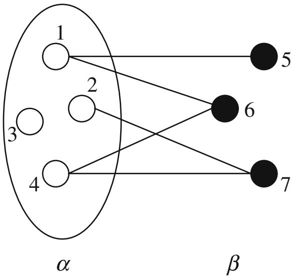


Fig. 1. Bipartite graph example

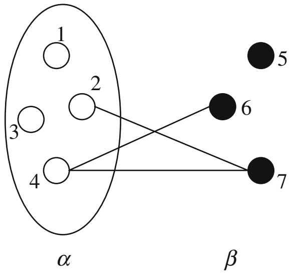


Fig. 2. Sample of bipartite graph

4. Multiple Capture-Recapture Estimators

Data collected from a B-graph sample can be understood as a multiple-capture sample in which each unregistered hidden population member (coloured vertex) captured via $i \in S$ is regarded as an independent trapping sample. Using this assumption, multiple-capture census estimators as discussed in Fienberg (1972), Bishop et al. (1988), and Cormack (1989; 1992) can be applied. However, the larger a population, the more sparse the total times unregistered hidden population members of subset β will be captured via registered members $i \in S$. Dependent on the sampling design and the assumptions about the population, various refinements of multiple-capture models have been introduced, especially for sparsely distributed animal populations. For a general review, we refer the reader to Seber (1986). The review article of Wilson and Collins (1992) merits special attention; it discusses the performance of 14 capture-recapture estimators. In this article, we discuss three capture-recapture estimators whose model assumptions are closely related to the assumptions of the proposed B-graph design: the moment estimator of Chao and a modified version of this estimator (Chao 1987; 1988; 1989) and the truncated Poisson estimator of Zelterman (1988).

Chao (1989) considered estimators for animal population size studies in which capture frequencies of the animals are low. In this study, we focus on the heterogeneity model-based estimator proposed by Chao (1989). This estimator has the following assumptions:

1. The animal population is closed, so there are no changes due to birth, death, emigration or immigration during the sampling period,
2. The probability of capturing an animal is independent of that animal's previous history,
3. Different animals are allowed to have different probabilities of capture.

The proposed B-graph design for human populations meets the assumptions of Chao's estimator. The 'closure' assumption refers to the definition of the inclusion criteria of the hidden population: who belongs to the population? To produce valid estimations, the definition of the hidden population must at least be bounded by strict relational, time and geographic criteria, that is, can you give me your friendly relations with people who have the same hidden variable in common as you, whom you have met during the last three months and who live in your town? The second assumption refers to the sampling procedure of the proposed B-graph design. The probability of an unregistered hidden population member $u \in \beta$ being mentioned by a registered hidden population member is independent of the previous capture history of $u \in \beta$. In the B-graph design, a simple random sample is drawn from the sampling frame. This implies that a multiple capture of an unregistered hidden population member is independent of the registered persons by whom he or she is mentioned. In Chao's terminology, the capture probability of unregistered $u \in \beta$ is independent of the sequence of the samples. The third assumption also applies to the proposed B-graph design. Each unregistered hidden population member $u \in \beta$ is assumed to have at least one contact with a registered hidden population member; this leads to positive inclusion probabilities for all $u \in \beta$ when drawing a sample from α. Accordingly, by random sampling from sampling frame α each $u \in \beta$ has a chance of being mentioned by an $i \in S$. However, different vertices have different probabilities of being mentioned, that is, the higher the degree of $u \in \beta$ in the total population, the higher the probability of being mentioned in the final sample.

For situations where s is not too small (≥ 5) and most unregistered hidden populations members are observed only one or two times, the following estimator of Chao (1988) can be employed:

$$\hat{m}_C = m(S) + \left[\frac{f_1^2}{2f_2}\right]. \tag{4}$$

Chao (1987, 1988) also proposed a biased-corrected version to correct for overestimation bias:

$$\tilde{m}_C = m(S) + \left[\frac{f_1(f_1 - 1)}{2(f_2 + 1)}\right] \tag{5}$$

The computation of the 95-percent confidence intervals of (4) and (5) are found in Chao (1989).

The idea behind Estimator (4) is that unregistered hidden population members of subset β with small capture probabilities (they have few relations in the network with members that are registered) are likely to be not mentioned (frequency class f_0) or only mentioned very few times by $i \in S$. This emphasis on the lower frequency classes makes Estimator (4) robust in the presence of heterogeneity. The influence of unregistered hidden population members mentioned very often is weighted down so that the presence of heterogeneity exercises a small influence on the estimate (Smit et al. 1997).

Based on the intuitive notion that 'individuals never seen are more similar to those rarely seen than those captured many times', Zelterman (1988, 227) formulated, independently of Chao, an estimator for the relative frequency of the unobservable zero class in a truncated Poisson distribution, that is,

$$\hat{m}_Z = \frac{m(S)}{1 - Q_1} \tag{6}$$

where $Q_1 = \exp[-2f_2/f_1]$.

The 95-percent confidence interval is given in Zelterman (1988).

Estimators (4) and (6) will produce about the same estimates, because both assume that the observed series of frequencies follows a Poisson distribution which is truncated below one (Smit et al. 1997). In a simulation study by Böhning (2010), in which the performance of Chao's estimator was compared with Zelterman's estimator, the author showed that the estimators are close if the ratio f_2/f_1 is small. He also showed that the biased-corrected estimator (5) of Chao performs best for small samples and small amounts of heterogeneity.

5. Illustration

To illustrate the B-graph sampling design, data from a social network study of the opiate-using population in the city of Utrecht, the Netherlands (Ten Den et al. 1995; Jansson and Spreen 1998) are used for secondary analyses. Utrecht is one of the largest cities (about 320,000 inhabitants) in the Netherlands and is geographically located in the middle of the country. At the time of the study, the opiate-using population in Utrecht caused a lot of nuisance for the general public, but there was also concern about specific health issues such as the relation between injecting drugs and contagious hepatitis, HIV, and sexually transmitted diseases. The goal of the study was to gain an insight into the nature of opiate use, such as types of users injecting drugs, lifestyles of opiate users, and so on. Another goal of this study was to gain insight in the total number of opiate users in Utrecht. Therefore several estimation techniques were used.

In Utrecht, local authorities managed several drug-assistance institutions that kept registration files of their clients, but worked more or less independently of each other. In the original study, the resulting sample of 101 opiate users was gathered by a random sample of 51 users from the registers of three drug-assistance organisations, by a convenience field work sample in which 37 users were found, and by a snowball sample in which 13 users were found via other users. Each interviewed opiate user was asked to mention other opiate users in Utrecht. Due to privacy reasons and to prevent a high rate of nonresponse, each opiate user was asked to give the first two letters of his or her first and family name, nickname, age, neighbourhood, and whether he or she was known as a client of the drug assistance by his or her fellow drug users. The identification of the respondents was done by a team of experienced field workers. Based on this sample, several estimation techniques were applied to estimate the prevalence of opiate users in Utrecht. It was possible to compute a Peterson-Lincoln estimate by using the registration files of the police and the largest drug-assistance organisation in Utrecht. The Petersen-Lincoln estimate was about 1,100 users. Furthermore, two extrapolation estimators (Smit et al. 1996) based on the registers of the largest drug-assistance organisation and the police were computed. Based on the first source, the estimate for the total population was about 1,000 users; for the second source (police data), the estimate was about 900 users. Finally, 69 users (51 of the random sample and 18 of the users found during field work) were evaluated as collected independently of each other, and served as the initial "random" sample for the Frank-Snijders estimators. Two network estimators of Frank and Snijders (1994) were reported (without standard errors) and resulted in estimates of 759 and 936 users. Finally, the researchers combined all different estimators and decided that the most likely estimate for the population size of the Utrecht opiate users population was about 950 users (Ten Den et al. 1995). The final report of Ten Den et al. (1995) did not provide the confidence intervals of the estimates.

To illustrate the B-graph sampling design, we were able to use the random sample of size 44 from the largest drug-assistance organisation. We call this the Regular Drug

Assistance (RDA). Note that our purpose is to estimate the number of opiate users who are not clients of the RDA but directly related to a user who is a client. In other words: how many opiate users in Utrecht are not known to the RDA but could be contacted via the RDA's clients? This is important information for the effectiveness of all kinds of health measures.

In Utrecht at the time of the study, 427 drug abusers were recorded as clients of the RDA, that is, $\alpha = \{1, 2 \ldots, 427\}$. A simple random sample without replacement S of size $s = 44$ was drawn and each $i \in S$ was asked to mention his/her contacts with other opiate users. This way, a respondent could mention not only other opiate users already registered by the RDA but also opiate users who were not registered on the RDA list. For each mentioned opiate user, the respondent gave individual and identifying characteristics. The criteria for opiate users to be included in the sample were:

1. the mentioned opiate user is a resident of the city of Utrecht or resides in Utrecht at least (at a minimum of) four days a week;
2. the mentioned opiate user has used opiates a minimum of 25 times in the past six months;
3. the respondent and the mentioned opiate user must know each other by first and family name.

Of the 44 selected clients, six refused to provide information about their opiate-using contacts. The remaining 38 clients mentioned 98 other opiate users who were not on the RDA list, that is, $m(S) = 98$. The 38 respondents reported 107 relations with the 98 mentioned opiate users. As a result, the observed frequency distribution of the sampled B-graph was rather sparse (see Table 1).

In Table 2 the three multiple-capture estimates and their 95-percent confidence intervals are given.

As expected, the estimates of the Chao and Zelterman estimator are close to each other, 538 and 535 respectively, because the ratio f_2/f_1 is rather small. Taking into account the 95-percent confidence intervals of the model-based estimators, we observe some differences. The underlying assumptions of the Chao and Zelterman estimators applied to this specific study can be regarded as plausible. The population can be considered closed, because respondents report only other opiate users whom they know by name and live in Utrecht and the practical sample was done in a time frame of three months. The number to be estimated can be understood as the number of opiate users directly connected to the clients of the RDA. The probability of capturing unregistered opiate user k via registered opiate user i is independent of registered user h because i and h are randomly selected from the register. The probability of capturing an unregistered opiate user is dependent on his or her amount of contacts with registered opiate users. The 95-percent confidence regions are rather large, but this is characteristic for sparse frequency distributions. The confidence

Table 1. Capture frequency distribution of mentioned opiate users

F_t	1	2
Counts	89	9

Table 2. Results of different estimators of opiate using population directly linked to clients of the RDA-lists

Estimator	Lower bound	Point estimate	Upper bound
Chao	306	538	1,031
Chao modified	293	490	886
Zelterman	340	535	1,307

region of the Zelterman estimator in particular is known to produce anomalous values caused by small standard errors close to zero (Wilson and Collins 1992).

Following various simulation studies, Chao (1989) concluded that her proposed moment estimator performed best for sparse populations. Furthermore, in a simulation study by Wilson and Collins (1992), Chao's estimator performed best in heterogeneous populations. Böhning (2010) showed in a simulation study that Chao's modified estimator performs best for small samples and small amounts of heterogeneity. In Table 2, the modified estimator has a smaller variance than the other two. However, these simulation results are based on slightly different sampling schemes. Based on the three estimators, we may conclude that a reasonable estimate of the number of opiate users directly linked to RDA opiate users in Utrecht is in the range of $500 - 550$. Compared to the estimations of the population size from the original study, the B-graph sampling design gives comparable point estimates ($500+427 = 927$; $550+427 = 977$; $490+427 = 917$), implying that the proportion of opiate users in Utrecht who are at a social distance of Step 2 from clients of the RDA (they know clients only via unregistered opiate users) is probably very small.

6. Discussion

In studies of hidden populations, sampling frames are often lacking, but sometimes the nature of the hidden trait will lead to the emergence of networks. In such research situations, Frank and Snijders (1994) proposed estimators that can be applied when one may assume an initial sample of individuals found independently of one another that resembles a random sample of the total network. Heckathorn (1997) elaborated RDS in which the recruitment of respondents is done by respondents, showing that "RDS produces samples that are independent of the initial subjects from which sampling begins" (p. 176). However, often partial sampling frames are available in studies of hidden populations. In this article, an alternative sampling design is introduced that makes use of the partial sampling frames by pooling them into one sampling frame. If this sampling frame is considered to cover a substantial part of the unknown hidden population by the local experts, one may draw a random sample of this sampling frame, asked the respondents who they know in the hidden population and estimate the number of persons who are not on the sampling frame. This proposed B-graph sampling design has some challenging features for hidden population research. First, in a lot of studies it is often interesting to know how many people with hidden activities are directly related to the registered group of known people. By random sampling from the registered population and application of the B-graph design, each member of the unknown directly related population has a chance to be in the sample. For instance, if a health organisation wants to know how many other possible "future" clients they can reach via their own clients for health education purposes, the B-graph design can be used. This way, "recruit" markets of criminal organisations, radicals, youth or street gangs, or networks of paedophiles can also be estimated. Furthermore, if capture-recapture estimates based on administrative sources are possible, a comparison can be made, revealing the size of the proportion of that part of the populations that is very difficult for institutions to reach. Another advantage of the B-graph design is that more qualitative information about the population of interest is collected, such as the quality of relations, lifestyles, and so on.

The B-graph sampling design can only be applied to populations with a network structure; the hidden activity must lead to network formation. As with capture-recapture or RDS studies, the practical problem of accurately identifying population members also remains for the B-graph design. Selected members have to disclose their relations. This is not a straightforward activity. Network members can be identified by a number of characteristics, such as the first two or three letters of first and family name, sex, age, neighbourhood, and so on. Reasons to work with identification variables are often to protect the privacy of users but also to reduce nonresponse. However, the remark of Chao et al. (2008, 957) for animal size studies also applies to human population size studies:

> "Careful sampling with proper marking (identifying) can provide more accurate estimates about the population size than an incomplete census."

7. References

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1988. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA, and London: The MIT Press.

Bogaerts, S. and A. Daalder. 2011. "Measuring Childhood Abuse and Neglect in a Group of Female Indoor Sex Workers in the Netherlands. A Confirmatory Factor Analysis of the Dutch Version of the Childhood Trauma Questionnaire-Short Form." *Psychological Reports* 108: 856–860. Doi: http://dx.doi.org/10.2466/02.10.13.16.PR0.108.3.856-860.

Böhning, D. 2010. "Some General Comparative Points on Chao's and Zelterman's Estimators of the Population Size." *Scandinavian Journal of Statistics* 37: 221–236. Doi: http://dx.doi.org/10.1111/j.1467-9469.2009.00676.x.

Böhning, D. and P. van der Heijden. 2009. "Recent Developments in Life and Social Science Applications of Capture-Recapture Methods." *AStA Advances in Statistical Analysis* 93: 1–3. Doi: http://dx.doi.org/10.1007/s10182-008-0097-7.

Brittain, S. and D. Böhning. 2009. "Estimators in Capture-Recapture Studies With Two Sources." *AStA Advances in Statistical Analysis* 93: 23–47. Doi: http://dx.doi.org/10.1007/s10182-008-0085-y.

Brugal, M.T., A. Domingo-Salvany, A. Maguire, J.A. Cayla, J.R. Villalbi, and R. Hartnoll. 1999. "A Small Area Analysis Estimating the Prevalence of Addiction to Opioids in Barcelona." *Journal of Epidemiology and Community Health* 53: 488–494. Doi: http://dx.doi.org/10.1136/jech.53.8.488.

Chao, A. 1987. "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability." *Biometrics* 43: 783–791. Doi: http://dx.doi.org/10.2307/2531532.

Chao, A. 1988. "Estimating Animal Abundance with Capture Frequency Data." *Journal of Wildlife Management* 52: 295–300. Doi: http://dx.doi.org/10.2307/3801237.

Chao, A. 1989. "Estimating Population Size for Sparse Data in Capture-Recapture Experiments." *Biometrics* 45: 427–438. Doi: http://dx.doi.org/10.2307/2531487.

Chao, A. 2001. "An Overview of Closed Capture-Recapture Models." *Journal of Agricultural, Biological, and Environmental Statistics* 6: 158–175. Doi: http://dx.doi.org/10.1198/108571101750524670.

Chao, A., H.Y. Pan, and S.C. Chiang. 2008. "The Petersen-Lincoln Estimator and its Extension to Estimate the Size of Shared Population." *Biometrical Journal* 50: 957–970. Doi: http://dx.doi.org/10.1002/bimj.200810482.

Cormack, R.M. 1989. "Log-Linear Models for Capture-Recapture." *Biometrics* 45: 395–413. Doi: http://dx.doi.org/10.2307/2531485.

Cormack, R.M. 1992. "Interval Estimation for Mark-Recapture Studies of Closed Populations." *Biometrics* 48: 567–576. Doi: http://dx.doi.org/10.2307/2532310.

Ten Den, C., B. Bieleman, E. De Bie, and J. Snippe. 1995. *Pijn in het hart*. Groningen and Rotterdam: Intraval.

Fienberg, S.E. 1972. "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables." *Biometrika* 59: 591–603. Doi: http://dx.doi.org/10.1093/biomet/59.3.591.

Frank, O. 1979. "Estimation of Population Totals by Use of Snowball Samples." In *Perspectives on Social Network Research*, edited by P.W. Holland and S. Leinhardt, 319–348. New York: Academic Press.

Frank, O. and T.A.B. Snijders. 1994. "Estimating the Size of Hidden Populations Using Snowball Sampling." *Journal of Official Statistics* 10: 53–67.

Goodman, L.A. 1961. "Snowball Sampling." *Annals of Mathematical Statistics* 32: 148–170.

Heckathorn, D.D. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems* 44: 174–199. Doi: http://dx.doi.org/10.2307/3096941.

Holland, R., R. Vivancos, V. Maskrey, J. Sadler, D. Rumball, I. Harvey, and L. Swift. 2006. "The Prevalence of Problem Drug Misuse in a Rural County of England." *Journal of Public Health* 28: 88–95. Doi: http://dx.doi.org/10.1093/pubmed/fdl009.

Jansson, I. and M. Spreen. 1998. "The Use of Local Networks in a Study of Heroin Users: Assessing Average Local Networks." *Bulletin de Méthodologie Sociologique* 59: 49–61. Doi: http://dx.doi.org/10.1177/075910639805900105.

Klovdahl, A.S. 1989. "Urban Social Networks: Some Methodological Problems and Possibilities." In *The Small World*, edited by M. Kochen, 176–210. Norwood, NJ: Ablex.

Kunst, M.J.J., F.W. Winkel, and S. Bogaerts. 2010. "Prevalence and Predictors of Posttraumatic Stress Disorder Among Victims of Violence Applying for State Compensation." *Journal of Interpersonal Violence* 2010: 1631–1654. Doi: http://dx.doi.org/10.1177/0886260509354591.

McCullough, D.R. and D.H. Hirth. 1998. "Evaluation of the Petersen-Lincoln Estimator for a White-tailed Deer Population." *Journal of Wildlife Management* 52: 534–544. Doi: http://dx.doi.org/10.2307/3801606.

Palusci, V.J., S.J. Wirtz, and T.M. Covington. 2010. "Using Capture-Recapture Methods to Better Ascertain the Incidence of Fatal Child Maltreatment." *Child Abuse & Neglect* 34: 396–402. Doi: http://dx.doi.org/10.1016/j.chiabu.2009.11.002.

Salganik, M.J. and D.D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34: 193–239. Doi: http://dx.doi.org/10.1111/j.0081-1750.2004.00152.x.

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

Seber, G.A.F. 1986. "A Review of Estimating Animal Abundance." *Biometrics* 42: 267–292. Doi: http://dx.doi.org/10.2307/2531049.

Smit, F., W. Brunenberg, and P. van der Heijden. 1996. "Het Schatten van Populatiegroottes." *Tijdschrift voor Sociale Gezondheidszorg* 74: 171–176.

Smit, F., J. Toet, and P. van der Heijden. 1997. "Estimating the Number of Opiate Users in Rotterdam Using Statistical Models for Incomplete Count Data." *In European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) Methodological Pilot Study of Local Prevalence Estimates*, Lisbon: EMCDDA.

Spreen, M. 1992. "Populations, Hidden Populations, and Link-Tracing Designs: What and Why?" *Bulletin de Méthodologie Sociologique* 36: 34–58. Doi: http://dx.doi.org/10.1177/075910639203600103.

Surjadi, B., J. van Horn, S. Bogaerts, and R. Bullens. 2010. "Internet Offending: Sexual and Non-Sexual Functions within a Dutch Sample." *Journal of Sexual Aggression* 16: 47–58. Doi: http://dx.doi.org/10.1080/13552600903470054.

Thompson, S.K. and O. Frank. 2000. "Model-Based Estimation with Link-Tracing Sampling Designs." *Survey Methodology* 26: 87–98.

Volz, E. and D.D. Heckathorn. 2008. "Probability Based Estimation Theory for Respondent Driven Sampling." *Journal of Official Statistics* 24: 79–97.

Watters, J.K. and P. Biernacki. 1989. "Targeted Sampling: Options for the Study of Hidden Populations." *Social Problems* 36: 416–430. Doi: http://dx.doi.org/10.2307/800824.

Wilson, R.M. and M.F. Collins. 1992. "Capture-Recapture Estimation with Samples of Size One Using Frequency Data." *Biometrika* 79: 543–553. Doi: http://dx.doi.org/10.1093/biomet/79.3.543.

Zelterman, D. 1988. "Robust Estimation in Truncated Discrete Distributions with Application to Capture-Recapture Experiments." *Journal of Statistical Planning and Inference* 18: 225–237. Doi: http://dx.doi.org/10.1016/0378-3758(88)90007-9.

Received September 2013
Revised November 2014
Accepted January 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 763–782, http://dx.doi.org/10.1515/JOS-2015-0044

Effects of Cluster Sizes on Variance Components in Two-Stage Sampling

Richard Valliant[1], *Jill A. Dever*[2], *and Frauke Kreuter*[3]

Determining sample sizes in multistage samples requires variance components for each stage of selection. The relative sizes of the variance components in a cluster sample are dramatically affected by how much the clusters vary in size, by the type of sample design, and by the form of estimator used. Measures of the homogeneity of survey variables within clusters are related to the variance components and affect the numbers of sample units that should be selected at each stage to achieve the desired precision levels. Measures of homogeneity can be estimated using standard software for random-effects models but the model-based intracluster correlations may need to be transformed to be appropriate for use with the sample design. We illustrate these points and implications for sample size calculation for two-stage sample designs using a realistic population derived from household surveys and the decennial census in the U.S.

Key words: Anticipated variance; measure of homogeneity; sample size calculation.

1. Introduction

Samples from finite populations are often selected in two or more stages for reasons of cost or operational necessity. For example, household samples in the U.S. may be selected through geographic areas like counties or groups of counties at the first stage, smaller areas like blocks at the second, and households at the last stage. Using multiple stages concentrates the sample in a limited number of areas, which is important when data are collected by personal interview at the respective households. In a survey of students, permission to conduct a survey may have to be obtained from school districts. Selecting districts first, then schools within sample districts, and finally a sample of students within a certain grade level within the school is operationally convenient and economical. Another example is a survey of employees in one or more business sectors, such as retail trade or services. Selecting establishments and then employees within establishments is a natural way of obtaining the sample.

Designing an efficient sample depends on estimating the contribution to the variance of an estimator associated with each stage of sampling. This involves estimating variance components for each stage that depend on the type of estimator and the types of units

[1] Universities of Michigan and Maryland – Joint Program for Survey Methodology, 1218 Lefrak Hall, College Park MD 20742 20742, U.S.A. Email: rvallian@umd.edu
[2] RTI International, Washington, District of Columbia, U.S.A. Email: jdever@rti.org
[3] University of Maryland – Joint Program for Survey Methodology, 1218 Lefrak Hall, College Park MD 20742 20742, U.S.A. Email: fkreuter@umd.edu

Acknowledgments: The authors are grateful to the editor and referees whose comments led to some important clarifications.

selected for the stage in question. This topic is covered in many standard texts on theoretical and applied sampling (Cochran 1977; Lohr 2010; Särndal et al. 1992). In the textbooks, formulae are available for the variance components for general sample designs; these formulae are usually specialized and simplified to obtain versions that facilitate sample size calculations. The relative sizes of the variance components are quite sensitive to how large the sampling units are at the different stages, how much variation there is among the sizes of the units, and the type of estimator used. Although this is implicit in the general variance-component formulae, this sensitivity is given little emphasis in most texts but can have a critical effect on calculated sample sizes and the achieved precision of estimators.

In certain applications, a survey designer has some control over the relative size of the sampling units. For example, in a household survey, extremely large metropolitan areas in the U.S., like New York or Chicago, are treated as strata and not as clusters of units. The first-stage units within such strata are groups of blocks defined by the U.S. Census Bureau for census taking and other survey data collections. Attempts are usually made to create groups by combining individual blocks so that the groups have about the same total population. In other applications, the survey designer has very little control over the units' sizes. In a school or establishment survey, the number of students or employees in each school or establishment is given. The survey must work with the existing sizes and combining these clusters further would not be meaningful.

In this article, we illustrate the effect of varying cluster sizes on design effects and measures of homogeneity within clusters for two-stage sampling. Section 2 discusses the variance-component formulae for two-stage sampling when the first-stage units are selected by either simple random sampling or probability proportional to size sampling. The effects of variation in cluster size are illustrated using an artificial, but realistic, population created using decennial census data from one county in the state of Maryland (Section 3). In Section 4, we describe how variance components from random-effects models can be used to calculate the measures of homogeneity needed for a two-stage sample. We summarize our results in the last section.

2. Background: Two-Stage Sampling

In this section, we present some background material for two-stage sampling and estimators of totals used for such designs. The units in the first stage of selection will be called primary sampling units (PSUs) or clusters. Units within PSUs are called elements and are the units for which data are collected. We use the following notation in the subsequent formulae:

$U =$ universe of PSUs
$M =$ number of PSUs in the universe
$U_i =$ universe of elements in PSU i
$N_i =$ number of elements in the population for PSU i
$N = \sum_{i \in U} N_i$, the total number of elements in the population
$\bar{N} = N/M$, the average number of elements per PSU
$m =$ number of sample PSUs
$n_i =$ number of sample elements in PSU i

s = set of sample PSUs
s_i = set of sample elements in PSU i
π_i = selection probability of PSU i
$\pi_{k|i}$ = selection probability of element k given PSU i was selected
y_k = value of a variable Y observed for element k
$\bar{y}_U = \sum_{i \in U} \sum_{k \in U_i} y_k / N$, the mean per element in the population
$\bar{y}_{Ui} = \sum_{k \in U_i} y_k / N_i$, the mean per element in the population in PSU i
$S_U^2 = \sum_{i \in U} \sum_{k \in U_i} (y_k - \bar{y}_U)^2 / (N - 1)$, the population variance of Y
$t_i = \sum_{k \in U_i} y_k$, the universe total of Y in PSU i
$t_U = \sum_{i \in U} t_i$, the universe total
$\bar{t}_U = t_U / M$, the average PSU total.

