
# Journal of Official Statistics
## vol. 32, i. 4 (2016)

**ARTICLES**

**BOOK REVIEW**



***EDITORIAL***

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 787–788, http://dx.doi.org/10.1515/JOS-2016-0040

# JOS Special Section on The Role of Official Statistics in Statistical Capacity Building – Editorial

*Irena Ograjenšek*[1]

The idea of developing a Special Section in the Journal of Official Statistics that would focus on the role of official statistics in statistical capacity building was formed at the 2013 World Statistics Congress in Hong Kong. It emerged within the framework of a Special Topic Session entitled *Official Statistics in the Service of Business and Industrial Statisticians* that I organized and chaired.

The papers presented in that session focused entirely on the relationship between official statistics on one hand, and business and industrial statistics on the other hand. However, the subsequent lively discussion (thanks to Special Topic Session discussant Carol Joyce Blumberg) went beyond the service relationship described in the session title. The issue of statistical capacity building – for many decades one of the primary goals of the International Statistical Institute and affiliated associations – and the role of official statistics in it came to the fore along with many other important related topics. Now, after the World Statistics Congress, to my great satisfaction, some of these topics are systematically addressed in the papers included in this Special Section.

The section is divided by theme into four related parts.

The first part deals with the role of official statistics in the 21st century. In this framework, Steve MacFeely addresses the need for continuous evolution of official statistics by emphasizing the fact that change always simultaneously presents threats and offers opportunities.

The second part is dedicated to official statistics and statistical capacity building in practice. Here, Sharleen Forbes and Alan Keegan present experiences from New Zealand with targeted efforts to raise the official statistics capability among government employees, while Tomi Deutsch uses the case of the Consumer Price Index to discuss statistical capacity building among official statisticians.

The third part contains what I like to refer to as the future outlook. In this framework, Shirley Coleman presents data mining opportunities in official statistics that small and medium-sized enterprises can use to enhance their business opportunities. Her paper is followed by Ron Kenett's and Galit Shmueli's description of a journey from 'generic' data quality to what they refer to as Information Quality (InfoQ) in official statistics. Last, but not least, Luciana Dalla Valle presents a flexible methodology which uses official statistics to deal with self-selection bias in survey data collection processes.

[1] University of Ljubljana, Faculty of Economics, Kardeljeva pl. 17, 1000 Ljubljana, Slovenia. Email: irena.ograjensek@ef.uni-lj.si

The fourth and final part contains two invited commentaries. In the first one, John Pullinger, the United Kingdom's National Statistician, presents his take on the role of official statistics in statistical capacity building. He systematically discusses all relevant aspects, starting with the historical perspective, continuing with the interplay of national and global capacity building efforts, and concluding with the data revolution challenge. In the second commentary, Marleen De Smedt, Advisor to the Director-General of Eurostat, explains how Eurostat currently addresses the needs of official statistics users.

This Special Section could not have been prepared without the support of very significant people. I am much indebted to the colleagues who accepted my invitation to review submissions and profoundly thank them for their careful work and insightful contributions. Special thanks go to JOS Editors-in-Chief Ingegerd Jansson and Annica De Groote, as well as Susanna Emanuelsson from the JOS Editorial Office, for their continuous support during the entire editorial process.

I hope you will find the Special Section informative and inspirational.

Irena Ograjenšek
Guest Editor

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 789–810, http://dx.doi.org/10.1515/JOS-2016-0041

# The Continuing Evolution of Official Statistics: Some Challenges and Opportunities

*Steve MacFeely*[1]

As economies, societies, and environments change, official statistics evolve and develop to reflect those changes. In reaction to disruptive innovations arising from globalisation, technological advances, and cultural changes, the pace of change of official statistics will accelerate in the future. The motivation for change may also be more existential than that of the past as official statisticians consider the survival of their discipline. This article examines some of the emerging developments and questions whether they present threats or offer opportunities.

*Key words:* Big data; globalisation; measuring progress; statistical literacy; interoperability.

## 1. Introduction

'The only thing that is constant is change' – Heraclitus

Ever since official statistics have been compiled, they have evolved to reflect changes in economy and society. Only a few years ago, reducing respondent burden was one of the dominant issues for official statistics in Europe and was the focus of many international statistical conferences. Yet today, following the financial crash of 2008 and the subsequent economic crisis, burden is no longer the dominant issue. Now a new range of more immediate pressures are confronting official statistics. Disruptive innovations such as social media and big data, measurement challenges arising from globalisation, and the need to understand social, economic, and environmental progress, are forcing official statistics to once again adapt and evolve.