The π-estimator of a population total weights the value for element k inversely by its selection probability, π_k. Särndal et al. (1992, Result 4.3.1) give a formula for the variance of the π-estimator for a very general two-stage sample design. However, the general formula is not useful for designing samples because it involves joint selection probabilities of units at each stage that do not explicitly involve sample sizes. In this section, we present the variance formulae for different two-stage sample designs where the variance of the estimated total is simple enough for use in sample size calculation. In the first, PSUs are selected by simple random sampling; in the second, PSUs are selected with varying probabilities. For both designs, we assume that elements within PSUs are selected by simple random sampling. We follow the discussions of the π-estimator and probability with replacement (*pwr*) estimator of a total in Subsections 2.1 and 2.2 with the ratio estimator of a total in Subsection 2.3.

2.1. Equal-Probability Sampling at Both Stages

Suppose the first stage is a simple random sample selected without replacement (*srswor*) of m PSUs from a population of M PSUs, and the second stage is a sample of n_i elements selected by *srswor* from the population of N_i. As a shorthand, denote this design by *srs/srs*. The selection probability of element k in PSU i is $\pi_k = \pi_i \pi_{k|i} = (m/M)(n_i/N_i)$. The π-estimator of a population total is

$$\hat{t}_\pi = \frac{M}{m} \sum_{i \in s} \frac{N_i}{n_i} \sum_{k \in s_i} y_k = \frac{M}{m} \sum_{i \in s} \hat{t}_i \tag{1}$$

where $\hat{t}_i = (N_i/n_i) \sum_{k \in s_i} y_k$, the estimate of the total for PSU i with a simple random sample. The design variance, that is, the variance computed with respect to repeated sampling, of the π-estimator is

$$V(\hat{t}_\pi) = \frac{M^2}{m} \frac{M-m}{M} S_{U1}^2 + \frac{M}{m} \sum_{i \in U} \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i} S_{U2i}^2 \tag{2}$$

where $S_{U1}^2 = \frac{\sum_{i \in U} (t_i - \bar{t}_U)^2}{M-1}$, and $S_{U2i}^2 = \frac{\sum_{k \in U_i} (y_k - \bar{y}_{Ui})^2}{N_i - 1}$, the unit variance of Y among the elements in PSU i.

The first component of (2), the "between" term, can also be written as a function of the variance among means per element within the PSUs. However, expressing the between term as a function of PSU totals as shown above allows a more intuitive explanation to be given for some subsequent results.

The relative variance (relvariance) of $\hat{t}_\pi$ is its variance divided by the square of the population total, $V(\hat{t}_\pi)/t_U^2$, and is especially useful for sample size calculation since the relvariance is unaffected by the scale of y. If the same number of sample elements, $n_i = \bar{n}$, is selected from each PSU, and the first-stage sampling fraction, m/M, and the second-stage sampling fraction, $\bar{n}/N_i$, are both small, the relvariance can be written as

$$\frac{V(\hat{t}_\pi)}{t_U^2} = \frac{B^2}{m} + \frac{W^2}{m\bar{n}} \tag{3}$$

where $B^2 = S_{U1}^2/\bar{t}_U^2$ is the unit relvariance among PSU totals and $W^2 = M^{-1}\sum_{i\in U}\left(\frac{N_i}{\bar{N}}\right)^2 \frac{S_{U2i}^2}{\bar{y}_U^2}$. A common simplification used in Cochran (1977) and Hansen et al. (1953a) is to further assume that all PSUs contain the same number of elements, that is, $N_i \equiv \bar{N}$, so that $W^2 = M^{-1}\sum_{i\in U} S_{U2i}^2/\bar{y}_U^2$. Roughly speaking, W^2 is an average relvariance per PSU with the per-PSU relvariance expressed as $S_{U2i}^2/\bar{y}_U^2$, that is, with the overall mean in the denominator. Expression (3) can be rearranged to give

$$\frac{V(\hat{t}_\pi)}{t_U^2} = \frac{\tilde{V}}{m\bar{n}} k[1 + \delta(\bar{n} - 1)] \tag{4}$$

where $\tilde{V} = S_U^2/\bar{y}_U^2$, $k = (B^2 + W^2)/\tilde{V}$, and $\delta = B^2/(B^2 + W^2)$, often referred to as a *measure of homogeneity*. With single-stage *srs* sampling of clusters from a population in which all clusters have the same size $\bar{N}$, δ is an *intraclass correlation* (see Cochran 1977, ch. 9; Lohr 2010 sec. 5.2.2) that can be computed as a type of Pearson correlation. With two-stage sampling, however, δ is not a correlation but still is related to the degree of homogeneity of elements within clusters. Note that an *fpc*, $1 - m\bar{n}/M\bar{N}$, is sometimes inserted into Expression (4) if the sampling fractions are not small, but this is an *ad hoc* addition that does not follow directly from rewriting (3).

The formula found in most textbooks is Expression (4) with $k = 1$, which comes from first writing the population variance of y as

$$(M\bar{N} - 1)S_U^2 = \sum_{i\in U} N_i\left(\frac{t_i}{Ni} - \frac{t_U}{M\bar{N}}\right)^2 + \sum_{i\in U}(N_i - 1)S_{U2i}^2.$$

Then, with some algebra (see Hansen et al. 1953a, sec. 6.6; Hansen et al. 1953b, sec. 6.5), it can be shown that when all clusters have the same size, $\bar{N}$, and both M and $\bar{N}$ are large,

$$\frac{S_U^2}{\bar{y}_U^2} = \frac{1 - M^{-1}}{1 - (M\bar{N})^{-1}}B^2 + \frac{1 - \bar{N}^{-1}}{1 - (M\bar{N})^{-1}}W^2 \quad \doteq B^2 + W^2 \tag{5}$$

that is, $k = 1$. In that case, (4) reduces to the relvariance of the estimated total in *srs*, $\tilde{V}/m\bar{n}$, times a design effect, $1 + \delta\ (\bar{n} - 1)$. The design-effect concept has been extended to more complex situations by Gabler et al. (1999), Lynn and Gabler (2005), and Park and Lee (2004).

The assumptions to obtain (5) that the number of population clusters and number of population elements per cluster are large is often reasonable, but assuming that the clusters all have the same size ($N_i = \bar{N}$) may not be. Although this special case is emphasized in texts like Kish (1965) and Lohr (2010), it can be misleading when clusters vary in size.

An alternative design for the second stage is to select elements at a fixed rate r within each cluster. The expected sample size in cluster i then is $n_i = rN_i$. This design might be preferred to *srs*/*srs* with a fixed-size sample at the second stage because all sample elements will have the same weight, $M/(mr)$. There are different ways of selecting such a sample. Bernoulli sampling is one; systematic sampling from a randomly ordered list is another. In the latter design, which we use here, the achieved sample size is either the integer floor or ceiling of rN_i. This type of systematic sample can reasonably be treated as *srswor* when the list is randomly ordered. Substituting $n_i = rN_i$ in (2), dividing by t_U^2, and using the equivalent expressions for the population total, $t_U = M\bar{t}_U = M\bar{N}\bar{y}_U$, gives the approximate relvariance as

$$\frac{V(\hat{t}_\pi)}{t_U^2} = \frac{B^2}{m} + \frac{\tilde{W}^2}{m\bar{n}^*} \tag{6}$$

where B^2 is the same quantity as in (3), $\bar{n}^* = r\bar{N}$, and $\tilde{W}^2 = M^{-1}\sum_{i\in U}\frac{N_i}{\bar{N}}\frac{S_{U2i}^2}{\bar{y}_U^2}$. There is some randomness in the achieved second-stage sample size when rN_i is not an integer. Note that $\bar{n}^*$ is an average cluster sample size in the sense that the average sample size over all clusters in the universe is $\sum_{i\in U} n_i/M = r\bar{N}$. The corresponding value of the measure of homogeneity is $\tilde{\delta} = B^2/(B^2 + \tilde{W}^2)$. The relvariance in (6) can also be written as

$$\frac{V(\hat{t}_\pi)}{t_U^2} = \frac{\tilde{V}}{m\bar{n}^*}\tilde{k}[1 + \tilde{\delta}(\bar{n}^* - 1)] \tag{7}$$

where $\tilde{k} = (B^2 + \tilde{W}^2)/\tilde{V}$. Note that (7) reduces to the usual textbook formula if $\tilde{k} = 1$, which requires that $S_U^2/\bar{y}_U^2 \doteq B^2 + \tilde{W}^2$. Since the design with a fixed sampling rate at the second stage may be more common in practice than one with a common $\bar{n}$ when the design is *srs/srs*, we concentrate on it in the numerical illustrations.

Expressions (4) or (7) are useful for sample size calculation since the number of sample PSUs, m, sample elements per PSU, $\bar{n}$, or the within-PSU rate, $r = \bar{n}^*/\bar{N}$, are explicit in the formula. Expressions like (4) and (7) often seem to be treated as if they apply regardless of how the samples of PSUs and elements within PSUs are selected. If, for example, a probability proportional to size (*pps*) sample of PSUs is selected, (4) and (7) do not reflect that feature. In Subsection 2.2 we therefore give a relvariance that is similar in form to (4) and (7) but is appropriate for *pps* sampling of PSUs.

When designing samples, practitioners sometimes use rough rules of thumb for values of $\tilde{\delta}$ (or δ), say $\tilde{\delta} \leq 0.10$, based on how "alike" elements within PSUs are thought to be. However, the form of S_{U1}^2 and, therefore, B^2 implies that the size of $\tilde{\delta}$ (or δ) can also be determined by the relative variability of the cluster totals, t_i. As we will illustrate, one way in which $\tilde{\delta}$ can be large is by having clusters that vary in size.

2.2. *Varying Probabilities at the First Stage*

Variances of estimators in designs more complicated than simple random sampling at each stage can also be written as a sum of components. However, the most general of these have limited value in determining sample sizes (e.g., see Särndal et al. 1992, result 4.3.1).

A more useful formulation is the case where PSUs are selected with varying probabilities but with replacement (*ppswr*), and the sample within each PSU is selected by *srswor*. We refer to this design as *ppswr/srs*. With-replacement designs may not often be used in practice but have simple variance formulae, which makes them useful for sample size calculation. The probability with-replacement (*pwr*) estimator of a total is

$$\hat{t}_{pwr} = \frac{1}{m} \sum_{i \in s} \frac{\hat{t}_i}{p_i}$$

where $\hat{t}_i$ was defined in Subsection 2.1 and p_i is the one-draw selection probability of PSU i. The variance of $\hat{t}_{pwr}$ is (Cochran 1977, 308-310)

$$V(\hat{t}_{pwr}) = \frac{1}{m} \sum_{i \in U} p_i \left(\frac{t_i}{p_i} - t_U \right)^2 + \sum_{i \in U} \frac{{N_i}^2}{m p_i n_i} \left(1 - \frac{n_i}{N_i} \right) S^2_{U2i}. \tag{8}$$

Making the assumption that $\bar{n}$ elements are selected in each PSU and that $\bar{n}/N_i$ is negligible, the variance reduces to

$$V(\hat{t}_{pwr}) = \frac{S^2_{U1(pwr)}}{m} + \frac{1}{m\bar{n}} \sum_{i \in U} \frac{N_i^2 S^2_{U2i}}{p_i}$$

where, in this case, $S^2_{U1(pwr)} = \sum_{i \in U} p_i \left(\frac{t_i}{p_i} - t_U \right)^2$ and S^2_{U2i} is defined for Expression (2). Dividing this by t_U^2 and simplifying, we obtain the relvariance of $\hat{t}_{pwr}$ as, approximately,

$$\frac{V(\hat{t}_{pwr})}{t_U^2} \doteq \frac{B_*^2}{m} + \frac{W_*^2}{m\bar{n}} = \frac{\tilde{V}}{m\bar{n}} k_* [1 + \delta_* (\bar{n} - 1)] \tag{9}$$

with $B_*^2 = \frac{S^2_{U1(pwr)}}{t_U^2}$, $W_*^2 = \frac{1}{t_U^2} \sum_{i \in U} N_i^2 \frac{S^2_{U2i}}{p_i}$, $k_* = \left(B_*^2 + W_*^2 \right) / \tilde{V}$, and $\delta_* = B_*^2 / \left(B_*^2 + W_*^2 \right)$. If $k_* = 1$, then (9) has the interpretation of an *srs* relvariance times a design effect, $1 + \delta_* (\bar{n} - 1)$.

The approximation in (9) does depend on the sampling fraction of elements within each sample cluster being small, and more importantly on using the with-replacement variance formula for the first stage. On the other hand, it does allow the number of population elements per cluster to vary, which is an important feature to account for in some populations.

A special case of the design above would be $p_i = N_i / N$, that is, probability proportional to the size of cluster i. If the weight of cluster i in a with-replacement sample of m clusters is $N/(mN_i)$ and an equal-probability sample of $\bar{n}$ elements are selected in each cluster, the sample is "self-weighting" as the weight of each sample element in the *pwr* estimator is the same: $(N/mN_i)(N_i/\bar{n}) = N/(m\bar{n})$. This combination of design and weighting method is common in household surveys where a practical goal is often to have an equal workload in each cluster and limit variation in weights.

A more general point to note is that the measures of homogeneity in (4), (7), and (9) depend on both the sample design and the estimator being used. This is because the decomposition of the variance of an estimator depends on both. A different decomposition would be needed for, say, the general regression (GREG) estimator of a total or an estimator of a mean that uses an estimate of N in its denominator.

2.3. *Ratio Estimator of a Total*

The π-estimator of a total may be inefficient in some designs compared to alternatives like the ratio estimator or a GREG estimator. In this section, we present the variance-component formula in the *srs/srs* design for the ratio estimator defined as

$$\hat{t}_R = \hat{t}_\pi \frac{N}{\hat{N}_\pi}$$

where $\hat{N}_\pi$ is the π-estimator of the number of elements in the population, N, defined as $\hat{N}_\pi = M\sum_s N_i/m$. (Note that, in probability proportional to N_i sampling with $p_i = N_i/N$, the estimated total number of elements is $\hat{N}_{pwr} = m^{-1}\sum_{i\in s}\hat{t}_i/p_i = N$, and there is no gain from ratio estimation.) Assuming that the sample size of clusters m is large and using a first-order linear approximation,

$$\hat{t}_R - t_U = \hat{t}_\pi - \bar{y}_U\hat{N}_\pi + O_p(M/m) \doteq \frac{M}{m}\sum_s \hat{t}_{zi} \qquad (10)$$

where $\hat{t}_{zi} = N_i\sum_{k\in s_i} z_k/n_i$ with $z_k = y_k - \bar{y}_U$. Expression (10) follows from assuming that $M^{-1}(\hat{t}_\pi - t_U)$ and $M^{-1}(\hat{N}_\pi - N)$ are $O_p(m^{-1/2})$ as they would be if $m^{1/2}(\hat{t}_\pi - t_U)/M$ and $m^{1/2}(\hat{N}_\pi - N)/M$ had asymptotic standard normal distributions. In that case the remainder in the first-order Taylor series approximation to $M^{-1}(\hat{t}_R - t_U)$ is $O_p([m^{-1/2}]^2) = O_p(m^{-1})$ (see Wolter 2007, Theorem 6.2.2). Under those conditions, $\hat{t}_\pi - \bar{y}_U\hat{N}_\pi = O_p(M/m^{1/2})$, that is, a higher order than the remainder term in (10). Approximation (10) has the same form as the π-estimator in (1). Consequently, a variance-component formula analogous to (2) and a relvariance formula similar to (3) can be derived. In particular,

$$V(\hat{t}_R) \doteq \frac{M^2}{m}\frac{M-m}{M}S^2_{Uz1} + \frac{M}{m}\sum_{i\in U}\frac{N_i^2}{n_i}\frac{N_i-n_i}{N_i}S^2_{U2zi}$$

with $S^2_{Uz1} = (M-1)^{-1}\sum_{i\in U}(t_{zi} - \bar{t}_{Uz})^2$, and $S^2_{U2zi} = (N_i-1)^{-1}\sum_{k\in U_i}(z_k - \bar{z}_{Ui})^2$ where $t_{zi} = \sum_{k\in U_i} z_k$, $\bar{t}_{Uz} = \sum_{i\in U} t_{zi}/M$, and $\bar{z}_{Ui} = \sum_{k\in U_i} z_k/N_i$. Assuming that the *fpc*s, $(M-m)/M$ and $(N_i - n_i)/N_i$, are approximately 1 and that the sample size in PSU i is rN_i, the relvariance formula is

$$V(\hat{t}_R) \doteq \frac{B_z^2}{m} + \frac{\tilde{W}_z^2}{m\bar{n}^*} = \frac{\tilde{V}}{m\bar{n}^*}k_z[1 + \delta_z(\bar{n}^* - 1)] \qquad (11)$$

where $B_z^2 = S^2_{Uz1}/\bar{t}^2_U$, $\tilde{W}_z^2 = M^{-1}\sum_{i\in U}(N_i/\bar{N})S^2_{U2zi}/\bar{y}^2_U$, $k_z = \left(B_z^2 + \tilde{W}_z^2\right)/\tilde{V}$, and $\delta_z = B_z^2/\left(B_z^2 + \tilde{W}_z^2\right)$. Compared to the (*srs/srs*, π-estimator) strategy the ratio estimator can reduce the measure of homogeneity, leading to more precise estimators as illustrated in Example 4 of Section 3.

3. Examples of Variance Components and Measures of Homogeneity

We created an example population based on U.S. Census counts from the year 2000 for Anne Arundel County in the state of Maryland and refer to this data set as `MDarea.pop`. The population is also included in the R package `PracTools` (Valliant et al. 2013, 2015). The population contains three continuous and two binary variables denoted by `y1, y2, y3, ins.cov,` and `hosp.stay`, respectively. The variables are generated using models, since individual-person data for small geographic areas is suppressed in the actual census for reasons of confidentiality. The variables in `MDarea.pop` were created by fitting models for several variables in the 2001-2002 National Health and Nutrition

Table 1. Descriptive statistics for the Maryland area population

	Tract population	BG population	y1	y2	y3
Minimum	86	52	− 62.7	− 2.9	32.6
1st quartile	2,728	780	18.7	2.3	66.7
Median	4,132	1,240	50.8	5.4	81.4
Mean	4,253	1,316	69.7	7.7	87.5
3rd quartile	5,684	1,732	104.4	10.7	101.2
Maximum	13,579	4,744	1163.7	101.1	479.2
Population CV	0.51	0.58	1.21	1.01	0.34

CV = coefficient of variation.

Examination Survey (Center for Disease Control and Prevention 2009) and 2003 National Health Interview Survey (Center for Disease Control and Prevention 2012) data sets to obtain regression means that depended on whether a person was Hispanic and on the person's gender and age. Person-level values were created using random-effects models that had error terms for tracts, block groups, and persons. The three continuous variables (y1, y2, y3) are positively skewed with mean values based on models for body weight, body mass index, and systolic blood pressure (although the scales of the generated variables do not match those of these physical measurements). The binary variables, ins.cov and hosp.stay, are based on the rates of insurance coverage and overnight hospital stay in a twelve-month period.

The geographic divisions used in this data set are tracts and block groups, which are geographic areas defined by the Census Bureau (U.S. Census Bureau 2011). Tracts are constructed to have a desired population size of 4,000 people. Block groups (BGs) are smaller, with a target size of 1,500 people. However, the sizes of both tracts and BGs vary because the Census Bureau also attempts to limit the geographic area covered by a BG. Counts of persons in the data set are the same for most tracts and BGs as in the 2000 Census; exceptions are five BGs that were augmented to have at least 50 persons each.

The example population contains 403,997 persons, 95 tracts, and 307 BGs. The proportion of persons with insurance coverage is 0.793; the proportion with a hospital stay in the prior twelve months is 0.072. Descriptive statistics for other variables are given in Table 1.

Because the tracts and BGs in the Maryland population are extremely variable in size, we created two other variables called PSU and SSU to demonstrate the effect of having equal-sized units. Each artificial PSU has approximately the same number of persons; likewise the SSUs were created to have about the same number of persons. The PSUs and SSUs were formed after sorting the file by tract and BG within tract, thus retaining geographic proximity of persons grouped together. Each PSU has about 5,050 persons while an SSU has about 1,010. Although the assumption of equal PSU size made to obtain (5) or to set $\tilde{k} = 1$ may seem innocuous, it is far from that, as we will illustrate below.

We use the Maryland population to illustrate the effects of using different sizes of primary and secondary sampling units on the measures of homogeneity for two-stage sampling. In all of the examples, calculations are made assuming that the entire population is in hand. This means that the theoretical values in the preceding formulae can be evaluated rather than estimated from a sample as would be required in practice.

When examining the effects of varying unit sizes, working with a population is an advantage as the complication of sampling variability is eliminated.

Example 1. Between- and within-variance components in srs/srs design. Using the variables in the Maryland population, we computed the unit relvariance of each variable $\left(S_U^2/\bar{y}_U^2\right)$, $B^2+\tilde{W}^2$ and $\tilde{k}$ for comparison, and $\tilde{\delta}=B^2/(B^2+\tilde{W}^2)$ for the *srs/srs* design and the *pwr*-estimator. (Note that the π-estimator and *pwr*-estimator in *srs/srs* have the same form when the first stage is selected with replacement. In the examples, we will refer to the (*srs/srs*, *pwr*-strategy).) First, the results are shown in Table 2 using the PSU and SSU variables as clusters. Values of $\tilde{\delta}$ range from 0.001 to 0.079 when PSUs are clusters. Deltas are somewhat larger when SSUs are clusters, reflecting the common phenomenon that smaller geographic areas are somewhat more homogeneous than large ones in household populations. The third through fifth columns show that the approximation that $S_U^2/\bar{y}_U^2 \doteq B^2+W^2$ works well in this case.

Next, to illustrate the dramatic effect that varying sizes of clusters can have, in Table 3 we present the same statistics as above using tracts and BGs within tracts as clusters. Values of δ range from 0.023 to 0.730 when tracts are clusters. When BGs are used as clusters, $\tilde{\delta}$s range from 0.032 to 0.791. The measures of homogeneity increase substantially when tracts or BGs are the first-stage clusters. For example, when PSUs are clusters, $\delta=0.005$ for y1 but is 0.152 when tracts are clusters. This is almost entirely due to the increase in the between-variance component, B^2, when units with highly variable sizes are used. For example, $B^2=0.0079$ for y1 when PSU is a cluster, but is 0.2605 when tract is a cluster. The third through fifth columns in Table 3 show that the approximation $S_U^2/\bar{y}_U^2 \doteq B^2+\tilde{W}^2$ does not work well when either tracts or BGs are clusters. This again is due to the clusters not all having the same size. This implies that when making advance estimates of the relvariance of an estimated total, $\tilde{k}$ cannot be safely set to 1 in (7) when PSUs vary in size.

Example 2. **Effect of incorrect measures of homogeneity on achieved precision.** If incorrect values of the measure of homogeneity are used to compute sample sizes, the sample can be much less efficient than anticipated. This example looks at the effect of

Table 2. Variance components and measures of homogeneity in the Maryland population using PSUs *and* SSUs *as clusters with an srs/srs design, the pwr-estimator, and a fixed sampling rate at the second stage*

	B^2	$\tilde{W}^2$	$S_U^2/\bar{y}_U^2$	$B^2+\tilde{W}^2$	$\tilde{k}$	$\tilde{\delta}$
PSUs as clusters						
y1	0.0079	1.4553	1.4627	1.4631	1.0003	0.005
y2	0.0069	1.0097	1.0163	1.0166	1.0003	0.007
y3	0.0090	0.1048	0.1136	0.1137	1.0012	0.079
ins.cov	0.0012	0.2599	0.2611	0.2611	1.0003	0.005
hosp.stay	0.0175	12.8831	12.8979	12.9006	1.0002	0.001
SSUs as clusters						
y1	0.0365	1.4277	1.4627	1.4642	1.0010	0.025
y2	0.0169	1.0004	1.0163	1.0173	1.0010	0.017
y3	0.0184	0.0954	0.1136	0.1137	1.0012	0.161
ins.cov	0.0032	0.2581	0.2611	0.2613	1.0010	0.012
hosp.stay	0.0558	12.8549	12.8979	12.9107	1.0010	0.004

Table 3. Variance components and measures of homogeneity in the Maryland population using tracts and block groups as clusters with an srs/srs design, the pwr-estimator, and a fixed sampling rate at the second stage

	B^2	$\tilde{W}^2$	$S_U^2/\bar{y}_U^2$	$B^2+\tilde{W}^2$	$\tilde{k}$	$\tilde{\delta}$
Tracts as clusters						
`y1`	0.2605	1.4539	1.4627	1.7144	1.1720	0.152
`y2`	0.2687	1.0058	1.0163	1.2745	1.2540	0.211
`y3`	0.2707	0.1001	0.1136	0.3707	3.2634	0.730
`ins.cov`	0.2624	0.2593	0.2611	0.5217	1.9985	0.503
`hosp.stay`	0.3078	12.8786	12.8979	13.1864	1.0224	0.023
Block groups as clusters						
`y1`	0.3489	1.4478	1.4627	1.7967	1.2283	0.194
`y2`	0.3485	0.9994	1.0163	1.3479	1.3263	0.259
`y3`	0.3492	0.0926	0.1136	0.4418	3.8887	0.791
`ins.cov`	0.3408	0.2574	0.2611	0.5982	2.2916	0.570
`hosp.stay`	0.4246	12.8567	12.8979	13.2813	1.0297	0.032

using $\tilde{\delta}$s computed as if clusters all had the same size when clusters actually vary. Suppose that the costs which vary with the number of sample clusters and elements can be written as $C = C_1 m + C_2 m\bar{n}$ where C_1 is the cost per cluster and C_2 is the cost per sample element. If the budget for variable costs is fixed at C and the relvariance is given by (7), the optimal numbers of elements and clusters are (cf. Hansen et al. 1953a sec. 16.6):

$$\bar{n}_{opt} = \sqrt{\frac{C_1}{C_2}\frac{1-\tilde{\delta}}{\tilde{\delta}}} \quad \text{and} \quad m_{opt} = \frac{C}{C_1 + C_2\bar{n}_{opt}}. \tag{12}$$

(The results in (12) hold for both $\tilde{k} = 1$ and a general value of $\tilde{k}$.) In this example, the cost assumptions are $C = \$100{,}000$, $C_1 = \$1,000$, and $C_2 = \$100$. Suppose that the sample sizes in (12) are computed using the $\tilde{\delta}$s in Table 2, assuming that clusters are `PSUs` or `SSUs` (i.e., clusters with the same size). These values of $\bar{n}_{opt}$ and m_{opt} are shown in Table 4 using the $\tilde{\delta}$s and values of $\tilde{k}$ from Table 2. The estimated coefficients of variation (CVs), that is, the square root of the estimated relvariances that would be obtained with the equal-size cluster $\tilde{\delta}$s, are in the fourth column, assuming that $\tilde{V} = 1$. Suppose that the correct $\tilde{\delta}$s and $\tilde{k}$s are in reality those in Table 3, which account for varying cluster sizes. The actual CVs that would be obtained with these $\tilde{\delta}$s are also shown in the sixth column of Table 4, again assuming that $\tilde{V} = 1$. The ratio of actual CVs with $\tilde{\delta}$s from Table 3 to the estimated CVs with $\tilde{\delta}$s from Table 2 range from 1.5 to 6.3. In other words, the actual CVs range from 50% to 530% higher than estimated because varying cluster sizes increase the measures of homogeneity and values of $\tilde{k}$. This implies that if the correct $\tilde{\delta}$s and $\tilde{k}$s were used, more clusters and fewer elements per cluster should be selected than the m_{opt} and $\bar{n}_{opt}$ values in Table 4.

Example 3. ***ppswr* at first stage, *srs* at second**. This example repeats the calculations in Example 1 for the variables in the Maryland area population but with a different sample design. Assume that clusters will be selected proportional to the count of persons N_i in each cluster and that an *srs* with a small sampling fraction is selected in each sample cluster, that is, a particular case of *ppswr/srs*. Table 5 shows the values of B_*^2, W_*^2, and δ_*

Table 4. Loss of precision from using incorrect measures of homogeneity with an srs/srs design, the pwr-estimator, and a fixed sampling rate at the second stage

	$\tilde{\delta}$ with equal-size clusters (Table 2)	$\bar{n}_{opt}$	m_{opt}	Estimated CV (%)	$\tilde{\delta}$ with varying-size clusters (Table 3)	Actual CV (%)	Ratio: Actual to estimated CVs
Tracts as clusters							
`y1`	0.005	43	19	3.9	0.123	10.3	2.7
`y2`	0.007	38	21	4.0	0.173	11.8	3.0
`y3`	0.079	11	48	5.8	0.681	22.6	3.9
`ins.cov`	0.005	47	18	3.8	0.443	24.1	6.3
`hosp.stay`	0.001	84	11	3.5	0.018	5.8	1.6
Block groups as clusters							
`y1`	0.024	20	33	4.7	0.151	9.3	2.0
`y2`	0.016	25	29	4.4	0.206	11.5	2.6
`y3`	0.160	7	58	6.9	0.740	23.4	3.4
`ins.cov`	0.011	30	25	4.3	0.498	22.6	5.3
`hosp.stay`	0.003	58	15	3.8	0.023	5.6	1.5

Table 5. Variance components and measures of homogeneity in the Maryland population using `PSUs` *and* `SSUs` *as clusters with a ppswr/srs design and the pwr-estimator*

	B_*^2	W_*^2	k_*	δ_*
`PSUs` as clusters				
`y1`	0.0078	1.4553	1.0002	0.005
`y2`	0.0068	1.0097	1.0002	0.007
`y3`	0.0088	0.1048	1.0002	0.078
`ins.cov`	0.0012	0.2599	1.0002	0.005
`hosp.stay`	0.0173	12.8831	1.0002	0.001
`SSUs` as clusters				
`y1`	0.0364	1.4277	1.0010	0.025
`y2`	0.0169	1.0004	1.0010	0.017
`y3`	0.0183	0.0954	1.0008	0.161
`ins.cov`	0.0032	0.2581	1.0010	0.012
`hosp.stay`	0.0557	12.8549	1.0010	0.004

when PSUs and SSUs are clusters. Because each PSU and SSU was formed to have almost the same number of persons, the values in Table 5 are virtually the same as the *srs/srs* results in Table 2.

Table 6 shows the results when tracts and BGs are used as clusters. With the *ppswr/srs* design, the between term is much smaller than the within term compared to the results in Example 1. This is true whether `PSU` and `SSU` are used as clusters or tracts and BGs are used. For example, with `y1`, $\delta = 0.152$ when tracts are clusters in the *srs/srs* design (Table 3). However, $\delta_* = 0.006$ for `y1` with tracts as clusters in the *ppswr/srs* design in Table 6. The measures of homogeneity for other variables are also substantially less in Table 6 than in Table 3.

When clusters are selected by *srs*, S_{U1}^2 is the variance of the cluster totals around the average cluster total. In contrast, with *pps* sampling of clusters, $S_{U1(pwr)}^2$ is the variance of the estimated population totals, t_i/p_i around the population total, t_U. When clusters are selected with probability proportional to N_i, $t_i/p_i = N_i\bar{y}_{Ui}/(N_i/N) = N\bar{y}_{Ui}$. If these

Table 6. Variance components and measures of homogeneity in the Maryland population using tracts and BGs as clusters with a ppswr/srs design and the pwr-estimator

	B_*^2	W_*^2	k_*	δ_*
Tracts as clusters				
`y1`	0.0092	1.4539	1.0002	0.006
`y2`	0.0107	1.0058	1.0002	0.011
`y3`	0.0136	0.1001	1.0002	0.119
`ins.cov`	0.0018	0.2593	1.0002	0.007
`hosp.stay`	0.0223	12.8786	1.0002	0.002
Block groups as clusters				
`y1`	0.0160	1.4478	1.0007	0.011
`y2`	0.0176	0.9994	1.0007	0.017
`y3`	0.0211	0.0926	1.0006	0.186
`ins.cov`	0.0039	0.2574	1.0007	0.015
`hosp.stay`	0.0509	12.8567	1.0008	0.004

Table 7. Variance components and measures of homogeneity in the Maryland population using tracts and block groups as clusters with an srs/srs design, a fixed rate at the second stage, and a ratio estimator of a total

	B_z^2	$\tilde{W}_z^2$	k_z	δ_z
Tracts as clusters				
`y1`	0.0093	1.8390	1.2636	0.005
`y2`	0.0114	1.2662	1.2571	0.009
`y3`	0.0143	0.1253	1.2285	0.102
`ins.cov`	0.0021	0.3260	1.2568	0.007
`hosp.stay`	0.0265	16.3171	1.2672	0.002
Block groups as clusters				
`y1`	0.0193	1.9499	1.3462	0.010
`y2`	0.0223	1.3338	1.3344	0.017
`y3`	0.0271	0.1220	1.3127	0.182
`ins.cov`	0.0052	0.3426	1.3324	0.015
`hosp.stay`	0.0681	17.2695	1.3442	0.004

1-cluster estimates of the population total are fairly accurate, as they are here, the B^2 term can be quite small. This leads to much smaller values of the measure of homogeneity in *pps* sampling of clusters, implying that the effect of clustering is less important in this population for a design that selects clusters with probabilities proportional to their population counts.

Practitioners habitually gravitate toward *pps* sampling of clusters rather than *srs*. This example makes it clear why this choice is often a good one.

Example 4. *srs/srs* design with ratio estimator of the total. Next, we consider whether use of the ratio estimator of the total in an *srs/srs* design reduces the effects of using clusters with varying sizes. Table 7 displays results for the variance components, B_z^2 and $\tilde{W}_z^2$, k_z, and δ_z defined in Subsection 2.3 when tracts or block groups are used as clusters. The values of δ_z in Table 7 are much lower than those of $\tilde{\delta}$ in Table 3, implying that use of a ratio estimator in *srs/srs* substantially reduces the effect of clustering compared to using the *pwr*-estimator. The values of δ_z are very close to those of δ_* in Table 6 for the *ppswr/srs* design combined with the *pwr*-estimator. However, the values of k_* in Table 6 are all near 1 while k_z in Table 7 ranges from about 1.23 to 1.35. Thus, for a given number of sample clusters m and elements $\bar{n}$ in the *ppswr/srs* case or $\bar{n}^*$ in the *srs/srs* fixed rate case, the (*ppswr/srs*, *pwr*-estimator) strategy will be more efficient than the (*srs/srs*, ratio estimator) strategy. For example, suppose that BGs are clusters, the total of `y1` is estimated and $\bar{n} = \bar{n}^* = 50$. If *srs/srs* and the *pwr*-estimator is used, then $\tilde{k}[1 + \tilde{\delta}\ (\bar{n} - 1)] = 1.2283[1 + 0.194(50 - 1)] = 12.905$ using the figures in Table 3. For the (*ppswr/srs*, *pwr*-estimator) strategy, $k_*[1 + \delta_*(\bar{n} - 1)] = 1.007[1 + 0.011(50 - 1)] = 1.550$ using the values in Table 6. For (*srs/srs*, ratio estimator) the corresponding value is $k_z[1 + \delta_z (\bar{n}^* - 1)] = 1.3462[1 + 0.010(50 - 1)] = 2.006$ using the figures in Table 7. Accordingly, the relvariance for (*srs/srs*, *pwr*-estimator) is 8.33 (12.905/1.550) times as large as that of (*ppswr/srs*, *pwr*-estimator), while the relvariance of (*srs/srs*, ratio estimator) is 1.29 (2.006/1.55) times as large. Using the ratio estimator in *srs/srs* is much better than using the *pwr*-estimator, but still is considerably less efficient than the (*ppswr/srs*, *pwr*-estimator) strategy.

4. Estimating Variance Components Using Anticipated Variances

In normal circumstances, only a sample is available from a population and variance components must be estimated. Design-based estimators can be found in Särndal et al. (1992, sec. 4.3.2) and will not be covered here. As noted earlier, the general formulae for estimation of variance components are specialized, complex, and difficult to use in practice. Being able to use the software routines that are available for variance-component estimation would be a real advantage if they estimate the components properly. The best of these routines use algorithms designed to handle a variety of numerical problems that are hard to anticipate in practice. Searle et al. (1992) review the methods available, including minimum variance quadratic unbiased estimation (MIVQUE0), maximum likelihood, and restricted maximum likelihood (REML). Note that these estimates are derived through a specified model and not a particular sample design.

Model variance components can be introduced by using an anticipated variance (Isaki and Fuller 1982) defined as

$$AV(\hat{t}) = E_M[E_\pi(\hat{t} - t_U)^2] - [E_M E_\pi(\hat{t} - t_U)]^2$$

where E_M is the theoretical expectation (or average) with respect to the specified population model and E_π is the (design-based) expectation under repeated sampling. If the estimator is design-unbiased or approximately so, then the anticipated variance is $AV(\hat{t}) = E_M[\text{var}_\pi(\hat{t} - t_U)]$ since $E_\pi(\hat{t}) = t_U$. Thus the model expectation of a formula like (3) or (4) can be computed, resulting in a formula that includes model variance components that can be estimated using standard software. An additional advantage to this approach is the clarification of the key role that PSU and SSU sizes play in determining the measures of homogeneity. Expressions (4), (7), (9), and (11) contain measures of homogeneity, δ, $\tilde{\delta}$, δ_*, and δ_z, respectively, that are critical determinants of sample sizes. However, δ, $\tilde{\delta}$, δ_*, and δ_z are not equal to the model correlation of elements in the same cluster, except in some special circumstances, as we will illustrate.

Examples in the literature of using model variance-component estimates in survey design seem limited, even though practitioners often use the technique. A few examples are Chromy and Myers (2001); Hunter et al. (2005); Judkins and Van de Kerckhove (2003); and Waksberg et al. (1993). We demonstrate the basic approach using a random-effects model.

In a clustered population, the simplest model to consider is one with common mean and random effects for clusters and elements:

$$y_k = \mu + \alpha_i + \varepsilon_{ik}, \quad k \in U_i, \tag{13}$$

with $\alpha_i \sim (0, \sigma_\alpha^2)$, $\varepsilon_{ik} \sim (0, \sigma_\varepsilon^2)$, and the errors being independent. The model correlation of any two elements in the same cluster is

$$corr(y_k, y_{k'}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2} \equiv \rho. \tag{14}$$

The model expectation of the design variance can be computed under this model, but for sample size calculation, only the approximate expectation of the between- and within-variance components for two-stage sampling are needed. First, take the case of an *srs/srs* design and the *pwr*-estimator where a common sampling rate r is used in all clusters. The approximate model expectations are needed for $B^2 = S_{U1}^2/\bar{t}_U^2$ and $\tilde{W}^2 = M^{-1}\sum_{i \in U} \frac{N_i}{\bar{N}} \frac{S_{U2i}^2}{\bar{y}_U^2}$ in (6). After some algebra, the model expectations of S_{U1}^2 and S_{U2i}^2 defined below (2) are:

$$E_M\left(S_{U1}^2\right) \doteq \left(\sigma_\alpha^2 + \mu^2\right)S_N^2 + \bar{N}^2\sigma_\alpha^2 + \sigma_\varepsilon^2$$

$$E_M\left(S_{U2i}^2\right) = \sigma_\varepsilon^2$$

where $\bar{N} = \sum_{i \in U} N_i/M$ is the average number of elements per cluster, and $S_N^2 = \sum_{i \in U}(N_i - \bar{N})^2/(M-1)$ is the population variance of the PSU sizes, N_i. We also assume that M is large so that $M - 1 \doteq M$. Assuming that the expectation of a ratio, like $S_{U1}^2/\bar{t}_U^2$, is approximately the ratio of the expectations, the model expectation of the measure of homogeneity $\tilde{\delta}$ in (7) is

$$E_M(\tilde{\delta}) \doteq \frac{\left(\sigma_\alpha^2 + \mu^2\right)\nu_N^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2/\bar{N}^2}{\left(\sigma_\alpha^2 + \mu^2\right)\nu_N^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2(1 + \bar{N}^{-2})} \tag{15}$$

where $\nu_N^2 = S_N^2/\bar{N}^2$ is the relvariance of the N_is. If $N_i = \bar{N}$, that is, all the clusters are the same size, then $\nu_N^2 = 0$ and (15) reduces to

$$E_M(\tilde{\delta}) \doteq \frac{\sigma_\alpha^2 + \sigma_\varepsilon^2/\bar{N}^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2(1 + \bar{N}^{-2})}. \tag{16}$$

If, in addition, $\bar{N}$ is sufficiently large for $\sigma_\varepsilon^2/\bar{N}^2$ to be negligible compared to σ_α^2, then $E_M(\tilde{\delta})$ does equal the model correlation in (14). However, when clusters vary in size, (15) will be a closer approximation to the measure of homogeneity needed for sample size calculation.

The result for δ in (4) is very similar. The model expectation of δ is equal to (15) but $1 + \bar{N}^{-2}$ in the denominator is replaced with $1 + \nu_N^2 + \bar{N}^{-2}$. Numerically, the model expectation of $\tilde{\delta}$ will be somewhat larger than that of δ. For δ_* and δ_z the calculations would have to be specialized to be appropriate to the forms of B_*^2, W_*^2, B_z^2, and $\tilde{W}_z^2$ used in the definitions of those measures of homogeneity. We consider only δ_* below.

Next, consider the *ppswr/srs* design where the one-draw probability of cluster i is proportional to its number of elements, that is, $p_i = N_i/M\bar{N}$. The model expectation of $S_{U1(pwr)}^2$ is

$$E_M\left(S_{U1(pwr)}^2\right) = (M\bar{N})^2\sigma_\alpha^2\left[1 - \frac{1}{M}\left(2 - \frac{1}{\bar{N}}\right)\left(\nu_N^2 + 1\right)\right] + M^2\bar{N}\sigma_\varepsilon^2.$$

The model expectation of δ_* is then approximately

$$E_M(\delta_*) \doteq \frac{\sigma_\alpha^2\left[1 - \frac{1}{M}\left(2 - \frac{1}{\bar{N}}\right)\left(\nu_N^2 + 1\right)\right] + \frac{\sigma_\varepsilon^2}{\bar{N}}}{\sigma_\alpha^2\left[1 - \frac{1}{M}\left(2 - \frac{1}{\bar{N}}\right)\left(\nu_N^2 + 1\right)\right] + \sigma_\varepsilon^2\left(1 + \frac{1}{\bar{N}}\right)} \tag{17}$$

If $N_i \equiv \bar{N}$, selecting PSUs with probability proportional to the sizes N_i is the same as equal-probability sampling. In that case, (17) reduces to approximately the same form as

Table 8. Intracluster correlations ρ from (14) under a simple random-effects model

	Values of model intracluster correlation ρ			
	Unit used for clusters			
Variable	PSUs	SSUs	Tracts	Block groups
y1	0.005	0.024	0.008	0.012
y2	0.007	0.016	0.013	0.017
y3	0.079	0.161	0.148	0.191
ins.cov	0.004	0.011	0.008	0.014
hosp.stay	0.001	0.003	0.002	0.003

(16), which is essentially equal to the model correlation in (14) when $N_i \equiv \bar{N}$ and the average cluster size is large.

Example 5. Anticipated variance components in two-stage sampling from a model. A number of software routines are available for estimating variance components – the R package lme4 (Bates et al. 2011), the SAS® procedure proc mixed, and the xtmixed routine in Stata® are examples. We used the function lmer in lme4 to estimate the variance components for the model in (13) and the corresponding intracluster correlations in (14). The type of sample design used (*srs/srs* or *ppswr/srs*) does not affect these estimates, since they are based strictly on the model in (13). The results for all variables using PSUs, SSUs, tracts, and BGs as clusters are shown in Table 8. The estimates for ρ when PSUs and SSUs are clusters are almost the same as the values of $\tilde{\delta}$ in Table 2 where *srs* is used at each stage. But when tracts and BGs of varying sizes are used as the clusters, the ρs in Table 8 are very different and much smaller than the $\tilde{\delta}$s in Table 3. As noted above, the design-based formula for $B^2/(B^2 + \tilde{W}^2)$ will estimate the same thing as the model-based calculation if the clusters have the same large size, but not otherwise.

Table 9 shows the measures of homogeneity computed from Formula (15) for an *srs/srs* design and Formula (17) for a *ppswr/srs* design, both with the *pwr*-estimator of a total. Values of $\tilde{\delta}$ in Table 9 for *srs/srs* when PSUs and SSUs are clusters are similar to those in Table 2 and Table 8. For example, $\tilde{\delta} = 0.005$ in Table 2 for y1 with PSUs as clusters and

Table 9. Measures of homogeneity $E_M(\tilde{\delta})$ and $E_M(\delta_*)$ estimated from Expression (15) for an (srs/srs, pwr-estimator) strategy and from Expression (17) for a (ppswr/srs, pwr-estimator) strategy

	PSUs	SSUs	Tracts	Block groups
$\tilde{\delta}$s for *srs/srs* design using (15)				
y1	0.005	0.024	0.159	0.198
y2	0.007	0.016	0.216	0.264
y3	0.079	0.161	0.738	0.797
ins.cov	0.004	0.011	0.503	0.569
hosp.stay	0.001	0.003	0.022	0.029
δ_*s for *ppswr/srs* design using (17)				
y1	0.005	0.025	0.008	0.012
y2	0.007	0.017	0.013	0.018
y3	0.077	0.161	0.144	0.190
ins.cov	0.005	0.012	0.008	0.015
hosp.stay	0.001	0.004	0.002	0.004

Table 10. Sample sizes of PSUs and elements computed with incorrect and correct measures of homogeneity

	Tracts		Block groups	
	m	$\bar{n}$	m	$\bar{n}$
ρs for *srs/srs* design using (14)				
`y1`	22	35	26	29
`y2`	27	28	29	24
`y3`	57	8	61	7
`ins.cov`	22	35	27	27
`hosp.stay`	12	71	15	58
δs for *srs/srs* design using (15)				
`y1`	58	7	61	6
`y2`	62	6	65	5
`y3`	84	2	86	2
`ins.cov`	76	3	78	3
`hosp.stay`	32	21	35	18

is 0.005 in both Tables 8 and 9. `PSUs` and `SSUs` have almost the same size, and therefore (15) reduces to the model formula for the correlation in (14). When tracts or BGs are clusters, the values of ρ in Table 8 and $\tilde{\delta}$ and δ_* in Table 9 are substantially different – for example, when tracts are clusters $\rho = 0.148$ for y3 but $\tilde{\delta} = 0.738$ for *srs/srs* in Table 9. However, 0.738 is close to the value of 0.730 for (tracts, *srs/srs*) in Table 3. That is, using the correlation estimated from the model in the variance formula for a total in (7) would be a mistake, as shown in Example 6 below. However, using the model correlation to calculate a measure of homogeneity in (15) works.

Example 6. Effect of using incorrect measure of homogeneity on sample size calculation. Suppose that the design is *srs/srs* with a fixed second-stage sampling rate and that tracts or BGs are used as PSUs. The cost assumptions are the same as those in Example 2. Table 10 in its upper bank lists the sample sizes computed from (12), assuming that the model correlations in (14) can be used for $\tilde{\delta}$. This would be appropriate if tracts and BGs were equal sized. The lower tier of Table 10 shows the sample sizes computed when the measures of homogeneity are the ones proper for tracts and BGs that are computed from (15). Since the correct $\tilde{\delta}$s in Table 9 are much larger than the model correlations in Table 8, the sample sizes of tracts and BGs in the lower tier are larger than in the upper tier of Table 10. The sample sizes of elements per tract or BG are correspondingly lower when the appropriate measures of homogeneity are used. All of the allocations in Table 10 respect the cost constraint of \$100,000, but the one in the lower tier will yield smaller CVs (i.e., more precise estimates), assuming that the values of $\tilde{\delta}$ from (15) are correct.

Results are different for a *ppswr/srs* design. The values of the model correlation ρ in Table 8 and the measures of homogeneity $\tilde{\delta}$ in Table 9 for tracts and BGs are almost identical. They are also very close to the design-based values in Table 6, resulting in relatively similar sample sizes for the two stages of the design using either method. As noted earlier, probability proportional to cluster-size sampling was extremely effective in reducing the between component of variance in the Maryland population. The upshot of this is that the measures of homogeneity and thus the sample sizes are quite similar to ones for a population in which clusters all have the same size.

5. Conclusion

Using variance components and measures of homogeneity are key parts of designing multistage samples. The relative sizes of the variance components are very sensitive to the sizes of the first-stage units or clusters themselves. Many textbooks present specialized variance formulae that assume that all clusters contain the same number of elements. However, varying cluster sizes can increase the measures of homogeneity that affect the precision of estimates from a two-stage sample. Having clusters that are more internally homogeneous will require more clusters and fewer elements per cluster to be sampled to achieve a desired level of precision. The effect of having variable-sized clusters also depends on the method of selecting clusters and the type of estimator that is used. Probability proportional to cluster-size sampling is more efficient than simple random sampling of clusters. Use of a ratio estimator when clusters are sampled via *srs* will temper some of the precision losses when cluster sizes vary, but still will be less efficient than *pps* sampling. As a result, recognizing the effects of varying cluster sizes is important for designing efficient samples and choosing estimators.

The variation of the tract sizes in the Maryland population used in our examples is considerably more than practitioners would prefer when defining PSUs for a household survey. For example, the range of the number of persons per tract is 86 to 13,579. Having such a large variation in PSU sizes leads to large differences in the cluster totals of analysis variables. This causes the between-cluster variance component to be large, which in turn leads to high measures of homogeneity and inefficiency if an equal-probability sample of clusters is selected. Standard practice would be to combine the small tracts or BGs so that all PSUs have some prescribed minimum number of persons. Although variation in cluster sizes can have a dramatic effect on the measures of homogeneity needed to design a sample, this seems to be rarely emphasized in sampling texts.

If the designer has some flexibility in forming the clusters, as would usually be the case in a household survey, clusters with nearly equal numbers of elements should definitely be created. In some surveys, however, the clusters are naturally occurring units, like schools, classrooms, or establishments. In those cases, one may have to live with the predefined units, but considering the variation in cluster size will be important when determining sample sizes. This will be true whether clusters are selected with equal probability or with probabilities proportional to their sizes as measured by counts of elements. Generally speaking, sampling unequal-sized clusters with probabilities proportional to their sizes will be more efficient as long as the measure of sizes (MOSs) are accurate and cluster totals of analysis variables are closely related to MOSs. If clusters are selected with equal probability, some efficiency can be recovered by using a ratio estimator of a total rather than a π-estimator; however, in the examples we presented, *pps* sampling will still be more efficient.

We have not covered several topics that are important in practice: three-stage sampling and nonlinear estimators more general than a ratio estimator. Three-stage sampling is used in many household surveys, but involves more complex variance formulae that we plan to address in a separate paper. Although we did not cover nonlinear estimators, such as the poststratification estimator or the general regression estimator, the analyses presented here will apply after forming a linear approximation to the nonlinear estimator (see, e.g., Binder

1995). The sizes of design effects for these nonlinear estimators can be quite different from those for the π-estimator, as pointed out by Park and Lee (2004).

Another important topic that we have omitted is domain estimation. The general technique of breaking the variance of an estimator into components will apply to subpopulation estimates. However, using the usual method of coding y to 0 for units not in the subpopulation will have an effect on the size of between- and within-variance components, which in turn affects the measures of homogeneity and sample size calculations. Whether a domain is spread over most clusters or present only in a subset of them will also affect the efficiency of sampling probability proportional to an MOS compared to equal-probability sampling of clusters.

Sample size calculation is an important aspect of survey design. Using formulae with assumptions that are not supported by the population at hand can result in either wasted project funding, an insufficient sample size with lower precision than desired, or inconclusive hypothesis tests. We demonstrated techniques not clearly specified in the literature to properly account for the variance components under two first-stage sample designs and the implications for assuming equal cluster sizes when in fact this is not the case. With knowledge in hand, survey statisticians are better equipped to design multistage surveys, and teachers will be better able to explain some of the nuances of sample design to students.

6. References

Bates, D., M. Maechler, and B. Bolker. 2011. *lme4: Linear Mixed-Effects Models Using S4 Classes*. Available at: http://CRAN.R-project.org/package=lme4. (accessed October 12, 2015).

Binder, D. 1995. “Linearization Methods for Single Phase and Two-Phase Samples: A Cookbook Approach.” *Survey Methodology* 22: 17–22.

Center for Disease Control and Prevention. 2009. *National Health and Nutrition Examination Survey: 1999–2010 survey content*. Washington, DC: Department of Health and Human Services. Retrieved from www.cdc.gov/nchs/data/nhanes/survey_content_99_10.pdf.

Center for Disease Control and Prevention. 2012. *National Health Interview Survey*. Retrieved from National Center for Health Statistics: http://www.cdc.gov/nchs/nhis.htm.

Chromy, J. and L. Myers. 2001. “Variance Models Applicable to the NHSDA.” In Proceedings of the Survey Research Methods Section: American Statistical Association, August 5–9, 2001. Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/sections/SRMS/Proceedings/. (accessed October 12, 2015).

Cochran, W. 1977. *Sampling Techniques*, (3rd edition). New York: John Wiley & Sons.

Gabler, S., S. Haeder, and P. Lahiri. 1999. “A Model Based Justification of Kish’s Formula for Design Effects for Weighting and Clustering.” *Survey Methodology* 25: 105–106.

Hansen, M., W. Hurwitz, and M. Madow. 1953a. *Sample Survey Methods and Theory*, (Vol. I) New York: John Wiley & Sons.

Hansen, M., W. Hurwitz, and W. Madow. 1953b. *Sample Survey Methods and Theory*, (Vol. II) New York: John Wiley & Sons.

Hunter, S., K. Bowman, and J. Chromy. 2005. "Results of the Variance Component Analysis of Sample Allocation by Age in the National Survey on Drug Use and Health." In Proceedings of the Survey Research Methods Section: American Statistical Association, August 7–11, 2005 (pp. 3132–3136). Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/sections/SRMS/Proceedings/. (accessed October 12, 2015).

Isaki, C. and W. Fuller. 1982. "Survey Design Under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77: 89–96. Doi: http://dx.doi.org/10.1080/01621459.1982.10477770.

Judkins, D. and W. van de Kerckhove. 2003. *Residential Energy Consumption Survey 2005 Optimization*. Washington, DC: Department of Energy.

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

Lynn, P. and S. Gabler. 2005. "Approximations to b^* in the Prediction of Design Effects Due to Clustering." *Survey Methodology* 31: 101–104.

Lohr, S. 2010. *Sampling: Design and Analysis*, (2nd edition). Boston, MA: Brooks/Cole CENGAGE Learning.

Park, I. and H. Lee. 2004. "Design Effects for the Weighted Mean and Total Estimators Under Complex Survey Sampling." *Survey Methodology* 30: 183–193.

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

Searle, S., G. Casella, and C. McCulloch. 1992. *Variance Components*. New York: John Wiley & Sons.

U.S. Census Bureau. 2011. *2010 Census Redistricting Data (Public Law 94–171) Summary File*. Washington, DC: Department of Commerce. Available at: http://www.census.gov/prod/cen2010/doc/pl94-171.pdf. (accessed October 12, 2015).

Valliant, R., J.A. Dever, and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Valliant, R., J.A. Dever, and F. Kreuter. 2015. `PracTools:` Tools for Designing and Weighting Survey Samples. R package version 0.3. Available at: http://CRAN.R-project.org/package=PracTools. (accessed November 25, 2015).

Waksberg, J., S. Sperry, D. Judkins, and V. Smith. 1993. "National Survey of Family Growth: Evaluation of Linked Design." *Vital and Health Statistics* 117: July 1993. 20pp. (PHS) 93-1391. PB94-103462. PC A04 MF A01. Available at: http://www.cdc.gov/nchs/products/series/series02.html (accessed October 12, 2015).

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

Received January 2013
Revised January 2015
Accepted January 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 783–807, http://dx.doi.org/10.1515/JOS-2015-0045

On Proxy Variables and Categorical Data Fusion

Li-Chun Zhang[1]

The problem of inference about the joint distribution of two categorical variables based on knowledge or observations of their marginal distributions, to be referred to as *categorical data fusion* in this paper, is relevant in statistical matching, ecological inference, market research, and several other related fields. This article organizes the use of proxy variables, to be distinguished from other auxiliary variables, both in terms of their effects on the uncertainty of fusion and the techniques of fusion. A measure of the gains of efficiency is provided, which incorporates both the identification uncertainty associated with data fusion and the sampling uncertainty that arises when the theoretical bounds of the uncertainty space are unknown and need to be estimated. Several existing techniques for generating fusion distributions (or datasets) are described and some new ones proposed. Analysis of real-life data demonstrates empirically that proxy variables can make data fusion more precise and the constructed fusion distribution more plausible.

Key words: Identification problem; sampling uncertainty; uncertainty analysis; fusion distribution; fusion data; proxy variable; relative efficiency.

1. Introduction

Some statistical problems are characterized by a lack of observations of interest. A familiar example is incomplete data due to survey nonresponse. Examples of other 'censoring' mechanisms that have received attention in the social sciences can be found in Manski (1995). In all these cases, the lack of observations of interest induces an *identification uncertainty* about any stipulated model assumptions that is not a question of the sample size but one of the data structure, such that "inference even from an infinite number of observations is subject" (Koopmans 1949, 132).

The particular situation to be considered in this article is inference about the joint distribution of two *target* categorical variables of interest based on knowledge or observations of their marginal distributions, to be referred to as *categorical data fusion*. The setting is readily recognizable in statistical matching (e.g., D'Orazio et al. 2006b; Rässler 2002), ecological inference (e.g., Wakefield 2004; King 1997), and several other related fields.

The first topic of interest in data fusion is uncertainty analysis. The identification problem implies that there exist a set of probability distributions of the target two-way contingency table, denoted by Θ and referred to as the *uncertainty space*, whose elements can be constrained by knowledge or observations of the table margins.

[1] University of Southampton, S3RI/Social Statistics and Demography, Highfield Southampton SO17 1BJ, UK and Statistics Norway, P.O. Box 8131 Dep. 0033 Oslo, Norway. Email: L.Zhang@soton.ac.uk.

Acknowledgment: I would like to thank three anonymous referees for their insightful comments, and an Associate Editor for helpful suggestions regarding the presentation.

The conceptualization and measure of uncertainty space for statistical matching have been considered in Kadane (1978), Moriarity and Scheuren (2001), D'Orazio et al. (2006a), Rässler and Kiesel (2009) and Conti et al. (2012, 2013).

The second topic of interest is data fusion techniques. Each element of the uncertainty space corresponds to a specific joint distribution. Identification is only possible by stipulation. The thus-identified joint distribution will be referred to as the *fusion distribution*. A fusion distribution should be regarded as a *pseudo* estimate of the target distribution, since the underlying assumption is not empirically verifiable. Sometimes, as is often the case in statistical matching, the practical interest is to construct a *fusion dataset* that conforms to the fusion distribution. It is natural to treat the two as the dual aspects of each data fusion technique. Indeed, D'Orazio et al. (2006b) refer to the construction of fusion distribution as statistical matching at the macro level and to fusion data as that at the micro level.

In this article we organize for the first time the use of proxy variables for categorical data fusion. We define a *proxy* variable to be similar in concept to the target variable and have the same support. For example, having a registered job-seeker status or not can be considered a proxy variable of being unemployed or not in the Labor Force Survey (LFS), but not whether a person is male or female even though both are binary variables. On the other hand, having a registered job-seeker status or not is not a proxy variable of the there-category LFS status (employed, unemployed, not in the labor force), because of the different support. It is helpful to distinguish between proxy and other auxiliary variables in data fusion both with regards to uncertainty and technique.

The rest of the article is arranged as follows. In the first place, when available, the proxy variables are usually the covariates that have the strongest association with the target ones. To facilitate a precise statement of this, in Section 2 we propose a measure of the relative efficiency of fusion with and without the proxy (or other auxiliary) variables, which builds on the measure of uncertainty space proposed by Conti et al. (2012), but here incorporates additionally the sampling uncertainty when the relevant theoretical uncertainty bounds are unknown and need to be estimated.

Next, existing methods, including conditional independence model, middle-of-bounds estimation and iterative proportional fitting, are discussed in Section 3. Note is given whether a technique can be more readily motivated depending on the availability of proxy variables. We also introduce some new methods, including a recursive derivation of the middle-of-bounds estimates, and in particular a flexible technique of *distribution calibration* for making use of proxy variables.

Thirdly, using real-life data on education, election turnout, and labor force status, we demonstrate empirically in Section 4 that proxy variables can potentially yield not only huge reduction of the identification uncertainty of data fusion, but also more plausible pseudo estimates of the target joint distribution. Finally, a short summary is given in Section 5.

2. Uncertainty Analysis

2.1. The Identification Problem

There is a general identification problem in data fusion due to the lack of joint observations of the target data. The problem can be characterized by the breakdown of likelihood-based inference of the uncertainty space Θ. Binary data can be used to provide an illustration.

Let $Y_1 = 0, 1$ and $Y_2 = 0, 1$ be the two target variables. Consider first the situation where Y_1 and Y_2 are separately observed in two independent and disjoint samples. This is a typical setting for statistical matching. Let n_1 and n_2 be the respective sample sizes, and let y_1 and y_2 be the respective numbers of $Y_1 = 1$ and $Y_2 = 1$. Let y_1 have the Binomial (n_1, ϕ_1) distribution where $\phi_1 = P(Y_1 = 1)$, and let y_2 have the Binomial (n_2, ϕ_2) distribution where $\phi_2 = P(Y_2 = 1)$. Note that the two outcomes y_1 and y_2 are independent of each other because they are observed in two independent samples of Y_1 and Y_2, respectively. The likelihood is then given by

$$\begin{aligned} L(\phi_1, \phi_2; y_1, y_2) &\propto \phi_1^{y_1}(1-\phi_1)^{n_1-y_1}\phi_2^{y_2}(1-\phi_2)^{n_2-y_2} \\ &= (\theta_{10}+\theta_{11})^{y_1}(\theta_{00}+\theta_{01})^{n_1-y_1}(\theta_{01}+\theta_{11})^{y_2}(\theta_{00}+\theta_{10})^{n_2-y_2} \propto L(\theta; y_1, y_2) \end{aligned}$$

where $\theta = (\theta_{ij})_{i,j=0,1}$, and $\theta_{ij} = P(Y_1 = i, Y_2 = j)$ for $i,j = 0, 1$. The maximum-likelihood estimate (MLE) of (ϕ_1, ϕ_2) is $(\hat{\phi}_1, \hat{\phi}_2) = (y_1/n_1, y_2/n_2)$. However, the MLE $\hat{\boldsymbol{\theta}}$ can not be uniquely identified, but is constrained to a region called the likelihood ridge (D'Orazio et al. 2006a) defined by

$$\hat{\theta}_{10} + \hat{\theta}_{11} = y_1/n_1 \qquad \text{and} \qquad \hat{\theta}_{01} + \hat{\theta}_{11} = y_2/n_2 \qquad \text{and} \qquad \sum_{ij} \hat{\theta}_{ij} = 1$$

Next, suppose a single sample of the target data of size n, where joint observations of Y_1 and Y_2 are unavailable due to some censoring mechanism. Instead, only the marginal totals y_1 of $Y_1 = 1$ and y_2 of $Y_2 = 1$ are observed. Let n_{ij} be the number of units with $(Y_1, Y_2) = (i,j)$, for $i,j = 0, 1$, where $y_1 = n_{11} + n_{10}$ and $y_2 = n_{01} + n_{11}$. Suppose the joint cell counts follow the multinomial distribution with parameters θ as defined above. The likelihood is then the sum of the probabilities of all possible joint cell counts subjected to the marginal constraints, that is,

$$\begin{aligned} L(\boldsymbol{\theta}; y_1, y_2) &\propto P(y_1, y_2) \\ &= \sum_{m=L_{11}}^{U_{11}} P(n_{11} = m, n_{10} = y_1 - m, n_{01} = y_2 - m, n_{00} = n - y_1 - y_2 + m) \\ &= \sum_{m=L_{11}}^{U_{11}} b_m \theta_{11}^{m} \theta_{10}^{y_1-m} \theta_{01}^{y_2-m} \theta_{00}^{n-y_1-y_2+m} \end{aligned}$$

where $L_{11} = \max\,(y_1 + y_2 - n, 0)$, and $U_{11} = \min\,(y_1, y_2)$, and the coefficient b_m is given by

$$b_m = \frac{n!}{m!(y_1-m)!(y_2-m)!(n-y_1-y_2+m)!}$$

A variation of the setting is when one of the margins is known, as is usual in ecological inference. Suppose the marginal distribution of Y_1, that is $\phi_1 = P(Y_1 = 1)$, is known. Conditional on y_1, n_{11} and n_{01} are now modelled as two independent binomial distributions, that is Binomial $(y_1, \theta_{11}/\phi_1)$ for n_{11}, and Binomial $(n_1 - y_1, \theta_{01}/(1 - \phi_1))$ for n_{01}. The likelihood is then given by

$$L(\theta; y_1, y_2) \propto P(y_2|y_1) = \sum_{m=L_{11}}^{U_{11}} P(n_{11} = m, n_{01} = y_2 - m|y_1)$$
$$= \sum_{m=L_{11}}^{U_{11}} P(n_{11} = m|y_1)P(n_{01} = y_2 - m|n - y_1)$$

This is the same likelihood as above, except that the coefficient b_m is replaced by

$$b_m^c = \left(\frac{y_1!}{m!(y_1 - m)!}\right)\left(\frac{(n - y_1)!}{(y_2 - m)!(n - y_1 - y_2 + m)!}\right) = b_m/\left(\frac{n!}{y_1!(n - y_1)!}\right)$$

that is $b_m^c \propto b_m$ for fixed (y_1, y_2, n). Plackett (1977) demonstrates that the MLE of the log odds ratio of this 2×2 table is either ∞ or $-\infty$. Equivalently, the MLE of either $P(Y_2 = 1|Y_1 = 1)$ or $P(Y_2 = 1|Y_1 = 0)$ is 0 or1, which are all on the boundary of the likelihood ridge.

The reason for the breakdown of likelihood-based inference above is not the sample size. The number of observations might as well be infinite in any of the settings, the problem would still remain. Identification of a particular θ is only possible by stipulation, which is thus associated with an identification uncertainty that is distinct from the sampling uncertainty. The former is due to the structure of the available data, whereas the latter is basically a function of the sample size. While the sampling uncertainty will become negligible as the sample size tends to infinity, the identification uncertainty could remain fundamentally unchanged. Therefore, for proper inference in data fusion, it is necessary to quantify the identification uncertainty.

2.2. *Measure of Identification Uncertainty*

A natural approach is to construct a measure of the uncertainty space Θ, in the sense that larger Θ would imply greater identification uncertainty and *vice versa*. Denote by $Y_1 = 1, \ldots, H$ and $Y_2 = 1, \ldots, J$ the target variables of interest. Let $\phi_i = P[Y_1 = i]$ and $\phi_j = P(Y_2 = j)$, where the simplified notation requires that one observe the notational correspondence between i and Y_1 and between j and Y_2. Let $\theta_{ij} = P[(Y_1, Y_2) = (i,j)]$ be the target joint distribution. The Fréchet inequalities for θ_{ij} are given as

$$\max\ (\phi_i + \phi_j - 1, 0) = L_{ij} \leq \theta_{ij} \leq U_{ij} = \min\ (\phi_i, \phi_j)$$

It should be noted that logical constraints among the variables may invalidate these bounds. Such situations of incoherence are excluded from the general discussion below (see e.g., Lindley et al. 1979, Vantaggi 2008 and Brozzi et al. 2012 for discussions).

The Fréchet inequalities can also be given for any subtable as follows. Let $R_1 \subseteq \{1, \ldots, H\}$ be a subset of categories of Y_1, and let $R_2 \subseteq \{1, \ldots, J\}$ be that of Y_2. Let $\theta_R = \sum_{i \in R_1} \sum_{j \in R_2} \theta_{ij}$ be the total measure of the subtable corresponding to $R_1 \times R_2$. Let ϕ_{Rj} and ϕ_{Rj} be the respective marginal probabilities of the subtable, satisfying

$\theta_R = \sum_{i \in R_1} \phi_{Ri} = \sum_{j \in R_2} \phi_{Rj}$, given which the Fréchet inequalities for θ_{ij}, where $i \in R_1$ and $j \in R_2$, are given as

$$\max\,(\phi_{Ri} + \phi_{Rj} - \theta_R, 0) = L_{Rij} \leq \theta_{ij} | \theta_R \leq U_{Rij} = \min\,(\phi_{Ri}, \phi_{Rj}) \qquad (1)$$

The full-table bounds thus correspond to the case of $\theta_R = 1$, $R_1 = \{1, \ldots, H\}$ and $R_2 = \{1, \ldots, J\}$.

Conti et al. (2012) propose using the interval width as a point-wise measure of Θ at θ_{ij}, that is,

$$\Delta_{ij} \overset{\text{def}}{=} U_{ij} - L_{ij} \qquad (2)$$

Below we derive two results Lemma 1 and Corollary 1 in the case of categorical (Y_1, Y_2).

Lemma 1 The point-wise measure Δ_{ij} given by (2) can be directly calculated as

$$\Delta_{ij} = \min\,(\phi_i, 1 - \phi_i, \phi_j, 1 - \phi_j) \qquad (3)$$

Proof. First, it is only necessary to consider the situation where $\phi_i \leq \phi_j$, since one can handle the situation where $\phi_i \geq \phi_j$ by exchanging the generic denotation of Y_1 and Y_2. Next, provided $\phi_i \leq \phi_j$, one only needs to distinguish between two situations: $\phi_i + \phi_j \leq 1$ or $\phi_i + \phi_j > 1$. The result (3) follows then from observing:

$$\phi_i \leq \phi_j \text{ and } \phi_i + \phi_j \leq 1 \;\Rightarrow\; \Delta_{ij} = \phi_i \text{ and } \phi_i \leq \min\,(\phi_j, 1 - \phi_j) \leq 1/2 \leq 1 - \phi_i$$

$$\phi_i \leq \phi_j \text{ and } \phi_i + \phi_j > 1 \;\Rightarrow\; \Delta_{ij} = 1 - \phi_j \text{ and } 1 - \phi_j < \phi_i \leq \phi_j \text{ and } 1 - \phi_j \leq 1 - \phi_i \quad \blacksquare$$

Corollary 1 The identification uncertainty (2) is the same everywhere for binary Y_1 and Y_2.

Proof. The binary outcome space can be specified as (i, i^c) and (j, j^c), respectively, such that $\phi_{i^c} = 1 - \phi_i$ and $\phi_{j^c} = 1 - \phi_j$. It follows from (1) that Δ_{ij} is the same for any (i, j). $\blacksquare$

Next, suppose there are additional categorical auxiliary variables X, and let $k = 1, \ldots, K$ be the levels arising from cross classifying all the variables in X. The joint distributions $\phi_{ik} = P(Y_1 = i, X = k)$ and $\phi_{jk} = P(Y_2 = j, X = k)$ are assumed to be observable or known, but not the target conditional distribution $\lambda_{ij}^k = P(Y_1 = i, Y_2 = j | X = k)$. Note that, in this paper, ϕ can designate any unconditional probability while θ will be reserved for that of (Y_1, Y_2). Note also the special tensor (or Einstein) notation for conditional probability λ_{ij}^k, which facilitates the summation convention *whenever an index appears both as superscript and subscript*. An index that appears only as subscript, or only as superscript, remains constant. Thus, for example, we have $\lambda_i^k = P(Y_1 = i | X = k)$, and $\phi_i = \lambda_i^k \phi_k = \sum_k P(Y_1 = i | X = k) P(X = k) = E_X\left(\lambda_i^k\right)$, where E_X denotes expectation over X.

As a measure of the conditional identification uncertainty given $X = k$, Conti et al. (2012) use

$$\Delta_{ij}^k \overset{\text{def}}{=} U_{ij}^k - L_{ij}^k \qquad (4)$$

where $L_{ij}^k = \max\left(\lambda_i^k + \lambda_j^k - 1, 0\right) \leq \lambda_{ij}^k \leq \min\left(\lambda_i^k, \lambda_j^k\right) = U_{ij}^k$. It follows from Lemma 1 that

$$\Delta_{ij}^k = \min\left(\lambda_i^k, 1 - \lambda_i^k, \lambda_j^k, 1 - \lambda_j^k\right)$$

Note that sharper bounds are available when Y_1 and Y_2 are *ordered* categorical variables (Conti et al., 2012, 2013). Note also that it is sometimes possible to achieve point-wise identifiability due to logical constraints between the target and auxiliary variables. For instance, let Y_1 be the employment status and let X contain the payroll records at the tax authority, then the presence of wage payment in X would imply null probability of Y_1 being other than employed.

To assess the contribution of the auxiliary information $\{\phi_{ik}\}$ and $\{\phi_{jk}\}$ on θ_{ij}, put

$$\bar{\Delta}_{ij} \overset{\text{def}}{=} E_X\left(\Delta_{ij}^k\right) = \phi_k \Delta_{ij}^k \tag{5}$$

$$\bar{L}_{ij} \overset{\text{def}}{=} \phi_k L_{ij}^k = E_X\left(L_{ij}^k\right) \leq \theta_{ij} = E_X\left(\lambda_{ij}^k\right) \leq E_X\left(U_{ij}^k\right) = \phi_k U_{ij}^k \overset{\text{def}}{=} \bar{U}_{ij} \tag{6}$$

One observes that $\phi_i = \phi_k \lambda_i^k = E_X\left(\lambda_i^k\right)$ and $\phi_j = \phi_k \lambda_j^k = E_X\left(\lambda_j^k\right)$. It follows from Jensen's inequality that $L_{ij} \leq \bar{L}_{ij}$ and $\bar{U}_{ij} \leq U_{ij}$ (Conti et al., 2009), such that

$$\bar{\Delta}_{ij} = \bar{U}_{ij} - \bar{L}_{ij} \leq \Delta_{ij} \tag{7}$$

The result (7) means that the bounds $(\bar{L}_{ij}, \bar{U}_{ij})$ are never wider but can only be narrower than $\left(L_{ij}, U_{ij}\right)$ due to the additional information $\{\phi_{ik}\}$ and $\{\phi_{jk}\}$. A measure of the *relative efficiency (RE)* of this additional information for θ_{ij} can thus be given as

$$\gamma_{ij} = \bar{\Delta}_{ij}/\Delta_{ij} \tag{8}$$

In particular, powerful auxiliary information is often the case when proxy values for the target ones are available, which can greatly reduce the identification uncertainty, as will be illustrated in Section 4. Moreover, the scope of data fusion techniques is widened by the proxy variables (Section 3).

Conti et al. (2012) propose combining the point-wise measure (4) to yield an overall measure of the identification uncertainty through a set of normalising weights, that is,

$$\bar{\Delta} = w_k^{ij} \Delta_{ij}^k \qquad \text{where} \quad w_k^{ij}/\phi_k = \lambda_i^k \lambda_j^k$$

and $w_k^{ij} = \tilde{\phi}_{ijk} = P[(Y_1, Y_2, X) = (i, j, k)]$ when Y_1 and Y_2 are independent conditional on X. But other choices may be possible. In particular, setting $w_k^{ij} = w^{ij}\phi_k$, where $1_{ij} w^{ij} = 1$, yields

$$\Delta = w^{ij}\Delta_{ij} \qquad \text{and} \qquad \bar{\Delta} = w^{ij}\bar{\Delta}_{ij} \qquad \text{and} \qquad \gamma = \bar{\Delta}/\Delta = \bar{w}^{ij}\gamma_{ij} \tag{9}$$

where $\bar{w}^{ij}/w^{ij} = \Delta_{ij}/\Delta$. The choice (9) expresses the overall RE $\gamma = \bar{\Delta}/\Delta$ as a weighted average of the point-wise RE γ_{ij}s. The weights may be set as $w^{ij} = \phi_i \phi_j$. Or they may be chosen to reflect the relative 'importance' of θ_{ij}, for example, both $\Delta =$ max Δ_{ij} and $\Delta =$ min Δ_{ij} can be accommodated by (9). Note that, in the special case of binary data without auxiliary data, Δ_{ij} is a constant of (i, j), so that the overall measure Δ does not depend on the choice of the weights.

2.3. *Estimation of Uncertainty Bound*

The uncertainty bounds (L_{ij}, U_{ij}) for the target θ_{ij} depend on the marginal probabilities ϕ_i and ϕ_j. In reality these may be unknown and need to be estimated. Consequently, in uncertainty analysis one *also* needs to take into consideration the sampling uncertainty.

Take first the case where observations of Y_1 and Y_2 are available in separate and independent samples. Assume asymptotic normal distributions of $\hat{\phi}_i$ and $\hat{\phi}_j$. The distribution of the max and min of bivariate normal random variables has been studied in the literature (e.g., Nadarajah and Kotz 2008; Cain 1994). These results apply directly to $\hat{U}_{ij}$, but further derivation is needed for $\hat{L}_{ij}$. An alternative is to directly evaluate the expectations and variances by Monte Carlo calculation.

Take next the situation with a single sample, where $\hat{\phi}_i$ and $\hat{\phi}_j$ are not independent. Without losing generality, it suffices to consider $\left(\hat{L}_{11}, \hat{U}_{11}\right)$ for cell $(1, 1)$ in a 2×2 table. Denote the true cell counts by $(n_{11}, n_{10}, n_{01}, n_{00})$ where n_{11} is the cell of concern. Let $n = \sum_{i=0}^{1}\sum_{j=0}^{1} n_{ij}$. The estimates $\hat{L}_{11}$, $\hat{U}_{11}$ and $\hat{\Delta} = \hat{\Delta}_{11} = \hat{U}_{11} - \hat{L}_{11}$ are, respectively, given as

$$\hat{L}_{11} = n^{-1} \max\,(n_{11} - n_{00}, 0)$$

$$\hat{U}_{11} = n^{-1}(n_{11} + \min\,(n_{10}, n_{01}))$$

$$\hat{\Delta} = \hat{\Delta}_{11} = n^{-1}(\min\,(n_{10}, n_{01}) + \min\,(n_{11}, n_{00}))$$

The expectation and variance of $\hat{L}_{11}$ can be evaluated *via* conditioning on $m = n_{11} + n_{00}$, for $m = 1, \ldots, n$. For convenience, denote by $\wp_{b;m,\psi}$ the generic binomial probability function, that is, $\wp_{b;m,\psi} = P(B = b)$ for $B \sim \text{Binomial}(m, \psi)$. Then,

$$E(\hat{L}_{11}) = n^{-1}\sum_{m=1}^{n} \mu^{+}_{m;\psi}\wp_{m;n,\xi}$$

$$V(\hat{L}_{11}) = n^{-2}\left(\sum_{m=1}^{n} \tau^{+}_{m;\psi}\wp_{m;n,\xi} - \left(\sum_{m=1}^{n} \mu^{+}_{m;\psi}\wp_{m;n,\xi}\right)^2\right)$$

where $\xi = \theta_{11} + \theta_{00}$, and $\mu^{+}_{m;\psi} = \sum_{b=k+1}^{m} (2b - m)\wp_{b;m,\psi}$ and $\tau^{+}_{m;\psi} = \sum_{b=k+1}^{m} (2b - m)^2\wp_{b;m,\psi}$, and $\psi = \theta_{11}/(\theta_{11} + \theta_{00})$, and $k = \lfloor m/2 \rfloor$ is the largest integer less or equal to $m/2$. An alternative, closed expression for $\mu^{+}_{m;\psi}$ can be given as $\mu^{+}_{m;\psi} = m(2\psi - 1)P(B \geq k + 1) + 2(k + 1)(1 - \psi)\wp_{k+1;m,\psi}$, where $B \sim \text{Binomial}(m, \psi)$, on noting the following result (Patel et al. 1976, 201):

$$\sum_{b=k}^{m} b\binom{m}{b}\psi^b(1 - \psi)^{m-b} = m\psi P(B \geq k) + k(1 - \psi)P(B = k)$$

Similarly for $\hat{U}_{11}$. Let $m = n_{10} + n_{01}$ and $\xi = \theta_{10} + \theta_{01}$. One obtains

$$E(\hat{U}_{11}) = \theta_{11} + n^{-1}\sum_{m=1}^{n} \mu_{m;\psi}\wp_{m;n,\xi}$$

$$V(\hat{U}_{11}) = n^{-1}\theta_{11}(1 - \theta_{11}) + n^{-2}\left(\sum_{m=1}^{n} \tau_{m;\psi}\wp_{m;n,\xi} - \left(\sum_{m=1}^{n} \mu_{m;\psi}\wp_{m;n,\xi}\right)^2\right)$$

$$+2n^{-2}\left(\sum_{m=1}^{n} \eta_m\mu_{m;\psi}\wp_{m;n,\xi} - n\theta_{11}\left(\sum_{m=1}^{n} \mu_{m;\psi}\wp_{m;n,\xi}\right)\right)$$

where $\eta_m = E(n_{11}|m) = (n - m)\theta_{11}/(\theta_{11} + \theta_{00})$, and

$$\mu_{m;\psi}=E(\min(A,B)|A+B=m,B\sim\text{Binomial}(m,\psi))=\sum_{b=1}^{k} b\wp_{b;m,\psi}+\sum_{b=1}^{m-k-1} b\wp_{b;m,1-\psi}$$

$$\tau_{m;\psi}=E\left(\min(A,B)^2|A+B=m,B\sim\text{Binomial}(m,\psi)\right)=\sum_{b=1}^{k} b^2\wp_{b;m,\psi}+\sum_{b=1}^{m-k-1} b^2\wp_{b;m,1-\psi}$$

Again, a closed expression can be given for $\mu_{m;\psi}=m\psi P(B\leq k)-(k+1)(1-\psi)\wp_{k+1;}$ $m,\psi+m(1-\psi)P(B\geq k+1)-(m-k)\psi_{\wp k;m,\psi}$. Finally, *via* the same conditioning on $m=n_{10}+n_{01}$, one obtains

$$E(\hat{\Delta})=n^{-1}\sum_{m=1}^{n}(\mu_{m;\psi_1}+\mu_{n-m;\psi_2})\wp_{m;n,\xi}$$

$$\begin{aligned}V(\hat{\Delta})=&n^{-2}\left(\sum_{m=1}^{n}\tau_{m;\psi_1}\wp_{m;n,\xi}-\left(\sum_{m=1}^{n}\mu_{m;\psi_1}\wp_{m;n,\xi}\right)^2\right)\\&+n^{-2}\left(\sum_{m=1}^{n}\tau_{n-m;\psi_2}\wp_{m;n,\xi}-\left(\sum_{m=1}^{n}\mu_{n-m;\psi_2}\wp_{m;n,\xi}\right)^2\right)\\&+2n^{-2}\left(\sum_{m=1}^{n}\mu_{m;\psi_1}\mu_{n-m;\psi_2}\wp_{m;n,\xi}-\left(\sum_{m=1}^{n}\mu_{m;\psi_1}\wp_{m;n,\xi}\right)\left(\sum_{m=1}^{n}\mu_{n-m;\psi_2}\wp_{m;n,\xi}\right)\right)\end{aligned}$$

where $\psi_1=\theta_{10}/(\theta_{10}+\theta_{01})$ and $\psi_2=\theta_{11}/(\theta_{11}+\theta_{00})$.

Now that the true target distribution $\boldsymbol{\theta}$ is not identifiable, one needs to *stipulate* a particular element in the uncertainty space $\tilde{\boldsymbol{\theta}} \in \Theta$, in order to evaluate the expectations and variances above. Various fusion distributions described in Section 3 can be used. As it will be illustrated in Section 4, the choice seems to matter little to the results. In other words, the identification uncertainty of the sampling uncertainty is usually small compared to the sampling uncertainty itself.

3. Fusion Techniques

Data fusion techniques depend not only on whether auxiliary data are available, but also the nature of the auxiliary data that are available. Note will be given whether a technique requires proxy variables or not. To focus on the identification that results from the underlying assumptions, the techniques will be described in terms of the relevant theoretical distributions. It is understood that some of these may be known while some may require estimation in a particular application.

3.1. Conditional Independence Assumption

Denote by $\{(X, Y_1), (X, Y_2)\}$ the setup where each target variable is separately observed with the auxiliary ones. The *conditional independence assumption (CIA)* is given by

$$\tilde{\lambda}_{ij}^{k} = \lambda_i^k \lambda_j^k \qquad (10)$$

The corresponding fusion distribution can be given in several expressions:

$$\tilde{\theta}_{ij} = \lambda_i^k \lambda_j^k \phi_k = \lambda_i^k \phi_{jk} = \lambda_j^k \phi_{ik} = \phi_{ik}\phi_{jk}/\phi_k$$

A schematic denotation of data fusion by the CIA is $Y_1 \coprod Y_2 | X$. The auxiliary data may or may not include proxy variables. However, the possibility of including a good proxy variable for at least one of the variables can be beneficial (Rässler 2002; D'Orazio et al. 2006b). The independence assumption (IA), that is, $Y_1 \coprod Y_2$ or $\tilde{\theta}_{ij} = \phi_i \phi_j$, can be considered as a special case of the CIA in the absence of auxiliary information.

To obtain categorical fusion data, some variant of the hot-deck imputation can be used (see e.g., Singh et al. 1993). Constraints of hot-deck imputation may easily be imposed when generating synthetic fusion data. For instance, starting from the dataset $\{(x_s, y_{1s}); s = 1, \ldots, n\}$, synthetic $\tilde{y}_{2s}$ can be generated randomly for each $s = 1, \ldots, n$ from the conditional distribution λ_j^k given $x_s = k$. However, one may wish to constrain the synthetic dataset $\{(x_s, y_{1s}, \tilde{y}_{2s}); s = 1, \ldots, n\}$ such that $\tilde{n}_{jk} = \sum_{s=1}^n I_{x_s=k} I_{\tilde{y}_{2s}=j} = n_k \lambda_j^k = n_k \phi_{jk}/\phi_k$ for all (j,k) and $n_k = \left(\sum_{s=1}^n I_{x_s=k}\right)$. This can be accomplished as follows: first, construct a vector of n_k components where $\tilde{n}_{jk}$ of them have value j, for $j = 1, \ldots, J$; then, assign any permutation of this vector to the units that have $x_s = k$. The difference between the unconstrained and constrained hot decks here is an example of the matching noise (see e.g., Conti et al. 2008 and Marella et al. 2008 for discussions).

It is convenient to merge separate datasets under the CIA. Okner (1972) is often cited as an early reference. But the CIA is understandably avoided in ecological inference, where it would have defeated its own purpose. It is interesting to note that the same assumption may be popular for generating fusion data, but disreputable when it comes to the construction of fusion distribution.

3.2. *Middle of Bounds*

To start with, consider the situation with no auxiliary data. The difference between the true θ_{ij} and any admissible $\tilde{\theta}_{ij}$, or the 'loss' of $\tilde{\theta}_{ij}$ as measured by $|\tilde{\theta}_{ij} - \theta_{ij}|$, has an upperbound

$$\Lambda_{ij} = \max\,(\tilde{\theta}_{ij} - L_{ij},\, U_{ij} - \tilde{\theta}_{ij}) = \Delta_{ij}/2 + |\tilde{\theta}_{ij} - \mu_{ij}|$$

where $\mu_{ij} = (L_{ij} + U_{ij})/2$ and $\Delta_{ij} = U_{ij} - L_{ij}$. In other words, Λ_{ij} is the *upper bound* of the identification error of $\tilde{\theta}_{ij}$. It attains the minimum value $\Delta_{ij}/2$ at

$$\tilde{\theta}_{ij} = \mu_{ij} = (L_{ij} + U_{ij})/2 \tag{11}$$

which is the *middle-of-bounds* (MoB) value that minimizes the maximum potential loss. Note that D'Orazio et al. (2006a, 2006b) define the 'middle-of-bounds' as the expectation of θ_{ij} with respect to a Bayesian distribution of the parameter. Theirs differs from the definition (11) and its minimax interpretation, except in the special case of binary Y_1 and Y_2.

The MoB fusion distribution $\tilde{\boldsymbol{\theta}}$ should be well defined and preserve all the margins of Y_1 and Y_2. Take first the binary base, and let Y_1 and Y_2 take values (i, i^c) and (j, j^c), respectively. Then,

$$
\begin{aligned}
2(\theta_{\tilde{i}j} + \theta_{\tilde{i}j^c}) &= (U_{ij} + L_{ij}) + (U_{ij^c} + L_{ij^c}) \\
&= \theta_{ij} + \min\,(\theta_{ij^c}, \theta_{i^c j}) + \max\,(\theta_{ij} - \theta_{i^c j^c}, 0) \\
&\quad + \theta_{ij^c} + \min\,(\theta_{ij}, \theta_{i^c j^c}) + \max\,(\theta_{ij^c} - \theta_{i^c j}, 0) = 2\phi_i
\end{aligned}
$$

since $\min\,(a, b) + \max\,(a - b, 0) \equiv a$ for any a and b. An MoB fusion distribution in the nonbinary case can be constructed recursively, by repeatedly referring to the basic binary case and the subtable bounds (1). Example 1 below suffices to illustrate the idea.

Example 1. Consider the target 3×3 table to the left in Table 1. The marginal ϕ_i and ϕ_j are as given, as well as the MoB values directly derived from them. Clearly, since these do not sum to the total measure $\theta_R = 1$, they do not yield a well-defined fusion distribution. However, starting from any of them, which is by definition an admissible value of the corresponding $\tilde{\theta}_{ij}$, one can construct the corresponding MoB fusion distribution *rooted* in the chosen cell. The choice of cell (1,1) is illustrated here. The initial $\tilde{\theta}_{11} = 1/8$ partitions the remaining MoB $\tilde{\theta}_{ij}$s into three groups:

1. Cell (1,2) and (1,3). The implied row margin is $\phi_{Ri} = \phi_i - \tilde{\theta}_{i1} = 3/8$ for $i = 1$. The column margins are $\phi_{Rj} = \phi_j = 1/8$ for $j = 2$ and $5/8$ for $j = 3$. The relevant subtable is given by deleting the initial first column since, whatever the values $(\tilde{\theta}_{21}, \tilde{\theta}_{31})$, they have no effect on $(\tilde{\theta}_{12}, \tilde{\theta}_{13})$ given ϕ_{Ri} and ϕ_{Rj}. Thus the total measure of the relevant subtable is $\theta_R = 1 - 1/4 = 3/4$. The MoB $(\tilde{\theta}_{12}, \tilde{\theta}_{13}) = (1/16, 5/16)$ follow from the subtable bounds (1).
2. Similarly for cell (2,1) and (3,1). The implied column and row margins are as given. The relevant subtable is given by deleting the initial first row, yield the corresponding subtable total measure $\theta_R = 1 - 1/2 = 1/2$. The MoB $(\tilde{\theta}_{21}, \tilde{\theta}_{31})$ follow from (1).
3. The remaining cells on deleting the initial row and column occupied by the root cell (1,1). The implied row margins are $\phi_{Ri} = \phi_i - \theta_{i1}$ and $\phi_{Rj} = \phi_j - \theta_{1j}$. The implied subtable total measure is $\theta_R = 1 - \sum_j \tilde{\theta}_{1j} - \sum_i \tilde{\theta}_{i1} + \tilde{\theta}_{11}$, which is $3/8$ in this case. Clearly, the initial problem is thus reduced to the smaller, remaining 2×2 table, which can be solved recursively.

Table 1. Illustration of MoB fusion distribution rooted in cell (1,1).

	$(\theta_R = 1)$	Y_2		
		1/4	1/8	5/8
	1/2	1/8	1/8	5/16
Y_1	1/3	1/8	1/16	1/6
	1/6	1/12	1/16	1/12

$$\xrightarrow{\tilde{\theta}_{11}=\frac{1}{8}}$$

Cell (1,2) & (1,3)

$(\theta_R = 3/4)$	1/8	5/8
3/8	1/16	5/16

Cell (2,1) & (3,1)

$(\theta_R = 1/2)$	1/8
1/3	1/16
1/6	1/16

Cell (2,2), (2,3), (3,2) & (3,3)

$(\theta_R = 3/8)$	1/16	5/16
13/48	1/32	23/96
5/48	1/32	7/96

The resulting MoB distribution is well defined and preserves all the margins of Y_1 and Y_2. ■

In the setting $\{(X, Y_1), (X, Y_2)\}$, the *conditional* binary MoB fusion distribution is given by

$$\tilde{\lambda}_{ij}^{k} = \mu_{ij}^{k} = \frac{1}{2}\Big(\max\left(\lambda_i^k + \lambda_j^k - 1, 0\right) + \min\left(\lambda_i^k, \lambda_j^k\right)\Big) \tag{12}$$

such that $\tilde{\theta}_{ij} = \phi_k \tilde{\lambda}_{ij}^{k}$, denoted by $\mu_{ij}|X$. The conditional nonbinary MoB fusion distribution can be constructed recursively as described above, separately for each $X = k$. Again, the auxiliary data may or may not include proxy variables, although the plausibility of the MoB distribution can be quite different with or without the latter.

The use of binary MoB fusion distribution has been considered, for example, by Chambers and Steel (2001) in the context of ecological inference, but rarely in statistical matching. The discussion above shows that the MoB fusion distribution is more complicated to handle than CIA when merging data files containing nonbinary and/or multiple target variables.

3.3. Structure-Preserving Estimation

Consider the setting $\{(X', Y_1), (X', Y_2)\}$, and suppose now the auxiliary data are $X' = (X, Z_1)$, where Z_1 is a proxy variable for Y_1 and X contains the rest of the nonproxy variables. Data fusion is yielded by turning Z_1 into $\tilde{Y}_1$, under certain distributional constraints derived from the knowledge or observations available. Denote by ϕ_{ijk} the joint distribution of (Y_1, Y_2, X), and by ϕ_{hjk} that of (Z_1, Y_2, X) where the proxy Z_1 is indexed by h, and by $\tilde{\phi}_{ijk}$ the fusion distribution $(\tilde{Y}_1, Y_2, X)$ where $\tilde{Y}_1$ has the same index as Y_1 but a distinction is made between ϕ and $\tilde{\phi}$.

Structure-preserving estimation (SPREE) operates by raking (or iterative proportional fitting) of the initial table $\{\phi_{hjk}\}$ towards certain sufficient margins that are available. To identify the constraints that may be imposed, one only needs to inspect, in a 'descending' order, the log-linear representation of the fusion distribution, that is,

$$\log\, \tilde{\phi}_{ijk} = \tilde{\alpha}_0 + \tilde{\alpha}_i + \tilde{\alpha}_j + \tilde{\alpha}_k + \tilde{\alpha}_{ij} + \tilde{\alpha}_{ik} + \tilde{\alpha}_{jk} + \tilde{\alpha}_{ijk}$$

Take first $\tilde{\alpha}_{ijk}$, which corresponds to the sufficient margin $\tilde{\phi}_{ijk}$. Since ϕ_{ijk} is unavailable, no constraint can be imposed on $\tilde{\alpha}_{ijk}$. Next, take $\tilde{\alpha}_{jk}$, for which ϕ_{jk} can be derived from $\{(X, Z_1, Y_2)\}$ and imposed through raking. The case similar for $\tilde{\alpha}_{ik}$, where ϕ_{ik} can be derived from $\{(X, Z_1, Y_1)\}$, but not $\tilde{\alpha}_{ij}$, which requires the knowledge of ϕ_{ij}. There is no need to go through the lower-order terms as these will be fixed through the constraints already included: $\{\phi_{ik}\}$ and $\{\phi_{jk}\}$. Note that one needs to ensure that these two distributions are consistent with each other if they are estimated from separate data sources. The fusion distribution by SPREE can be characterized by the proxy interactions, derived from (Z_1, Y_2, X), which are preserved by raking

$$(\tilde{\alpha}_{ij}, \tilde{\alpha}_{ijk}) = (\alpha_{hj}, \alpha_{hjk}) \qquad \text{and} \qquad \tilde{\lambda}_i^{jk} = \tilde{\phi}_{ijk} / \phi_{jk} \tag{13}$$

A schematic representation of SPREE (13) is $(Z_1, Y_2, X) \longrightarrow (\tilde{Y}_1, Y_2, X)\big|(Y_2, X)\&(Y_1, X)$.

Singh et al. (1993) consider a similar approach of exploring proxy data through log-linear constraints in the setting of merging three data files. The term SPREE, however, is taken directly from the small-area estimation literature that dates further back (e.g., Purcell and Kish 1980).

Two other generic settings for SPREE are worth noting. First, consider $\{(X', Y_1), (X', Y_2)\}$ where $X' = (X, Z_1, Z_2)$, that is, proxy variables are available for both Y_1 and Y_2. The SPREE can either turn (Z_1, Z_2, X, Y_2) into $(\tilde{Y}_1, Z_2, X, Y_2)$, or (Z_1, Z_2, X, Y_1) into $(Z_1, \tilde{Y}_2, X, Y_1)$. Afterwards, the 'redundant' proxy variable can be integrated out to obtain the fusion distribution $\tilde{\phi}_{ijk}$, that is, Z_2 out of the distribution of $(\tilde{Y}_1, Z_2, X, Y_2)$ or Z_1 out of the distribution of $(Z_1, \tilde{Y}_2, X, Y_1)$. For instance, the SPREE turns Z_2 (indexed by g) into $\tilde{Y}_2$ by raking of $\{\phi_{higk}\}$ towards $\{\phi_{hik}\}$ and $\{\phi_{hjk}\}$, that is, $(Z_1, Z_2, X, Y_1) \rightarrow (Z_1, \tilde{Y}_2, X, Y_1)|(Z_1, Y_1, X)\&(Z_1, Y_2, X)$, which is characterized by

$$(\tilde{\alpha}_{hij}, \tilde{\alpha}_{ijk}, \tilde{\alpha}_{hijk}) = (\alpha_{hig}, \alpha_{hgk}, \alpha_{higk}) \qquad \text{and} \qquad \tilde{\lambda}_j^{hik} = \tilde{\phi}_{hijk}/\phi_{hjk} \tag{14}$$

In the second case, consider $\{Y_1, Y_2, (X, Z_1, Z_2)\}$, where there are no joint observations of the target and auxiliary variables of any kind, but there do exist joint proxy variables among the auxiliaries. The SPREE remains operative by raking of $\{\phi_{hgk}\}$ towards $\{\phi_i\}$, $\{\phi_j\}$ and $\{\phi_k\}$, that is, $(Z_1, Z_2, X) \rightarrow (\tilde{Y}_1, \tilde{Y}_2, X)|Y_1\&Y_2\&X$, under which

$$(\tilde{\alpha}_{ij}, \tilde{\alpha}_{ik}, \tilde{\alpha}_{jk}, \tilde{\alpha}_{ijk}) = (\alpha_{hg}, \alpha_{hk}, \alpha_{gk}, \alpha_{hgk}) \tag{15}$$

It is instructive to note that neither the CIA (10) nor the MoB (12) is able to utilize the auxiliary data (X, Z_1, Z_2) in this setting.

3.4. *Distribution Calibration*

To start with, observe the setting $\{Y_1, Z_1\}$, where the target Y_1 and proxy Z_1 are separately available. To turn Z_1 into $\tilde{Y}_1$ that has the same distribution as Y_1, one only needs to identify an $H \times H$ matrix $\xi = \{\xi_i^h; i, h = 1, \ldots, H\}$, where $1^i \xi_i^h = 1$, such that

$$\tilde{\phi}_i = \xi_i^h \phi_h = \phi_i$$

Morever, being a gross-flow matrix from Z_1 to $\tilde{Y}_1$, ξ tells one how to generate a set of values $\{\tilde{Y}_{1s}; s = 1, \ldots, n\}$ from the initial proxy values $\{z_{1s}; s = 1, \ldots, n\}$ by constrained hot deck. Subjected to rounding, Tr(nξ) initial proxy values will then remain the same, where n is the diagonal matrix of $(n\phi_h)_{h=1,\ldots,H}$, while the rest $n -$ Tr(nξ) will be changed. By contrast, with $\delta = \{\delta_i^h\}$ where $\delta_i^h = 1$ if $i = h$ and 0 otherwise, no proxy values will be changed at all. This suggests as a well defined approach to obtain some minimum-change $\tilde{\xi}$ by solving the following optimization problem:

$$\min_{\xi} D(\xi, \delta) \qquad \text{subject to} \quad \phi_i = \xi_i^h \phi_h \quad \text{and} \quad 1^i \xi_i^h = 1 \quad \text{and} \quad \xi_i^h \geq 0 \tag{16}$$

where $D(\xi, \delta)$ is the distance function of choice. For instance, to minimize the number of changes of the initial proxy values, one can put $D = \text{Tr}(\mathbf{n}\delta) - \text{Tr}(\mathbf{n}\xi) = n - \text{Tr}(\mathbf{n}\xi)$. Or, a squared Euclidean distance function between ξ and δ is given by $D = \sum_{i,h} (\xi_i^h - \delta_i^h)^2$.

Provided additional nonproxy auxiliary data, *distribution calibration (DC)* defined by (16) can be applied conditionally. Suppose the setting $\{(X, Y_1), (X, Z_1)\}$.

Conditional distribution calibration (CDC) from Z_1 to $\tilde{Y}_1$ for each $X = k$ yields $\tilde{\xi}_k = \{\tilde{\xi}_i^{hk}; i, h = 1, \ldots, H\}$, such that

$$\tilde{\lambda}_i^k = \lambda_h^k \tilde{\xi}_i^{hk} = \lambda_i^k \qquad \text{and} \qquad \tilde{\phi}_i = \phi_k \tilde{\lambda}_i^k = \phi_k \lambda_i^k = \phi_i \qquad \text{and}$$

$$\tilde{\phi}_{ik} = \phi_k (\phi_{ik}/\phi_k) = \phi_{ik}$$

Note that a different distribution $\tilde{\phi}_{ik}$ of $(\tilde{Y}_1, X)$ would be generated by unconditional DC, that is, $\phi_i = \tilde{\xi}_i^h \phi_h$, since $\tilde{Y}_1$ is then independent of X given Z_1, such that $\tilde{\phi}_{ik} = \phi_{hk} \tilde{\xi}_i^h \neq \phi_{ik}$.

Given the relevant proxy variables, DC and CDC can be used to generate a fusion distribution, whether or not there are joint observations of the target and proxy variables. Consider again the setting $\{Y_1, Y_2, (X, Z_1, Z_2)\}$. A scheme of DC can be as follows:

$$\left.\begin{array}{lll} Z_1 \overset{\text{DC}}{\rightarrow} \tilde{Y}_1 & \Rightarrow & \tilde{\phi}_i = \phi_h \xi_i^h = \phi_i \\ Z_2 \overset{\text{DC}}{\rightarrow} \tilde{Y}_2 & \Rightarrow & \tilde{\phi}_j = \phi_g \xi_j^g = \phi_j \end{array}\right\} \Rightarrow \quad \tilde{\phi}_{ijk} = 1^{gh} \tilde{\phi}_{ghijk}^{Z_2 Z_1 \tilde{Y}_1 \tilde{Y}_2 X} = \lambda_k^{gh} \phi_{gh} \xi_i^h \xi_j^g$$

where the last expression follows since $\tilde{Y}_1$ is independent of the other variables given Z_1 and similarly for $\tilde{Y}_2$ given Z_2. This is a different fusion distribution than that by SPREE (15).

It is worth noting that DC and CDC can be useful for generating fusion data prescribed by another fusion technique. Take the SPREE (15) under the setting $\{Y_1, Y_2, (X, Z_1, Z_2)\}$. It is not immediately clear how to generate the fusion data it implies. However, let $\tilde{\lambda}_p^k$ be the fusion conditional probability of $p = (i, j)$ given $X = k$. Let $q = \big(h, g\big)$ index $\big(Z_1, Z_2\big)$ in accordance. Then, CDC satisfying $\tilde{\lambda}_p^k = \lambda_q^k \xi_p^{qk}$ yields the gross-flowmatrix that can turn (Z_1, Z_2) into the SPREE $(\tilde{Y}_1, \tilde{Y}_2)$ with minimum changes given $X = k$. As another example, consider CDC under the setting $\{(X, Y_1), (X, Y_2), (X, Z_1, Z_2)\}$:

$$\left.\begin{array}{lllll} Z_1 \overset{\text{CDC}}{\longrightarrow} \tilde{Y}_1 | X & \Rightarrow & \tilde{\lambda}_i^k = \lambda_i^k & \Rightarrow & \tilde{\lambda}_{ik} = \phi_{ik} \\ Z_2 \overset{\text{CDC}}{\longrightarrow} \tilde{Y}_2 | X & \Rightarrow & \tilde{\lambda}_j^k = \lambda_j^k & \Rightarrow & \tilde{\lambda}_{jk} = \phi_{jk} \end{array}\right\} \quad \Rightarrow \quad \tilde{\phi}_{ijk} = \frac{\tilde{\phi}_{ik} \tilde{\phi}_{jk}}{\phi_k} = \frac{\phi_{ik} \phi_{jk}}{\phi_k}$$

that is, exactly the same fusion distribution as that of the CIA in the setting $\{(X, Y_1), (X, Y_2)\}$. But CDC can yield different fusion data. For instance, suppose $\{(X, Y_1), (X, Y_2)\}$ represent two separate sample datasets, while $\{(X, Z_1, Z_2)\}$ is a population register dataset. On the one hand, a population fusion dataset can be generated by CDC; on the other hand, a synthetic CIA population fusion dataset can be obtained by randomly and separately generating $\tilde{Y}_1$ and $\tilde{Y}_2$ conditional on X in the population. Both datasets will have the same fusion distribution, but the CDC data will resemble the real population much more than the CIA data.

4. Two Cases

Two real-life datasets involving education, election turnout, and labor force status variables are used to illustrate the approach to uncertainty analysis and the fusion techniques described above, and to empirically evaluate the relative efficiency of the available proxy data.

4.1. Education and Election Turnout: Binary Data

Both the highest level of education and election turnout are collected in the Norwegian Election Survey 2005, to be treated as Y_1 and Y_2, respectively. A level of education can also be compiled based on the register information available at Statistics Norway, denoted as Z_1, while the true head count can be obtained from the local electoral offices, denoted by Z_2. Both Z_1 and Z_2 can be linked to the survey at the individual level, and the observed four-way table for the respondents in Election Survey 2005 provides all the data for this illustration. For ease of exposition, only two categories "Low" and "High" are coded for the education variable.

Various settings of the data are given in Table 2. All the cross counts of Y_1 and Y_2 are given in parentheses and assumed to be unobserved. In the top block, the overall unconditional counts of (Y_1, Y_2) are given to the left, and those of (Z_1, Z_2) to the right. Together they provide the setting $\{Y_1, Y_2, (Z_1, Z_2)\}$. The next block gives the setting $\{(Z_1, Y_1), (Z_1, Y_2)\}$, where Z_1 is the only auxiliary data. The case is similar for $\{(Z_2, Y_1), (Z_2, Y_2)\}$ in the third block. Lastly, the bottom block provides the setting $\{(Z_1, Z_2, Y_1), (Z_1, Z_2, Y_2)\}$.

Table 3 illustrates the results of uncertainty analysis for $P[(Y_1, Y_2) = (Low, No)]$. The first row corresponds to the setting $\{Y_1, Y_2, (Z_1, Z_2)\}$. The estimated lower and upper bounds are $(0.0, 0.104)$. The estimated width of the uncertainty space at this point is 0.104. Since Θ measures the same everywhere in the case of binary data, as previously noted for (2), 0.104 is also the estimated overall measure of the uncertainty space. The relative efficiency is unity by definition. The associated sampling uncertainty is evaluated as described in Subsection 2.3, for which it is necessary to stipulate a joint distribution. Three alternatives are illustrated in Table 3. The first one is the true sample distribution of (Y_1, Y_2) given in Table 2; the second one is the CIA fusion distribution; and the last one is the MoB fusion distribution. It is seen that the estimated standard errors (SEs) are virtually the same using any of the three alternatives.

In a similar manner, the other rows of Table 3 provide the results under different settings of jointly available auxiliary data. It is seen that with only Z_1 available, the identification uncertainty is reduced by 17% (that is, RE $= 0.83$), whereas the reduction is 62% (that is, RE $= 0.38$) with Z_2, so that it is much more informative than Z_1. With both proxy variables available, the estimated uncertainty bounds are $(0.074, 0.095)$, strictly narrower than the initial $(0.0, 0.104)$ on both sides. The width of the interval is 0.021, which is about one fifth of that without (Z_1, Z_2). Taking into account the sampling uncertainty, an approximate 95% confidence interval of the width of the identification uncertainty interval is $(0.014, 0.028)$. In comparison, had the joint sample of (Y_1, Y_2) been available, the width of the approximate 95% confidence interval of $P[(Y_1, Y_2) = (Low, No)]$ would have been 0.027. Thus, *in this respect*, there is at least as much information about $P[(Y_1, Y_2) = (Low, No)]$ in $\{(Z_1, Z_2, Y_1), (Z_1, Z_2, Y_2)\}$ as that in $\{(Y_1, Y_2)\}$.

Table 4 illustrates a number of (pseudo) estimates of $P[(Y_1, Y_2) = (Low, No)]$ together with their respective identification assumptions. The first one (from the top) is based on the true data of (Y_1, Y_2). The next five are derived under the setting

$\{(Z_1, Z_2, Y_1), (Z_1, Z_2, Y_2)\}$. Note the difference between the two CIAs. The two situations of single proxy variable follow next. In the last setting where (Z_1, Z_2) are not jointly observed with any of the target variables, only SPREE and DC can make use of them. A few general impressions can be noted.

- All the different SPREE estimates appear reasonable here; the best ones (that is, 0.0877 and 0.0876) yield an estimated cell count 153 after rounding, which is almost identical to the true observation 154. Adjusting Z_1 towards Y_1 gives better results than adjusting Z_2 towards Y_2. But at this stage of knowledge one is unable to *deduce* this from the higher association between Z_2 and Y_2 compared to that between Z_1 and Y_1.

Table 2. Education and election turnout data.

	Y_2				Z_2		
Y_1	No	Yes		Z_1	No	Yes	
Low	(154)	(885)	1039	Low	210	920	1130
High	(28)	(676)	704	High	44	569	613
	182	1561			254	1489	
	$Z_1 = \text{Low}$ Y_2				$Z_1 = \text{High}$ Y_2		
Y_1	No	Yes		Y_1	No	Yes	
Low	(149)	(854)	1003	Low	(5)	(31)	36
High	(9)	(118)	127	High	(19)	(558)	577
	158	972			24	589	
	$Z_2 = \text{No}$ Y_2				$Z_2 = \text{Yes}$ Y_2		
Y_1	No	Yes		Y_1	No	Yes	
Low	(140)	(61)	201	Low	(14)	(824)	838
High	(26)	(27)	53	High	(2)	(649)	651
	166	88			16	1473	
	$(Z_1, Z_2) = \text{(Low, No)}$ Y_2				$(Z_1, Z_2) = \text{(Low, Yes)}$ Y_2		
Y_1	No	Yes		Y_1	No	Yes	
Low	(136)	(59)	195	Low	(13)	(795)	808
High	(8)	(7)	15	High	(1)	(111)	112
	144	66			14	906	
	$(Z_1, Z_2) = \text{(High, No)}$ Y_2				$(Z_1, Z_2) = \text{(High, Yes)}$ Y_2		
Y_1	No	Yes		Y_1	No	Yes	
Low	(4)	(2)	6	Low	(1)	(29)	30
High	(18)	(20)	38	High	(1)	(538)	539
	22	22			2	567	

Table 3. Estimated lower and upper bounds for $\boldsymbol{P}[(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = (\boldsymbol{Low}, \boldsymbol{No})]$ *and associated standard error (SE) using true data, CIA or MoB fusion distribution as basis of evaluation, estimated width of uncertainty space and true SE in parentheses, relative efficiency (RE) of proxy data.*

Joint proxy variable	Bound (Lower, Upper)	1,000 × SE of Bound (Lower, Upper)			Width	RE
		True	CIA	MoB		
-	(0.000, 0.104)	(0.0, 7.3)	(0.0, 7.3)	(0.0, 7.3)	0.104 (0.0073)	1
Z_1	(0.018, 0.104)	(9.1, 7.2)	(8.9, 7.2)	(7.1, 7.2)	0.086 (0.0065)	0.83
Z_2	(0.065, 0.104)	(6.2, 4.9)	(5.7, 4.9)	(6.2, 4.9)	0.039 (0.0044)	0.38
(Z_1, Z_2)	(0.074, 0.095)	(4.6, 4.7)	(4.4, 4.7)	(4.6, 4.7)	0.021 (0.0034)	0.20

- The CIA results are worse than SPREE in every setting for this dataset. The advantage of SPREE is particularly useful in cases without any joint observations between the proxy and target variables, where it makes much better use of the auxiliary information.
- The MoB estimates are quite reasonable as long as Z_2 is available, and Z_1 appears to bring little improvement either on its own or in addition to Z_2. The effect of the proxy data is evident if 0.0846 given (Z_1, Z_2) is compared to 0.0521 in the absence of (Z_1, Z_2).
- The Euclidean distance is used to generate the DC. The result is worse than the SPREE, but better than CIA and MOB, which are unable to make use of the proxy variables in this setting.

Table 4. Illustrated (pseudo) estimates of $\boldsymbol{P}\left[(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = (\boldsymbol{Low}, \boldsymbol{No})\right]$.

Setting	Estimate	Identification assumptions
$\{(Z_1, Z_2, Y_1, Y_2)\}$	0.0884	True sample
$\{(Z_1, Z_2, Y_1),$ $(Z_1, Z_2, Y_2)\}$	0.0856	CIA: $Y_1 \coprod Y_2 \mid (Z_1, Z_2)$
	0.0761	CIA: $Y_1 \coprod (Y_2, Z_2) \mid Z_1$ and $Y_2 \coprod (Y_1, Z_1) \mid Z_2$
	0.0846	MoB: $\mu_{ij} \mid (Z_1, Z_2)$
	0.0876	SPREE: $(Z_1, Y_2, Z_2) \rightarrow (\tilde{Y}_1, Y_2, Z_2) \mid (Y_1, Z_2) \& (Y_2, Z_2)$
	0.0863	SPREE: $(Z_1, Y_1, Z_2) \rightarrow (Z_1, Y_1, \tilde{Y}_2) \mid (Y_1, Z_1) \& (Y_2, Z_1)$
$\{(Z_1, Y_1), (Z_1, Y_2)\}$	0.0813	CIA: $Y_1 \coprod Y_2 \mid Z_1$
	0.0592	MoB: $\mu_{ij} \mid Z_1$
	0.0877	SPREE: $(Z_1, Y_2) \rightarrow (\tilde{Y}_1, Y_2) \mid Y_1 \& Y_2$
$\{(Z_2, Y_1), (Z_2, Y_2)\}$	0.0805	CIA: $Y_1 \coprod Y_2 \mid Z_2$
	0.0845	MoB: $\mu_{ij} \mid Z_2$
	0.0833	SPREE: $(Y_1, Z_2) \rightarrow (Y_1, \tilde{Y}_2) \mid Y_1 \& Y_2$
$\{Y_1, Y_2, (Z_1, Z_2)\}$	0.0622	IA: $Y_1 \coprod Y_2$
	0.0521	MoB: μ_{ij}
	0.0833	SPREE: $(Z_1, Z_2) \rightarrow (\tilde{Y}_1, \tilde{Y}_2) \mid Y_1 \& Y_2$
	0.0794	DC: $Z_1 \xrightarrow{DC} \tilde{Y}_1$ and $Z_2 \xrightarrow{DC} \tilde{Y}_2$

Finally, it may be reiterated that the choice of a particular fusion distribution is empirically unverifiable within the identification uncertainty bounds. Indeed, under each of the four settings considered in Table 4, the *same* uncertainty analysis, as given in Table 3 for the corresponding datasetting, should be reported for all the different pseudo estimates.

4.2. Labor Force Gross Flows

Labor force gross flows are of concern for both policy makers and researchers. Let the labor force status be classified as "employed (E)", "unemployed (U)" and "not in the labor force (N)" for each eligible person in some given age range. Let Y_1 be the status at time point t_1 and Y_2 that at t_2, then gross flow (i,j) refers here to the probability $\theta_{ij} = P[Y_1 = i, Y_2 = j]$. Together these form the 3×3 matrix, where the row margins $\phi_i = \sum_j \theta_{ij}$, for $i = 1, 2, 3$, form the marginal distribution of Y_1 and the column margins $\phi_j = \sum_i \theta_{ij}$, for $j = 1, 2, 3$, that of Y_2. Further classification by region, age, and so on may be of practical interest, but will not be considered here.

Countries that conduct the LFS typically apply some form of rotating panel design, so that joint observation (or panel data) of Y_1 and Y_2 are available for various combinations of t_1 and t_2. However, concerns for response burden and cost of following the same person over time will place a practical limit on the length of LFS participation, so that joint observations are not available if the difference between t_1 and t_2 is beyond that limit. For instance, in the Norwegian LFS (NLFS), each sample person participates in eight successive quarters, such that panel data are available for any two time points within a two-year span but not otherwise.

Two questions are considered below. Subsection 4.2.1 studies the efficiency of proxy data for labor force gross flows. To this end, proxy labor force status, denoted by Z_1 and Z_2 respectively, are compiled based on the various administrative data available to Statistics Norway (SN) and linked to the NLFS yearly panel between 2011 and 2012. The sources include employer/employee and self-employer registration, administration of job seekers, related health and welfare, payroll tax records, military services, and so on. Essentially the same proxy labour force status is used for the register-based census 2011. At the same time, it is acknowledged that at the individual level the proxy values will not always coincide with those that could be collected in the NLFS.

The second question to be considered is data fusion of (Y_1, Y_2), for which no joint observations are available. In particular, there is then an issue of how to make use of the data that are available for the time period between t_1 and t_2. For instance, although one does not have panel data between 2011 and 2013, one does have data between 2011 and 2012 and between 2012 and 2013, respectively. Various fusion methods can be used. For instance, under the CIA between (Y_1, Z_1) in 2011 and (Y_2, Z_2) in 2013 conditional on (Y_t, Z_t) in 2012, it becomes possible both to generate the fusion distribution of (Y_1, Y_2) and to assess the associated sampling uncertainty. However, this would not be appropriate if the identification uncertainty surrounding the CIA is ignored (Subsection 4.2.2).

4.2.1. Relative Efficiency of Proxy Labor Force Status

The data between 2011 and 2012 are given in Table 5. All joint observations of (Y_1, Y_2) are given in parentheses and assumed to be unobservable. The proxy register variables

(Z_1, Z_2) are jointly available with either of the target status, that is, the generic setting $\{(Z_1, Z_2, Y_1), (Z_1, Z_2, Y_2)\}$.

The target NLFS sample gross flows of (Y_1, Y_2) and the proxy flows of (Z_1, Z_2) are given in Table 6, together with four fusion distributions by the CIA, MoB and two SPREE methods, respectively. Comparisons between the target and proxy joint distribution show that the register flow is higher for the stable employed persons (E, E), but lower for the stable unemployed persons (U, U) and 'inactive' ones (N, N). The largest relative deviations among the off-diagonal flows occur for (U, E) and (N, U). The causes for these differences are complex. For instance, persons who are on the way back into the labor force from N may be classified as U or E if interviewed in the NLFS, but they may well remain as N in the register sources until they first become E (possibly lagging behind the NLFS), which can be a cause for register underestimation of (N, U).

Focusing on the results of data fusion, it may be noted that all the techniques adjust the proxy flows (E, E) and (N, N) downwards. The adjustment of the proxy flow (U, U) differs across the method. In particular, the off-diagonal proxy flows are all adjusted upwards, and the flows (U, E) and (N, U) are no longer the ones that relatively deviate most from the target flows. Overall, the CIA results are worse than the others, especially for the diagonal flows, whereas the MoB results may seem slightly better than the two SPREE. Indeed, compared to the average of the two SPREE results, the MoB fusion distribution is closer to the target distribution for five out of nine flows.

Still, regardless of how plausible the fusion distributions may seem compared to the direct register-based proxy distribution, they can only be treated as potentially useful pseudo estimates. Proper inference is only facilitated by uncertainty analysis. Table 7 provides the estimated identification uncertainty bounds and the associated SE with and without the proxy variables as auxiliary data. The SEs are evaluated here on the basis of the true sample distribution, but any of the fusion distributions would have yielded virtually the same results. Again, the identification uncertainty matters little to the assessment of the sampling uncertainty.

It can be seen that the sampling uncertainty is relatively small compared to the identification uncertainty, especially in terms of the width of the identification uncertainty interval. The proxy variables are most effective for reducing the identification uncertainty of the 'corner' flows (E, E), (E, N), (N, E) and (N, N). As these four measure over 95% of the outcome space, the overall measure of the uncertainty space is greatly reduced in the presence of the proxy variables. Depending on the choice of w^{ij} in the calculation of $\hat{\Delta} = w^{ij}\hat{\Delta}_{ij}$ and $\hat{\hat{\Delta}} = w^{ij}\hat{\hat{\Delta}}_{ij}$, one obtains $\hat{\Delta} = 0.266, 0.263$ or 0.269 when w^{ij} is set to the true θ_{ij}, the CIA or MoB $\tilde{\theta}_{ij}$, and $\hat{\hat{\Delta}} = 0.069, 0.069$ or 0.070 in correspondence. The overall relative efficiency of the proxy variables is 0.26 by all means.

4.2.2. Making Use of Available Data in Data Fusion

Where observations are unavailable for gross flows (Y_1, Y_2) over t_1 and t_2, various fusion distributions can be generated based on the intermediate observable target and proxy data. For instance, the gross flows between 2011 and 2013 can be derived from the observable flows between 2011 and 2012 and that between 2012 and 2013, under assumption that the labor force status in 2011 is independent of that in 2013 conditional on the status in 2012.

Table 5. Proxy and survey labor force status 2011 – 2012. Source: NLFS and registers at SN.

	$(Z_1, Z_2) = (\text{E, E})$					$(Z_1, Z_2) = (\text{E, U})$					$(Z_1, Z_2) = (\text{E, N})$			
	Y_2					Y_2					Y_2			
Y_1	E	U	N		Y_1	E	U	N		Y_1	E	U	N	
E	(5289)	(15)	(124)	5428	E	(23)	(1)	(3)	27	E	(97)	(2)	(140)	239
U	(5)	(15)	(13)	33	U	(0)	(2)	(0)	2	U	(2)	(1)	(3)	6
N	(84)	(6)	(138)	228	N	(3)	(1)	(3)	7	N	(12)	(2)	(92)	106
	5378	36	275			26	4	6			111	5	235	
	$(Z_1, Z_2) = (\text{U, E})$					$(Z_1, Z_2) = (\text{U, U})$					$(Z_1, Z_2) = (\text{U, N})$			
	Y_2					Y_2					Y_2			
Y_1	E	U	N		Y_1	E	U	N		Y_1	E	U	N	
E	(8)	(0)	(0)	8	E	(10)	(0)	(1)	11	E	(6)	(0)	(0)	6
U	(0)	(0)	(0)	0	U	(2)	(5)	(4)	11	U	(1)	(2)	(3)	6
N	(0)	(1)	(2)	3	N	(2)	(6)	(4)	12	N	(0)	(0)	(16)	16
	8	1	2			14	11	9			7	2	19	
	$(Z_1, Z_2) = (\text{N, E})$					$(Z_1, Z_2) = (\text{N, U})$					$(Z_1, Z_2) = (\text{N, N})$			
	Y_2					Y_2					Y_2			
Y_1	E	U	N		Y_1	E	U	N		Y_1	E	U	N	
E	(146)	(4)	(10)	160	E	(23)	(1)	(1)	25	E	(140)	(8)	(66)	214
U	(3)	(3)	(3)	9	U	(1)	(0)	(1)	2	U	(2)	(1)	(3)	6
N	(67)	(3)	(44)	114	N	(5)	(9)	(26)	40	N	(73)	(38)	(1487)	1598
	216	10	57			28	11	28			215	47	1556	

However, analysis of the register-based status overtime suggests that such a CIA is unattainable. Moreover, even if the CIA had seemed reasonable for the proxy gross flows, it would only have yielded plausible pseudo estimates of the target gross flows, due to the fact that identification is not verifiable empirically but is only achieved on the strength of stipulation.

To illustrate data fusion under alternative settings in this situation, a synthetic dataset has been constructed as follows. Denote by $(Y_1, Z_1) = (i, h)$ the data in 2011 and by $(Y_t, Z_t) = (k, l)$ the data in 2012, with the joint sample distribution ϕ_{hikl}. Assume the CIA and the same conditional transition probabilities from 2012 to 2013 as from 2011 to 2012, that is, put $\lambda_{jg}^{lk} = \phi_{lkjg}/\phi_{lk}$ equal to the corresponding $\lambda_{kl}^{hi} = \phi_{hikl}/\phi_{hi}$, for $j, g = 0, 1$. The synthetic joint distribution over 2011, 2012, and 2013 is then given by $\phi_{hikljg} = \phi_{hikl}\phi_{lkjg}/\phi_{lk}$, from which the synthetic joint distribution of (Z_1, Y_1, Y_2, Z_2) can be obtained by integrating out (Y_t, Z_t), that is, $\phi_{hijg} = 1^{kl}\phi_{hikljg}$, and so on.

Consider three settings: (i) ignore (Y_t, Z_t) and assume the setting $\{(Z_1, Z_2, Y_1), (Z_1, Z_2, Y_2)\}$, that is, with joint auxiliary data (Z_1, Z_2), (ii) assume the setting $\{(Z_1, Z_2, Z_t, Y_t, Y_1), (Z_1, Z_2, Z_t, Y_t, Y_2)\}$, that is, with joint auxiliary data (Z_1, Z_2, Z_t, Y_t), and (iii) ignore (Z_1, Z_2, Z_t) and assume the setting $\{(Y_t, Y_1), (Y_t, Y_2)\}$, where Y_t may be considered a proxy for Y_1 as well as for Y_2.

The respective theoretical uncertainty bounds and width of the nine gross flows between Y_1 and Y_2 are given in Table 8. It is clear that using all the available joint auxiliary data, that is (Z_1, Z_2, Z_t, Y_t) here, provides the narrowest uncertainty bounds. There is more

Table 6. Target, proxy and fusion labor force gross flows by CIA, MoB and SPREE

	Target gross flows				Proxy gross flows		
	Y_2				Y_2		
Y_1	E	U	N	Y_1	E	U	N
E	0.6736	0.0057	0.0402	E	0.6846	0.0049	0.0410
U	0.0083	0.0030	0.0052	U	0.0030	0.0020	0.0037
N	0.0400	0.0065	0.2176	N	0.0480	0.0020	0.2107
	CIA: $Y_1\coprod Y_2\|(Z_1, Z_2)$				MoB: $\mu_{ij}\|(Z_1, Z_2)$		
	Y_2				Y_2		
Y_1	E	U	N	Y_1	E	U	N
E	0.6460	0.0059	0.0675	E	0.6628	0.0075	0.0501
U	0.0078	0.0013	0.0074	U	0.0078	0.0058	0.0076
N	0.0681	0.0080	0.1880	N	0.0510	0.0068	0.2058
	SPREE: $Z_1 \rightarrow \tilde{Y}_1\|(Y_1, Z_2)\&(Y_2, Z_2)$				SPREE: $Z_2 \rightarrow \tilde{Y}_2\|(Y_1, Z_1)\&(Y_2, Z_1)$		
	Y_2				Y_2		
Y_1	E	U	N	Y_1	E	U	N
E	0.6530	0.0053	0.0612	E	0.6567	0.0084	0.0543
U	0.0081	0.0022	0.0062	U	0.0077	0.0022	0.0065
N	0.0609	0.0077	0.1956	N	0.0575	0.0045	0.2022

Table 7. Estimated identification bounds and width (SE in parentheses; all numbers in 10^4)

Proxy		$(Y_1, Y_2) =$ (E, E)			$(Y_1, Y_2) =$ (E, U)			$(Y_1, Y_2) =$ (E, N)	
variable	Lower	Upper	Width	Lower	Upper	Width	Lower	Upper	Width
-	4413	7194	2781	0	151	151	0	2630	2630
	(92)	(48)	(48)	(0)	(13)	(13)	(0)	(48)	(48)
(Z_1, Z_2)	6276	6980	703	12	138	126	155	857	702
	(38)	(28)	(26)	(5)	(12)	(12)	(18)	(27)	(25)
Proxy		$(Y_1, Y_2) =$ (U, E)			$(Y_1, Y_2) =$ (U, U)			$(Y_1, Y_2) =$ (U, N)	
Variable	Lower	Upper	Width	Lower	Upper	Width	Lower	Upper	Width
-	0	165	165	0	151	151	0	165	165
	(0)	(14)	(14)	(0)	(12)	(12)	(0)	(14)	(14)
(Z_1, Z_2)	12	143	131	1	115	114	5	148	143
	(3)	(13)	(12)	(3)	(11)	(11)	(3)	(13)	(13)
Proxy		$(Y_1, Y_2) =$ (N, E)			$(Y_1, Y_2) =$ (N, U)			$(Y_1, Y_2) =$ (N, N)	
variable	Lower	Upper	Width	Lower	Upper	Width	Lower	Upper	Width
-	0	2642	2642	0	151	151	0	2630	2630
	(0)	(48)	(48)	(0)	(13)	(13)	(0)	(47)	(47)
(Z_1, Z_2)	155	866	711	0	136	136	1695	2422	726
	(14)	(27)	(26)	(0)	(12)	(12)	(27)	(28)	(26)

Table 8. Theoretical identification bounds and width given auxiliary data (all numbers in 10^4)

	$P(E,E) = 6,411$			$P(E,U) = 71$			$P(E,N) = 685$		
Joint auxiliary data	Lower	Upper	Width	Lower	Upper	Width	Lower	Upper	Width
(Z_1, Z_2)	6,038	6,738	702	19	149	130	357	1,065	708
(Z_1, Z_2, Z_t, Y_t)	6,287	6,718	431	26	137	111	376	811	435
Y_t	6,241	7,164	923	0	151	151	0	857	857

	$P(U,E) = 98$			$P(U,U) = 9$			$P(U,N) = 60$		
Joint auxiliary data	Lower	Upper	Width	Lower	Upper	Width	Lower	Upper	Width
(Z_1, Z_2)	19	151	132	0	94	94	9	146	137
(Z_1, Z_2, Z_t, Y_t)	40	148	108	2	86	84	11	120	109
Y_t	0	167	167	0	138	138	0	167	167

	$P(N,E) = 710$			$P(N,U) = 71$			$P(N,N) = 1,885$		
Joint auxiliary data	Lower	Upper	Width	Lower	Upper	Width	Lower	Upper	Width
(Z_1, Z_2)	381	1,086	705	0	131	131	1,505	2,216	711
(Z_1, Z_2, Z_t, Y_t)	400	835	435	2	119	117	1,754	2,199	445
Y_t	0	880	880	0	151	151	1,721	2,630	909

information about the target distribution of (Y_1, Y_2) in the register proxy (Z_1, Z_2) than in the survey proxy Y_t, as witnessed by the widths of the uncertainty bounds. In other words, there is more information in the concurrent proxy variables that are of a different definition than in the proxy variable that has the same definition but is from a different reference time point. Although the actual figures in Table 3 are obtained on a synthetic dataset, the basic results appear to reinforce the message that in data fusion one should strive to make use of all available auxiliary data.

5. Summary

The usefulness of proxy variables for categorical data fusion is considered above. A measure of the relative efficiency with and without proxy (or other auxiliary) variables is proposed. In practice, the uncertainty analysis must also take into account the sampling uncertainty in cases where the identification uncertainty bounds are unknown and need to be estimated. A flexible technique of distribution calibration is introduced for making use of proxy variables, which can be useful for constructing the fusion distribution as well as the fusion dataset. Empirical results demonstrate that proxy variables can play two beneficial roles at the same time: not only do they provide a general means for reducing the uncertainty associated with data fusion, they also widen the scope of plausible pseudo estimates of the target joint distribution.

6. References

Brozzi, A., A. Capotorti, and B. Vantaggi. 2012. "Incoherence Correction Strategies in Statistical Matching." *International Journal of Approximate Reasoning* 53: 1124–1136. Doi: http://dx.doi.org/10.1016/j.ijar.2012.06.009.

Conti, P.L., D. Marella, and M. Scanu. 2008. "Evaluation of Matching Noise for Imputation Techniques Based on Nonparametric Local Linear Regression Estimators." *Computational Statistics & Data Analysis* 53: 354–365. Doi: http://dx.doi.org/10.1016/j.csda.2008.07.041.

Conti, P.L., M. Di Zio, D. Marella, and M. Scanu. 2009. "Uncertainty Analysis in Statistical Matching." *Paper given at the First Italian Conference on Survey Methodology (ITACOSM09), June 10–12, 2009, Siena*

Conti, P.L., D. Marella, and M. Scanu. 2012. "Uncertainty Analysis in Statistical Matching." *Journal of Official Statistics* 28: 69–88.

Conti, P.L., D. Marella, and M. Scanu. 2013. "Uncertainty Analysis for Statistical Matching of Ordered Categorical Variables." *Computational Statistics & Data Analysis* 68: 311–325. Doi: http://dx.doi.org/10.1016/j.csda.2013.07.004.

Cain, M. 1994. "The Moment-generating Function of the Minimum of Bivariate Normal Random Variables." *The American Statistician* 48: 124–125. Doi: http://dx.doi.org/10.1080/00031305.1994.10476039.

Chambers, R.L. and R.G. Steel. 2001. "Simple Methods for Ecological Inference in 2 x 2 Tables." *Journal of the Royal Statistical Society Series A* 164: 175–192. Doi: http://dx.doi.org/10.1111/1467-985X.00195.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006a. "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints." *Journal of Official Statistics* 22: 137–157.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006b. *Statistical Matching: Theory and Practice*. Chichester: Wiley.

Kadane, J.B. 1978. "Some Statistical Problems in Merging Data Files." *In 1978 Compendium of Tax Research*, (pp. 159–171). Washington, D.C. Department of Treasury. (Reprinted in Journal of Official Statistics 17: 423–433.).

King, G. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

Koopmans, T. 1949. "Identification Problems in Economic Model Construction." *Econometrica* 17: 125–144. Doi: http://dx.doi.org/10.2307/1905689.

Lindley, D.V., A. Tversky, and R.V. Brown. 1979. "On the Reconciliation of Probability Assessments (incl. discussions)." *Journal of the Royal Statistical Society Series A* 142: 146–180. Doi: http://dx.doi.org/10.2307/2345078.

Manski, C.F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Marella, D., P.L. Conti, and M. Scanu. 2008. "On the Matching Noise of Some Nonparametric Imputation Procedures." *Statistics and Probability Letters* 78: 1593–1600. Doi: http://dx.doi.org/10.1016/j.spl.2008.01.020.

Moriarity, C. and F. Scheuren. 2001. "Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure." *Journal of Official Statistics* 17: 407–422.

Nadarajah, S. and S. Kotz. 2008. "Exact Distribution of the Max/Min of Two Gaussian Random Variables." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 16: 210–212. Doi: http://dx.doi.org/10.1109/TVLSI.2007.912191.

Okner, B.A. 1972. "Constructing a New Microdata Base From Existing Microdatasets: the 1966 Merge File." *Annals of Economic and Social Measurement* 1: 325–342.

Patel, J.K., C.H. Kapadia, and D.B. Owen. 1976. *Handbook of Statistical Distributions*. New York: Marcel Dekker.

Plackett, R.L. 1977. "The Marginal Totals of a 2 x 2 Table." *Biometrika* 64: 37–42. Doi: http://dx.doi.org/10.1093/biomet/64.1.37.

Purcell, N.J. and L. Kish. 1980. "Postcensal Estimates for Local Areas (or Domains)." *International Statistical Review* 48: 3–18. Doi: http://dx.doi.org/10.2307/1402400.

Rässler, S. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*, Vol. 168 of Lecture Notes in Statistics. New York: Springer Verlag.

Rässler, S. and H. Kiesl. 2009. "How Useful Are Uncertainty Bounds? Some Recent Theory With an Application to Rubin's Causal Model." *In Proceedings of the 57th Sessions of the International Statistical Institute.* (2009) CD-ROM. Durban, South Africa.

Singh, A.C., H. Mantel, M. Kinack, and G. Rowe. 1993. "Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption." *Survey Methodology* 19: 57–79.

Vantaggi, B. 2008. "Statistical Matching of Multiple Sources: A Look Through Coherence." *International Journal of Approximate Reasoning* 49: 701–711. Doi: http://dx.doi.org/10.1016/j.ijar.2008.07.005.

Wakefield, J. 2004. "Ecological Inference for 2 x 2 Tables (incl. discussions)." *Journal of the Royal Statistical Society Series A* 167: 385–445. Doi: http://dx.doi.org/10.1111/j.1467-985x.2004.02046.x.

Received July 2013
Revised August 2015
Accepted September 2015

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 809–812, http://dx.doi.org/10.1515/JOS-2015-0046

Book Review

Carina Cornesse[1] *and Annelies G. Blom*[2]

Mario Callegaro, Reginald P. Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, Paul J. Lavrakas (Eds). *Online Panel Research: A Data Quality Perspective*. 2014. Chichester, UK: John Wiley and Sons. ISBN: 978-1-119-94177-4, 508 pp., £55.00.

Online Panel Research. A Data Quality Perspective by Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas is an edited volume that brings together state-of-the-art findings on various aspects of online panel research. It presents evidence on a diverse set of research questions on detecting and correcting for different kinds of errors arising in online panels. The book also gives advice on practical aspects of conducting online panels and new developments regarding web panel software.

The book is a valuable addition to and extension of the existing literature on online surveys. Other books in this area typically focus on survey design and on the practical implementation of web surveys (see for example Couper 2008 and Tourangeau et al. 2013). *Online Panel Research* is different in two main respects: firstly, while previous literature has a broad focus on all kinds of web survey research, this book concentrates exclusively on survey methodological research on online panels; secondly, this book focuses particularly on errors and biases in online panels.

In structural terms the book follows a Total Survey Error logic (see Groves et al. 2009). It is a compact collection of findings on the most important issues in online panel research including studies from various countries. The book is very comprehensive and highly instructive for survey methodological research, and is particularly valuable for survey practitioners either already conducting or still aiming to build an online panel.

However, there are some caveats that the reader should be aware of: first, the book is generally written from a commercial data collection rather than an academic perspective. This becomes apparent in the language used in several chapters and section introductions throughout the book where "customer" and "client" interests are emphasized and survey "companies" (see p. 9) are addressed. Second, most chapters apply a limited definition of representativeness, that is, the authors assume that online panels need only be representative of the online population. The reason for this might be that most online panels simply do not include non-Internet users. Some probability-based online panels aim to be representative of the general population and include previously offline persons (so-called offliners) by providing them with the necessary equipment. This aspect of increasing representativeness by including offliners is not discussed in the book, not even

[1] Collaborative Research Center "Political Economy of Reforms" (SFB 884), University of Mannheim, Germany. Email: carina.cornesse@uni-mannheim.de
[2] Department of Political Science, School of Social Sciences, University of Mannheim, Germany

in Chapter 2, where panels with and without the inclusion of offliners are compared in terms of their data quality (see p. 26).

The book is structured logically. It begins with a general introduction followed by one section each on coverage, nonresponse, measurement error, weighting adjustments, special domains (such as smartphone usage in online panels), and operational issues (such as online panel software). Each section contains a short introduction written by the editors of the book. In the following, we briefly discuss each section in turn.

The general introduction contains a brief overview of topics and steps important in online panel research. It lists state-of-the-art findings with additional references to more detailed literature. This section consists of two chapters. The first, written by the editors of the book, is especially helpful regarding the collection of standards, associations' guidelines, and advisory groups presented. Chapter 2 by Callegaro et al. provides a detailed overview of studies comparing online panels to other panels and benchmark surveys. This chapter also offers a rich typology of comparison studies on data quality (in particular on measurement error) and provides a range of examples.

The coverage section contains Chapter 3 by Struminskaya et al. and Chapter 4 by Grönlund and Strandberg, which both assess the representativeness of online panels. While Chapter 3 offers valuable and detailed practical insights into the design and implementation of an online panel as well as recent findings on the representativeness of probability-based online panels, Chapter 4 focuses on the effect of panel attrition on the representativeness of panel survey results. Both chapters are highly instructive and transparent regarding the models estimated and the conceptual as well as analytical decisions taken. In Chapter 5, McCutcheon et al. provide the results of a survival analysis model of members in a multimode consumer panel. The analysis is very easy to follow, especially because of the helpful graphical presentation of results.

The nonresponse section of the book is very diverse in terms of the questions raised and the methods used to assess nonresponse. In Chapter 6, Lugtig et al. present an instructive latent class analysis to investigate the different behavioral patterns involved in panel attrition. Göritz presents results of logistic regression analyses, including hypotheses on and indicators of survey nonresponse, in Chapter 7. All variables are examined regarding their influence on the starting propensity as well as the completion propensity of a survey wave within a panel. In Chapter 8, Keusch et al. provide insight into the motives and value characteristics of participants in a nonprobability online panel. Among other findings, they show how important both intrinsic and extrinsic motivation is to participation, concluding that both types of motives need to be addressed and encouraged by the panel provider. In Chapter 9, Scherpenzeel and Toepoel present various experimental studies to assess the effect of nonmonetary incentives and encouragement strategies on panel participation. They conclude that survey practitioners should not take for granted that feedback and small acknowledgments have a significant positive effect on panel retention.

In the measurement error section, Hillygus et al. in Chapter 10 provide evidence on a wide range of indicators concerning the response behavior of professional respondents, that is, respondents participating in multiple panels. In Chapter 11, Greszki et al. focus on the magnitude and intensity of the effect that speeders have on data quality. Unfortunately, this study compares two panels which differ from one another in more than one respect, which limits the generalizability of their results. Both chapters offer rich descriptions of the underlying theories and methods used to assess measurement error.

The chapters on measurement error are followed by a section on weighting adjustments. In Chapter 12, Steinmetz et al. take a very thoughtful and critical view of propensity-weighting adjustments. They show the advantages and challenges of using reference surveys to calculate propensity weights for nonprobability panels. Their description of the process of applying weights is very detailed and easy to understand. However, they generalize their results on the representativeness of one specific panel (the Dutch WageIndicator Survey) to all nonprobability panels, although their panel may well attract a very specific group of panelists (see p. 286). Chapter 13 by Zhang gives an overview of imputation approaches and their advantages and disadvantages. Zhang gives detailed instructions on when and how to use imputations, as well as providing interesting insights into the impact of such imputation procedures on the representativeness of results (see p. 305).

The next section contributes to understanding how nonresponse and measurement error interact. The analyses in Chapter 14 by Malhotra et al. and Chapter 15 by Roberts et al. complement one another as they both look at the interdependence between nonresponse error and measurement error. In particular, they study the effects of nonresponse reduction in the recruitment phase. Chapter 14 focuses on the comparison between hard-to-recruit and easy-to-recruit respondents and their response behavior. The authors of this chapter use various different indicators of measurement error (pp. 326). Chapter 15 looks at the long-term effects of nonresponse reduction strategies. Their findings on the correlation between recruitment effort and conditioning during the later panel waves are particularly informative (p. 356).

The special domains section of this book consists of two very different chapters. Drewes in Chapter 16 presents interesting findings about smartphone users, their attitudes towards smartphones and web surveys, and the differences in their response behavior compared to users of conventional web devices, such as PCs and laptops. In Chapter 17, Napoli et al. report alarming facts concerning the history and development of Internet ratings panels (see e.g., pp. 388). Internet ratings panels systematically collect data on their participants' online behavior. They use special hardware and software to capture Internet usage patterns directly. The authors of this chapter conclude that to date the findings reported by these panels are not reliable and not representative of the online population (see p. 402) and frequently collect very sensitive information without explicitly informing their panelists (see for example p. 389 and p. 397).

The last section of the book covers new procedures for solving practical problems involved in conducting online panels. In Chapter 18, Macer provides information on recent developments regarding web panel software. The chapter covers software solutions for the complete survey process, from questionnaire development and panel management to monitoring panelists. In the subsequent chapter, Baker et al. provide evidence on the effectiveness of procedures to validate panelists' identities. Although the authors show that unvalidated respondents tend to produce data with a little lower quality than validated respondents, the authors conclude that respondent validation may lead to smaller and less representative samples without the answer quality being substantively better (see p. 450).

Therefore every researcher has to decide for themselves whether these procedures can and should be applied to their specific online panel project.

This edited volume serves as a very valuable introduction to online panel research, since it provides comprehensive information on the definitions, typologies, guidelines, and basic formulae necessary for starting research on online panels as well as building a new online panel. Our primary criticisms concern the definitions of representativeness adopted and the commercial perspective portrayed in some of the chapters. Nonetheless, the book is an important addition to the survey methodological literature, because it offers state-of-the-art research in the field of online panel research. We thus highly recommend this book to academic survey methodologists and practitioners in the field of online panels alike.

References

Couper, M. 2008. *Designing Effective Web Surveys*. New York: Cambridge University Press.

Groves, R.M., F.J. Fowler, Jr., M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.

Tourangeau, R., F. Conrad, and M. Couper. 2013. *The Science of Web Surveys*. Oxford: Oxford University Press.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 813–815, http://dx.doi.org/10.1515/JOS-2015-0047

Book Review

Mariano Ruiz Espejo[1]

Richard Valliant, Jill A. Dever and Frauke Kreuter. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer, 2013. ISBN 978-1-4614-6448-8, 670 pp. $68.34.

This book is directed at students, survey statisticians, social scientists, and other survey practitioners, presenting statistical thought and steps taken to design, select, and weight random survey samples. Following a first chapter on "An Overview of Sample Design and Weighting", which contains the background and the basic terminology used, the book is divided into four parts: I: Designing Single-Stage Sample Surveys (Ch. 2-7), II: Multistage Designs (Ch. 8-11), III: Survey Weights and Analyses (Ch. 12-16), and IV: Other Topics (Ch. 17-18).

Parts I-III describe examples of projects similar to those that might be encountered in practice. After introducing each project, the authors present the tools for accomplishing their work in the subsequent chapters. The last chapters in Parts I-III, Chapters 7, 11, and 16, provide one way of meeting the goals of the example project but with solutions that are not unique. The authors explain that "there are likely to be many ways of designing a sample and creating weights that will, at least approximately, achieve the stated goals... Practitioners need to be comfortable with the solutions they propose. They need to be able to defend decisions made along the way and understand the consequences that alternative design decisions would have. This book will prepare you for such tasks."

Part I addresses techniques that are valuable in designing single-stage samples. Chapter 2 presents a straightforward project to design a personnel survey. The subsequent chapters concentrate on methods for determining the sample size and allocating it among different groups in the population. Chapter 3 presents a variety of ways of calculating a sample size to meet stated precision goals for estimates for the full population. Chapter 4 covers various methods of computing sample sizes based on power requirements, which is common in epidemiological applications when the goal is to find a sample size that will detect with a high probability some prespecified difference in means, and so on, between subgroups or between groups at two different time periods. All of these goals substitute inference criteria or make deliberate use of approximations that can be seen as pragmatic or arbitrary and not scientific criteria or principles (for these, see Cochran 1977; Ruiz Espejo 1986, 1987, 2013).

Chapters 3 and 4 focus on sample-size decisions made based on optimizing precision and power for one single variable at a time. To meet multiple goals and respect cost constraints, the authors suggest that the methods in Chapters 3 and 4 could be applied by

[1] Catholic University of Murcia, Avenida Jerónimos 135, 301 07 Guadalope, Murcia, Spain. Email: mruiz033@alu.ucam.edu

trial and error in the hopes of finding an acceptable solution. A better approach is to use mathematical programming techniques that allow optimization across multiple variables.

Chapter 5 presents some multicriteria programming methods that can be used to solve these more complicated problems. These algorithms are better known to operations researchers and management scientists than to survey statisticians, and they allow more realistic treatment of complicated allocation problems involving multiple response variables and constraints on costs, precision, and sample sizes for subgroups. In Chapter 6, adjustments need to be made to the initial sample size to account for such circumstances.

Part II concerns the design of clustered samples in order to efficiently collect data, and therefore sample-design decisions are required in multiple stages. Chapter 8 begins with a moderately complex project to design an area sample and allocate units to geographic clusters in such a way that the size of the sample of persons is controlled relative to some important demographic groups. Chapters 9 and 10 cover the design of samples of those geographic clusters. Chapter 11 gives a solution to the area sample design.

Part III discusses the computation of survey weights and their use in some analyses. Chapter 12 begins with a project on calculating weights for a personnel survey, like the one designed in Project 1 of Chapter 2. Chapters 13 and 14 describe the steps for calculating base weights, making adjustments for ineligible units, nonresponse, and other sample losses, and for using auxiliary data to adjust for deficient frame coverage and to reduce variances. Some of the important techniques for using auxiliary data are the general regression estimator and calibration estimation, which provide biased estimators and their variances are not usually unbiasedly estimable. But software is now available to do some of the computations. Chapter 13 sketches the rationale behind the nonresponse weight-adjustment methods, which requires thinking about models for response and other methods that omit some units. Applications of calibration estimation, including poststratification, raking, and general regression estimation are covered in Chapter 14. More discussion of objective unbiased variance estimation could have been included here (Ruiz Espejo et al. 2006; Ruiz Espejo 2013, 2015). Weight trimming using quadratic programming and other more *ad hoc* methods are also dealt with in this chapter. Chapter 15 covers the major approaches to variance estimation in surveys. Chapter 16 gives a solution to weighting the personnel survey.

Part IV covers the specialized topics of multiphase sampling (Ch. 17) and quality control (Ch. 18).

My opinion of the book is that, while it does not resolve inference problems which arise in survey sampling theory, it does provide pragmatic ideas and solutions in applied designs and statistical weighting in sample surveys. Many examples of useful code written in R are provided throughout the book. The book is oriented towards practice but without the developed pure science behind it. For this reason, I believe that it is more useful for survey statisticians and social survey practitioners interested in practical solutions in the survey design than those interested in development of sampling theory.

References

Cochran, W.G. 1977. *Sampling Techniques*, 3rd ed. New York, NY: Wiley.

Ruiz Espejo, M. 1986. “Estimable Parametric Functions in Sampling Theory.” *Estadística Española* 28(112–113): 69–73.

Ruiz Espejo, M. 1987. "On UMV and UMMSE Estimators in Finite Populations." *Estadística Española* 29(115): 105–111.

Ruiz Espejo, M. 2013. *Exactness of Inference in Finite Populations*. Madrid: Bubok.

Ruiz Espejo, M. 2015. "Objective Unbiased Estimation for Nonresponse." *Estadística Española*, 57(186): 29–37.

Ruiz Espejo, M., M. Delgado Pineda, and H.P. Singh. 2006. "Postgrouped Sampling Method of Estimation." *Test* 15: 209–226.

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 817–819, http://dx.doi.org/10.1515/JOS-2015-0048

Book Review

Timothy Michael Mulcahy[1]

Louise Corti, Veerle Van den Eynden, Libby Bishop, and Mathew Woollard (eds). *Managing and Sharing Research Data: A Guide to Good Practice*. 2014. Los Angeles, London, New Delhi, Singapore, Washington, DC: SAGE Publications. 222 pp., ISBN 978-1-4462-67264, £25.99.

Rapid technological changes are reshaping the way we discover, access, and manipulate data and are spawning new and innovative processes that facilitate the effective management and exchange of information between agencies, independent of their geographic location or proprietary infrastructure (Heus and Gregory 2010). Increasingly, however, these changes call into question traditional statistical data management approaches and systems. As government and private-sector data producers move away from traditional survey approaches (for example, 'design-collect-analyze-publish') toward more integrated structures, the challenge for producers of statistics is to exploit recent technological advances to develop new and innovative ways to manage, access, process, discover, and visualize data (Lorenc et al. 2013). Old paradigms need to be revisited, prior assumptions reviewed, and traditional research methods re-evaluated and updated as appropriate.

Corti el al.'s (2014) book, *Managing and Sharing Research Data: A Guide to Good Practice*, addresses these daunting challenges. This clearly written, easy-to-follow handbook highlights best practices in research data management and sharing. However, it is not simply an anthology of data management and data sharing best practices, it is a comprehensive and contemporary primer on nearly every aspect of the research process firmly rooted in the research data lifecycle (Humphrey 2006).

The book is organized into eleven related chapters and applies to a wide range of audiences, covering topics such as management and sharing of data, the research data lifecycle, data management planning, documenting and providing context for data, formatting and organizing data, storing and transferring data, legal and ethical issues, rights relating to research data, strategies for collaborative research, secondary research, publishing, and citing research data. The book is particularly timely in that funding agencies around the globe are formulating new laws, policies, and procedures that strongly encourage, if not require, researchers and grantees to provide data management plans or strategies as an integrated component of their research proposals. While these efforts are sound and aim to encourage knowledge sharing, collaboration, and open access to publicly-funded research data, they also raise new questions and challenges for data producers and end users that are adeptly addressed in this practical handbook.

[1] NORC at the University of Chicago, Bethesda, MD, U.S.A. mulcahy-tim@norc.org

In Chapter 1 the authors introduce a recurring theme that permeates the book: the need to ensure high quality, sustainable research in a responsible and efficient manner, and to do so in a way that meets the replication standard (King 1995) and ensures the ability to share and reuse research data in perpetuity. The authors also highlight the benefits of managing and sharing research data effectively, noting concerns voiced by some researchers, and providing tangible examples of successfully implemented data management plans. Chapter 2 notes the critical importance of properly indexing, archiving, and curating data to facilitate future uses and reuses of the data. Readers are encouraged to utilize a data lifecycle approach, one that is at once specific to the detailed nuances of each stage of the research process while also generic enough to apply more widely to disciplines outside of social science.

In Chapter 3, the authors emphasize the critical importance of early planning and formulating data management plans to help design and implement successful research efforts, ensure that adequate resources are in place, and clearly articulate, assign, and manage individual roles and responsibilities. Examples are provided from the UK and US for planning, documenting, formatting, storing, confidentiality, ethics, consent, copyright, and sharing.

Chapters 4–8 directly address the myriad challenges involved in data sharing and provide tangible strategies to overcome these limitations. Chapter 4 emphasizes that data alone are of limited usefulness and that they must be accompanied by proper documentation and context. Data are not simply collected to serve our current needs, they are meant to be preserved in perpetuity to allow future use and reuse. Chapters 5 and 6 extend this theme to the importance of formatting and organizing data for long-term use, storage, and transfer. Chapter 7 discusses the legal and ethical frameworks involved in data sharing and shares valuable insights into the process of obtaining informed consent (including unknown future uses), modern techniques for perturbing and anonymizing data for safe use, and regulating access to data. Chapter 8 focuses on rights related to research data, and provides an in-depth discussion on Intellectual Property Rights in data, including copyrights and exemptions, database rights, freedom of information, and licensing.

In Chapter 9, the authors point out that collaborative research is becoming an increasingly common phenomenon and caution that this newly emerging paradigm presents particular challenges to ensuring high-quality, sustainable, and reusable research. Although collaborative research heightens the need to develop and implement sound data management plans, the authors provide effective strategies to assist in this regard, for example, developing standard procedures, protocols, and policies, and clearly assigning roles and responsibilities across collaborating entities.

Chapter 10 reinforces the notion that archived data at some later point may become important historic research materials and discusses the opportunities and limitations to conducting secondary analyses. On the one hand, respondent burden is reduced, data linking is made possible, and new data sets and derivative products may be created. In addition, reanalyzing data facilitates reinterpretation and allows new questions to be asked of archived data. This provides a mechanism for validating and replicating prior work – a critical check and balance on policy and programming decisions derived from earlier findings. On the other hand, the authors caution that truly replicating historical research is hardly a trivial task, if possible at all, as studies rarely canvass identical social phenomena.

The final chapter discusses the nuances involved in publishing and citing research data, yet another area that has borne witness to dramatic changes in recent years. The authors conclude by providing rich examples of data centers, institutional data repositories, traditional and digital data archives, and associations that are on the cutting edge of publishing and data citing.

In the main, practitioners and students will find this book to be extremely valuable. Written by and for researchers, professional researchers and students alike would do well to keep a copy of this guide book close at hand. Each chapter is self-contained, allowing readers to peruse individual sections to learn more and/or refresh their knowledge and understanding of a particular issue or topic of interest and is accompanied by a comprehensive bibliography that serves as a useful guide for further research. Finally, it provides empirical case studies and "hands-on" exercises and activities that drive home the core concepts of the modern-day research process.

References

Heus, P. and A. Gregory. 2010. Maximizing the Potential of Data – Modern IT Tools, Best Practices, and Metadata Standards for SBE Sciences. Open Data Foundation. Version 1.0.

Humphrey, C. 2006. "e-Science and the Life Cycle of Research."

King, G. 1995. "Replication, Replication." *Ps: Political Science and Politics* 28: 443–499. Available at: http://j.mp/jCyfF1.

Lorenc, B., P. Biemer, I. Jansson, J. Eltinge, and A. Holmberg. 2013. "Prelude to the Special Issue on Systems and Architectures for High-Quality Statistics Production." *Journal of Official Statistics* 29: 1–4. DOI: http://dx.doi.org/10.2478/jos-2013-0012.

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 821–826, http://dx.doi.org/10.1515/JOS-2015-0049

Editorial Collaborators

The editors wish to thank the following referees who have generously given their time and skills to the Journal of Official Statistics during the period October 1, 2014–September 30, 2015. An asterisk indicates that the referee served more than once during the period.

Aizcorbe*, Ana, Virginia Bioinformatics Institute, Arlington, VA, U.S.A.
Alam, Moudud, Dalarna University, Borlänge, Sweden
Alwin, Duane F., Pennsylvania State University, University Park State College, PA, U.S.A.
Andersson, Per Gösta, Stockholm University, Stockholm, Sweden
Antoni, Manfred, Institute for Employment Research, Nuremberg, Germany
Antoun, Christopher, University of Michigan, Ann Arbor, MI, U.S.A.
Baffour*, Bernard, University of Queensland, Brisbane, Australia
Baker, Reginal, Market Strategies Inc. Ann Arbor, MI, U.S.A.
Bakker, Bart F.M., Statistics Netherlands, Den Haag, Netherlands
Bauer*, Cici Chen, Brown University, Providence, RI, U.S.A.
Bauer, Johannes, IAB, Nuremberg, Germany
Bautista, René, NORC, Chicago, IL, U.S.A.
Bavdaž, Mojca, University of Ljubljana, Ljubljana, Slovenia
Beaumont*, Jean-Francois, Statistics Canada, Ottawa, U.S.A.
Benedetti, Roberto, University of Chieti Pescara, Pescara, Italy
Bergsma, Wicher P., London School of Economics and Political Science, London, UK
Bianconcini, Silvia, University of Bologna, Bologna, Italy
Biffignandi, Silvia, University of Bergamo, Bergamo, Italy
Bijak*, Jakub, University of Southampton, Southampton, UK
Bilgen*, Ipek, NORC, University of Chicago, Chicago, IL, U.S.A.
Bivand, Roger, Norwegian School of Economics, Bergen, Norway
Blasius*, Jörg, University at Bonn, Bonn, Germany
Blom, Annelies G., University of Mannheim, Mannheim, Germany
Booth, Tom, University of Edinburgh, Edinburgh, UK
Bosnjak, Michael, GESIS Leibniz-Institute for the Social Sciences, Mannheim, Germany
Brandolini, Andrea, Bank of Italy, Rome, Italy
Bredl*, Sebastian, Justus Liebig University, Giessen, Germany
Bregar, Lea, University of Ljubljana, Ljubljana, Slovenia
Brick, Michael, Westat, Rockville, MD, U.S.A.
Buelens, Bart, Statistics Netherlands, Heerlen, Netherlands
Bycroft*, Christine, Statistics New Zealand, Christchurch, New Zealand
Callegaro, Mario, Google, London, UK
Charest, Anne-Sophie, Université Laval, Quebec, Canada
Chen, Qixuan, Columbia University, New York, NY, U.S.A.
Chipperfield*, James Oliver, Australian Bureau of Statistics, Belconnen, Australia
Chowdhury, Sadeq, Agency for Healthcare Research and Quality, Rockville, MD, U.S.A.
Coleman, Shirley, Newcastle University, Newcastle upon Tyne, UK
Cornelis, Éric, University of Namur, Namur, Belgium

Crawford, Forrest, Yale University, New Haven, CT, U.S.A.
Creel, Darryl V., RTI, Rochville, MD, U.S.A.
Crequer, John, Statistics New Zealand, Christchurch, New Zealand
Cruyff, Maarten, University Utrecht, Utrecht, Netherlands
Curtin, Richard, University of Michigan, Ann Arbor, MI, U.S.A.
Cyr, André, Statistics Canada, Ottawa, Canada
Dale, Trine, TNS Gallup, Oslo, Norway
Dalla Valle, Luciana, Plymouth University, Plymouth, UK
Davidson, Russell, McGill University, Montreal, Canada
De Guili, Elena, University of Pavia, Pavia, Italy
Delden van*, Arnout, Statistics Netherlands, The Hague, Netherlands
Demirhan, Haydar, Hacettepe University, Ankara, Turkey
Deutsch, Tomi, University of Ljubljana, Ljubljana, Slovenia
Di Consiglio, Loredana, Istat, Rome, Italy
Diallo, Mamadou, Westat, Rockville, MD, U.S.A.
Dixon, John, U.S. Bureau of Labor Statistics, Washington, DC, U.S.A.
Dolson, David, Statistics Canada, Ottawa, Canada
Dominitz, Jeff, Resolution Economics Group, Beverly Hills, CA, U.S.A.
Donze*, Laurent. University of Fribourg, Fribourg, Switzerland
Dreassi, Emanuela, University of Florence, Florence, Italy
Durand, Claire, University of Montreal, Montreal, Quebec, Canada
Durrant*, Gabrielle, University of Southampton, Southampton, UK
Dykema, Jennifer, University of Wisconsin-Madison, Madison, WI, U.S.A.
Edwards, Michelle, Texas Christian University, Fort Worth, TX, U.S.A.
Elezovic, Suad, Statistics Sweden, Stockholm, Sweden
English, Ned, NORC, University of Chicago, Chicago, IL, U.S.A.
Erkens, Greg, US Bureau of Labor Statistics, Washington, DC, U.S.A.
Fabrizi*, Enrico, Catholic University, Piacenza, Italy
Farčnik, Daša, University of Ljubljana, Ljubljana, Slovenia
Foschi, Flavio, Istat, Rome, Italy
Fowler, Floyd Jackson Jr, University of Massachusetts, Boston, MA, U.S.A.
Frazis, Harley, U S Bureau of Labor Statistics, Washington, DC, U.S.A.
Freiman*, Michael, U.S. Census Bureau, Washington, DC, U.S.A.
Frey, Jesse C., Villanova University, Villanova, PA, U.S.A.
Fricker, Scott S., U.S. Bureau of Labor Statistics, Washington, DC, U.S.A.
Gargiulo, Floriana, LISC, France
Garner, Thesia, U.S. Bureau of Labor Statistics, Washington, DC. U.S.A.
Gibson, John, University of Waikato, Hamilton, New Zealand
Giesen, Deirdre, Statistics Netherlands, Heerlen, Netherlands
Giusti, Caterina, University of Pisa, Pisa, Italy
Glorieux, Ignace, Vrije University Brussel, Brussel, Belgium
Golinelli, Daniela, RAND, Santa Monica, CA, U.S.A.
Griffin, Richard, U.S. Census Bureau, Washington, DC, U.S.A.
Groen*, Jeffrey A., U.S. Bureau of Labor Statistics, Washington, DC, U.S.A.
Gulyá, Ágnes, Canterbury Christ Church University, Canterbury, UK
Göritz, Anja, University of Freiburg, Freiburg, Germany
Haan*, Jan de, Statistics Netherlands, The Hague, Netherlands
Handwerker, Elizabeth Weber, Bureau of Labor Statistics, Washington, DC, U.S.A.

Harel, Ofer, University of Connecticut, Storrs, CT, U.S.A.
Haslett, Stephen J., Massey University, Manawatu, New Zealand
Haunberger*, Sigrid, University of Applied Sciences, Olten, Switzerland
Heckathorn, Douglas D., Cornell University, Ithaca, NY, U.S.A.
Himelein, Kristen, World Bank, Washington, DC, U.S.A.
Hogan, Howard R., U.S. Census Bureau, Washington, DC, U.S.A.
Israel, Glenn D., University of Florida, Gainesville, FL, U.S.A.
Jans, Matt, University of Los Angeles, Los Angeles, CA, U.S.A.
Jiongo, Valery Dongmo, Statistics Canada, Ottawa, Canada
Juriová, Jana, Infostat, Bratislava, Slovakia
Kabzinska, Ewa, University of Southampton, Southampton, UK
Karon, John M., Emergint Corporation, Louisville, KY, U.S.A.
Kao, Fei-Fei, Ming Chuan University, Taipei, Taiwan
Kavonius*, Ilja Kristian, European Central Bank, Frankfurt am Main, Germany
Keegan, Alan, Statistics New Zealand, Wellington, New Zealand
Kemper, Christoph, University of Luxembourg, Luxembourg
Kim, Hang J, Duke University, Durham, NC, U.S.A.
Kim*, Jae-Kwang, Iowa State University, Ames, IA, U.S.A.
King, Thomas, Newcastle University, Newcastle upon Tyne, UK
Kinney*, Satkartar, NISS, Research Triangle Park, NC, U.S.A.
Krapavickaite, Danute, Vilnius Gediminas Technical University, Vilnius, Lithuania
Krenzke*, Thomas R., Westat, Rockville, MD, U.S.A.
Kuhn, Ursina, FORS, University of Lausanne, Lausanne, Switzerland
Kuusela*, Vesa, University of Helsinki, Helsinki, Finland
Larose, Chantal, University of Connecticut, Storrs, CT, U.S.A.
Larsen*, Michael D., George Washington University, Rockville, MD, U.S.A.
Laurie, Heather, ISER, University of Essex, Colchester, UK
Lavrakas*, Paul J., NORC at the University of Chicago, Chicago, IL, U.S.A.
Lee, Sunghee, University of Michigan, Ann Arbor, MI, U.S.A.
Li, Feng, Stockholm University, Stockholm. Sweden
Liberts, Mārtiņš, Central Statistical Bureau of Latvia, Riga, Latvia
Liu, Benmei, National Institutes of Health, Rockville, MD, U.S.A.
Liu, Mingnan, Survey Monkey, Palo Alto, CA, U.S.A.
Macchia, Stefania, ISTAT, Rome, Italy
MacFeely, Steve, Central Statistics Office, Cork, Ireland
Malmros*, Jens, Stockholm University, Stockholm, Sweden
Manrique-Vallier, Daniel, Indiana University, Bloomington, IN, U.S.A.
Maples, Jerry J., U.S. Census Bureau, Washington, DC, U.S.A.
Martin*, Peter, Anna Freud Centre, London, UK
Mathä*, Thomas, The Central Bank of Luxembourg, Luxembourg
Mavletova, Aigul, National Research University, Moscow, Russia
Measure, Alexander, Bureau of Labor Statistics, Washington DC, U.S.A.
Meekins*, Brian J., U.S. Bureau of Labor Statistics, Washington DC, U.S.A.
Meng, Xiao-Li, Harvard University, Cambridge, MA, U.S.A.
Menold, Natalja, GESIS, Mannheim, Germany
Messer*, Benjamin, Research Into Action, Portland, OR, U.S.A.
Miller, Peter V., US Census Bureau, Washington, DC, U.S.A.
Mittag, Nikolas, CERGE-EI, Prague, Czech Republic

Mohler, Peter Ph., University of Mannheim, Mannheim, Germany
Morren, Meike, University of Amsterdam, Amsterdam, Netherlands
Moon, Nick, GFK NOP Social Research, London, UK
Moors, Guy, Tilburg University, Tilburg, Netherlands
Moura, Fernando A. da Silva, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
Muennich, Ralf Thomas, University of Trier, Trier, Germany
Mulrow, Edward, NORC at the University of Chicago, Bethesda, MD, U.S.A.
Murphy, Joe, RTI International, Research Triangle Park, NC, U.S.A.
Mussard, Stéphane, University of Montpellier, Montpellier, France
Nagaraja, Chaitra, Fordham University, New York, NY, U.S.A.
Nascimento Silva do, Pedro, IBGE, Rio de Janeiro, Brazil
Nandram, Balgobin, Worcester Polytechnic Institute, Worcester, MA, U.S.A.
Niedomysl, Thomas, Lund University, Lund, Sweden
Norberg, Anders, Statistics Sweden, Stockholm, Sweden
Nuijten, Michèle, Tilburg University, Tilburg, Netherlands
Oganyan, Anna, Georgia Southern University, Statesboro, GA, U.S.A.
O'Hara*, Kieron, University of Southampton, Southampton, UK
Ostasiewicz, Katarzyna, Wrocław University of Economics, Wroclaw, Poland
Padilla, Alberto, Bank of Mexico, Mexico
Pahor, Marko, University of Ljubljana. Ljubljana, Slovenia
Pastor, Manuel, University of Southern California, Los Angeles, CA, U.S.A.
Perales, Francisco, University of Queensland, Brisbane, Australia
Petocz, Peter, Macquarie University Statistics, North Ryde, New South Wales, Australia
Pforr, Klaus, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany
Pinheiro, Eliane, University of São Paulo, São Paulo, Brazil
Pisani, Caterina, University di Siena, Siena, Italy
Polettini, Silvia, University of Rome, Rome, Italy
Pratesi, Monica, University of Pisa, Pisa, Italy
Puts, Marco J., Statistics Netherlands, Heerlen, Netherlands
Raghunathan, Trivellore E., University of Michigan, Ann Arbor, MI, U.S.A.
Rambaldi, Alicia, University of Queensland, Brisbane. Australia
Raymer, James, Australian National University, Canberra, Australia
Read, Janet, University of Central Lancashire, Lancashire, UK
Redek, Tjaša, University of Ljubljana, Ljubljana, Slovenia
Redline, Cleo D., NCES, Washington, DC, U.S.A.
Reiter, Jerome P., Duke University, Durham, NC, U.S.A.
Reis, Marco, University of Coimbra, Coimbra, Portugal
Revilla, Melanie, Universitat Pompeu Fabra, Barcelona, Spain
Rivest, Louis-Paul, University of Laval, Quebec, Canada
Roberts*, Margaret, University of California, San Diego, CA, U.S.A.
Robison, Edwin L., Bureau of Labor Statistics, Washington, DC, U.S.A.
Rolls, David, University of Melbourne, Melbourne, Australia
Rönning, Gerd, University of Tuebingen, Tuebingen, Germany
Ruspini, Elisabetta, University of Milano-Bicocca, Milan, Italy
Salvati*, Nicola, University of Pisa, Pisa, Italy
Sánchez-Fernández, Juan, University of Granada, Granada, Spain
Santos, Cristiano, IBGE, Rio de Janeiro, Brazil
Schaik van, Paul, Teesside University, Middlesbrough, UK

Scheffer*, Fredrik, Statistics Sweden, Stockholm, Sweden
Schmich, Patrick, Robert Koch Institute, Berlin, Germany
Schober, Michael F., New School for Social Research, New York, NY, U.S.A.
Scholtus, Sander, Statistics Netherlands, The Hague, Netherlands
Schonlau, Matthias, University of Waterloo, Waterloo, Ontario, Canada
Sebastiani, Fabrizio, Qatar Computing Research, Doha, Qatar
Semaan, Salaam, Centers for Disease Control and Prevention, Atlanta, GA, U.S.A.
Shmueli, Galit, Indian School of Business, Hyderabad, India
Sikkel, Dirk, Sixtat, Leidschendam, Netherlands
Silber, Jacques, Bar-Ilan University, Ramat-Gan, Israel
Silver, Mick, IMF, Washington, DC, U.S.A.
Sindoni, Giuseppe, Istat, Rome, Italy
Smith, Paul A., University of Southampton, Southampton, UK
Smith, Peter W.F., University of Southampton, Southampton, UK
Snijkers, Ger, Statistics Netherlands, Heerlen, Netherlands
Spreen, Marinus, Applied University Stenden, Leeuwarden, Netherlands
Stander, Julian, Plymouth University, Plymouth, UK
Stasny*, Elizabeth A., The Ohio State University, Columbus, OH, U.S.A.
Steorts*, Rebecca, Carnegie Mellon University, Pittsburgh, PA, U.S.A.
Stern, Michael, University of Chicago, Chicago, IL, U.S.A.
Stoyanchev, Svetlana, Interactions, New York, NY, U.S.A.
Strandell, Gustaf, Statistics Sweden, Örebro, Sweden
Ståhl, Olivia, University of Stockholm, Stockholm, Sweden
Swires-Hennessy, Ed, Newport, UK
Tang, Cheng Yong, Temple University, Philadelphia, PA, U.S.A.
Tille,Yves, University of Neuchatel, Neuchatel, Switzerland
Tomas, Amber, University of Virginia, Charlottesville, VA, U.S.A.
Toninelli, Daniele, University of Bergamo, Bergamo. Italy
Tuoto, Tiziana, ISTAT, Rome, Italy
Tzavidis, Nikos, University of Southampton, Southampton, UK
Valliant, Richard, University of Maryland, College Park, MD, U.S.A.
Van der Laan, Paul, Statistics Netherlands, The Hague, Netherlands
Vannieuwenhuyze*, Jorre, KU Leuven, Leuven, Belgium
Vehovar, Vasja, University of Ljubljana, Ljubljana, Slovenia
Vilhuber, Lars, Cornell University, Ithaca, NY, U.S.A.
Vincent, Kyle Shane, Bank of Canada, Ottawa, Canada
Wackerow, Joachim, GESIS, Mannheim, Germany
Wagner, James. R., University of Michigan, Ann Arbor, MI, U.S.A.
Watson, Nicole, University of Melbourne, Melbourne, Australia
Weidman, Pheny, United States Department of Agriculture, Washington, DC, U.S.A.
Wejnert*, Cyprian, Centers for Disease Control and Prevention, Atlanta, GA, U.S.A.
Wenemark, Marika, University of Linköping, Linköping, Sweden
West, Brady, University of Michigan, Ann Arbor, MI, U.S.A.
Wieczorek, Jerzy, Carnegie Mellon University, Pittsburgh, PA, U.S.A.
Vieira, Marcel De Toledo, University of Juiz de Fora, Juiz de Fora, Brazil
Willis, Gordon B., NCI, Bethesda, MD, U.S.A.
Yan, Ting, Institute for Social Research, Ann Arbor, MD, U.S.A.
Yao, Jing, University of Glasgow, Glasgow, UK

Zaslavsky, Alan M., Harvard University, Boston, MA, U.S.A.
Zenga, Michele Mario, University of Milano-Bicocca, Milan, Italy
Zhang, Chan, University of Michigan, Ann Arbor, MI, U.S.A.
Žiberna, Aleš, University of Ljubljana, Ljubljana, Slovenia
Zieschang, Kimberly D., IMF, Washington, DC, U.S.A.
Zuell, Cornelia, GESIS, Mannheim, Germany
Zwick, Markus, European Commission, Luxembourg, Luxembourg

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 827–830, http://dx.doi.org/10.1515/JOS-2015-0050

Index to Volume 31, 2015

Contents of Volume 31, Numbers 1–4

Articles, see Author Index
Book Reviews 139, 141, 143, 147, 809, 813, 817
Editorial Collaborators 821
Index 827
Preface 149, 349

Author Index

Alberts, I., Alho, J., Bijak, J., Bryant, J., Buettner, T., Falkingham, J., Forster, J. J., Gerland, P., King, T., Onorante, L., Keilman, N., O'Hagan, A., Owens, D., Raftery, A., Ševčíková, H., and Smith, W.F. Letter to Editor . . . . 537 – 544
Alho, J., See Alberts, I.
Antal, L., See Shlomo, N.
Astleithner, F., See Schnetzer, M.
Bakker, B.F.M., See Gerritse, S.C.
Bakker, B.F.M., van der Heijden, P.G.M., and Scholtus, S. Preface . . . . 349 – 355
Barron, M., Davern, M., Montgomery, R., Tao, X., Wolter, K.M., Zeng, W., Dorell, C., and Black, C. Using Auxiliary Sample Frame Information for Optimum Sampling of Rare Populations . . . . 545 – 557
Bavdaž, M., Giesen, D., Korenjak Černe, S., Löfgren, T., and Raymond-Blaess, V. Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutions . . . . 559 – 588
Black. C. See Barron, M.
Berger, Y.G., See Oguz Alper, M.
Berger Y.G. and Muñoz, J.F. On Estimating Quantiles Using Auxiliary Information . . . . 101 – 119
Biffignandi, S., see Karlberg, M.
Bijak, J., See Alberts, I.
Blackwell, L., Charlesworth, A., and Rogers, N.J. Linkage of Census and Administrative Data to Quality Assure the 2011 Census for England and Wales . . . . 453 – 473
Blake, A. See Raymer, J.
Bogaerts, S., See Spreen, M.
Brancato, G. See Signore, M.
Brion, P., Gros, E. Statistical Estimators Using Jointly Administrative and Survey Data to Produce French Structural Business Statistics . . . . 589 – 609
Broome, J. First Impressions of Telephone Survey Interviewers . . . . 611 – 625
Bryant, J., See Alberts, I.
Bryant, J.R. and Graham, P. A Bayesian Approach to Population Estimation with Administrative Data . . . . 475 – 487
Buelens, B., See Daas, P.J.H.
Buettner, T., See Alberts, I.
Burger, J., van Delden, A., and Scholtus, S. Sensitivity of Mixed-Source Statistics to Classification Errors . . . . 489 – 506
Cetkovic, P., See Schnetzer, M.
Chambers, R.L., Discussion . . . . 507 – 513
Chambers, R.L., See Chipperfield, J.O.

Charlesworth, A., See Blackwell, L.
Chipperfield, J.O. and Chambers, R.L. Using the Bootstrap to Account for Linkage Errors when
Analysing Probabilistically Linked Categorical Data . . . . 397 – 414
Cuevas, Á., Quilis, E. M., and Espasa, A. Quarterly Regional GDP Flash Estimates by Means of
Benchmarking and Chain Linking . . . . 627 – 647
Daas, P.J.H., see Karlberg, M.
Daas P.J.H., Puts, M.J., Buelens, B., and van den Hurk, P.A.M. Big Data as a Source for Official
Statistics . . . . 249 – 262
Davern, M., See Barron, M.
Dever, J.A., See Valliant, R.
Di Consiglio, L. and Tuoto, T. Coverage Evaluation on Probabilistically Linked Data . . . . 415 – 429
Dorell, C., See Barron, M.
Elliot, M., See Shlomo, N.
Elliott, M.R., See Mneimneh, Z.M.
Espasa, A., See Cuevas, Á.
Falkingham, J., See Alberts, I.
Fienberg, S.E. Discussion . . . . 527 – 535
Fleishman, L., Gubman, Y., and Tur-Sinai, A. Dwelling Price Ranking versus Socioeconomic
Clustering: Possibility of Imputation . . . . 205 – 229
Forster, J. J., See Alberts, I.
Gabrielli, L., See Marchetti, S.
Gerland, P., See Alberts, I.
Gerritse, S.C., van der Heijden, P.G.M., and Bakker, B.F.M. Sensitivity of Population Size Estimation
for Violating Parametric Assumptions in Log-linear Models . . . . 357 – 379
Giannotti, F., See Marchetti, S.
Giesen, D., See Bavdaž, M.
Giusti, C., See Marchetti, S.
Grafström, A., Matei, A. Coordination of Conditional Poisson Samples . . . . 649 – 672
Graham, P., See Bryant, J.R.
Gros, E., See Brion, P.
Gubman, Y., See Fleishman, L.
Heeringa, S.G., See Mneimneh, Z.M.
Holmberg, A. Discussion . . . . 515 – 525
Holmberg, A., See Karlberg, M.
Hulliger, B., See Karlberg, M.
Humer, S., See Schnetzer, M.
Jacques, P., See Karlberg, M.
Kappelhof, J.W.S. Face-to-Face or Sequential Mixed Mode Surveys Among Non-Western
Miniorities in the Netherlands: The Effect of Different Survey Designs on the Possibility
of Nonresponse Bias . . . . 1 – 30
Karlberg, M., Biffignandi, S., Daas, P.J.H., Holmberg, A., Hulliger, B., Jacques, P., Lehtonen, R.,
Münnich, R.T, Shlomo, N., Silberman, R., and Stoop, I. Preface . . . . 149 – 153
Karr, F.A., See Kim, J.H.
Keilman, N., See Alberts, I.
Kim, J.H, Karr, F.A, and Reiter, J.P. Statistical Disclosure Limitation in the
Presence of Edit Rules . . . . 121 – 138
King, T., See Alberts, I.
Kirchner, A. Validating Sensitive Questions: A Comparison of Survey and Register Data . . . . 31 – 59
Korenjak Černe, S., See See Bavdaž, M.
Kreuter, F., See Valliant, R.
Lehtonen, R., See Karlberg, M.
Lenk, M., See Schnetzer, M.
Li, J. and Valliant, R. Linear Regression Diagnostics in Cluster Samples . . . . 61 – 75
Mathew, T., see Young, D.S.
Löfgren, T., See Bavdaž, M.,
Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S.,
Pappalardo, L., and Gabrielli, L. Small Area Model-Based Estimators Using Big
Data Sources . . . . 263 – 281

Matei, A. See Grafström, A.
Mneimneh, Z.M., Tourangeau, R., Pennell, B.-E., Heeringa S.G. and Elliott. M.R. Cultural Variations in the Effect of Interview Privacy and the Need for Social Conformity on Reporting Sensitive Information . . . . 673 – 697
Montgomery, R., See Barron, M.
Moser, M., See Schnetzer, M.
Muñoz, J.F, see Berger, Y.G.
Münnich, R.T., See Karlberg, M.
Oguz Alper, M. and Berger, Y.G. Variance Estimation of Change in Poverty Rates: an Application to the Turkish EU-SILC Survey . . . . 155 – 175
O'Hagan, A., See Alberts, I.
Onorante, L., See Alberts, I.
Owens, D., See Alberts, I.
Pappalardo, L., See Marchetti, S.
Pedreschi, D., See Marchetti, S.
Pennell, B.-E., See Mneimneh, Z.M.
Pratesi, M., See Marchetti, S.
Puts, M.J., See Daas, P.J.H.
Quilis, E. M., See Cuevas, Á.
Raftery, A., See Alberts, I.
Raymer, J., Rees, P., and Blake, A. Raymond-Blaess, V., See Bavdaž, M., Frameworks for Guiding the Development and Improvement of Population Statistics in the United Kingdom . . . . 699 – 722
Rees, P., See Raymer, A.
Reiter, J.P., See Kim, J.H.
Rinzivillo, S., See Marchetti, S.
Rogers, N.J., See Blackwell, L.
Salvati, N., See Marchetti, S.
Scanu, M., See Signore, M.
Schnetzer, M., Astleithner, F., Cetkovic, P., Humer, S., Lenk, M., and Moser, M. Quality Assessment of Imputations in Administrative Data . . . . 231 – 247
Scholtus, S., See Bakker, B.F.M.
Scholtus, S., See Burger, J.
Ševčíková, H., See Alberts, I.
Shlomo, N., Antal, L., and Elliot, M. Measuring Disclosure Risk and Data Utility for Flexible Table Generators . . . . 305 – 324
Shlomo, N., See Karlberg, M.
Signore, M., Scanu, M., and Brancato, G. Statistical Metadata: a Unified Approach to Management and Dissemination . . . . 325 – 347
Silberman, R., See Karlberg, M.
Smith, P.W.F, See Alberts, I.
Smith, P.W.F., See Yildiz, D.
Snijkers, G., See Torres van Grinsven, V.
Spreen, M., and Bogaerts, S. B-Graph Sampling to Estimate the Size of a Hidden Population . . . . . 723 – 736
Stoop, I., See Karlberg, M.
Tao, X., See Barron, M.
Templ, M. Quality Indicators for Statistical Disclosure Methods: A Case Study on the Structure of Earnings Survey . . . . 737 – 761
Torres van Grinsven, V. and Snijkers, G. Sentiments and Perceptions of Business Respondents on Social Media: an Exploratory Analysis . . . . 283 – 304
Tourangeau, R., See Mneimneh, Z.M.
Tuoto, T., See Di Consiglio, L.
Tur-Sinai, A., See Fleishman, L.
Valliant, R., see Li, J.
Valliant, R., Dever, J.A., and Kreuter, F. Effects of Cluster Sizes on Variance Components in Two-Stage Sampling . . . . 763 – 782
Van Delden, A., See Burger, J.
Van der Heijden, P.G.M., See Bakker, B.F.M.
Van der Heijden, P.G.M., See Gerritse, S.C.

Van den Hurk, P.A.M., See Daas P.J.H.
Wolter, K.M., See Barron, M.
Yildiz, D. and Smith, P.W.F. Models for Combining Aggregate-Level Administrative Data in the Absence of a Traditional Census . . . 431 – 451
Young, D.S. and Mathew, T. Ratio Edits Based on Statistical Tolerance Intervals . . . 77 – 100
Zardetto, D. ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys . . . 177 – 203
Zeng, W., See Barron, M.
Zhang, L.-C. On Modelling Register Coverage Errors . . . 381 – 396
Zhang, L.-C. On Proxy Variables and Categorical Data Fusion . . . 783 – 807

Book Reviews

Online Panel Research: A Data Quality Perspective
Cornesse, C., and Blom, A. G. . . . 809 – 812
Survey Data Collection and Integration
Earp, M.S. . . . 139 – 140
Practical Tools for Designing and Weighting Survey Samples
Ruiz Espejo, M. . . . 813 – 815
Managing and Sharing Research Data: A Guide to Good Practice
Mulcahy, T.M. . . . 817 – 819
Register-based Statistics
Resnick, D.M. . . . 141 – 142
Improving Surveys with Paradata: Analytic Uses of Process Information
Walejko, G.K. . . . 143 – 145
Sampling and Choosing Cases in Qualitative Research
Willis, G. . . . 147 – 148

Printed in December 2015