Over time official statistics have become recognised as a cornerstone of democracy, underpinning transparency and accountability. This recognition is important, as the collection and compilation of statistics are costly both to the public, who pay for them through their taxes, and to businesses because of the administrative costs of supplying data. As recompense, most nations view official statistics as a public good and disseminate them free of charge to encourage the public and businesses to use those data to support their decision making. This is sensible and in line with the United Nations Fundamental Principals of Statistics (United Nations 2014). But many users, including governments, appear to misunderstand the real cost of 'public goods', leading to the paradox that exists today where there is an expectation of higher standards of quality and precision yet an

[1] United Nations Conference on Trade and Development, Geneva, Switzerland and University College Cork, College Road, Cork T12 YN60, Ireland. Email: steve.macfeely@unctad.org

unwillingness to pay for that precision. If adequate funding is not secured, then National Statistical Institutes (NSIs) and other producers of official statistics may be forced to make compromises that will ultimately undermine the independence and quality of official statistics.

Official statistics faces many tests, and they cannot all be dealt within a single article. This article limits discussion to the challenges arising from globalisation, measuring progress, and big data. Why these and not others? The issues presented in this article have been selected owing to their importance, not only for NSIs but for governments, businesses, and citizens also: modern globalisation is complex and our understanding of it tentative at best, yet it has far-reaching implications – corporate taxation policy, trade policy including tariff and non tariff barriers, understanding international trade and value chains, offshoring and outsourcing – all with implications for World Trade Organization (WTO) negotiations, employment, and real incomes. The need for environmental sustainability and economic stability should be obvious to all. The policy implications arising from a wider concept of 'progress' have not always been as clear, but the post-2015 Sustainable Development Goals will have gone a long way to removing this ambiguity. The requirement for an integrated information system that straddles and balances the three pillars of economy, society, and environment has now been clearly articulated – this will be a huge challenge for national, regional, and international statistical organisations. The success or failure of that information system to hold governments, businesses, and peoples to account may be of profound existential importance for 'people, planet, and prosperity'. Finally, the use of big data raises weighty questions for governments, businesses, and citizens alike – not least regarding privacy, ethics, and trust. For NSIs, big data also present a dilemma as they are both a threat and an opportunity. How NSIs deal with this will be of great importance for everyone.

There are several reasons for this article: to provide a general context and illustrate or highlight some of the issues that NSIs are grappling with; to provide a brief explanation of why these issues are important, complex and not easily solved; perhaps also to provoke some debate and discussion; and finally to reassure readers that while much still remains to be done, considerable thought and work have already been dedicated to these issues. The article is presented in six sections: Sections 2 – 4 deal in turn with the changing landscape arising from the emergence of new data sources, the measurement challenges posed by globalisation and measuring progress; Section 5 discusses two cross-cutting issues that stem from the challenges outlined in the first three sections, specifically the need to take statistical literacy seriously and the difficulties of moving from legacy production systems to more modern interoperable systems. The final section presents a short conclusion.

## 2. A Changing Landscape

According to Ball (2004), the etymology of the term *statistics* comes from the Latin term *statisticum collegium*, meaning 'council of state', and refers to *science of the state.* However, there is no shared view or agreement on when official statistics first emerged, largely reflecting differences in how statistics and official statistics are defined. What is clear is that the first set of national accounts was produced by William Petty in Ireland in 1652 during the Cromwellian occupation (Canny 1987). Ever since, statistics have served national administrations, whether free market or centrally planned, and have continually evolved to remain relevant. Hence Quetelet's description of statistics as "that particularly governmental science" (Mazower 2012, 100).

Today, the landscape in which official statistics operate is changing, and changing fast. In a world where our increasing day-to-day dependence on technology, social media, and electronic transactions are leaving significant 'digital footprints' in their wake, and where business models for many electronic service providers generate revenue from advertising rather than the provision of a core service, the monopoly held by official statistics for so long, namely to provide timely and high-quality statistics for free, is being challenged. The power of these new technologies should not be underestimated. Anyone doubting their reach should consider the following: there are an average 500 million tweets per day (Krikorian 2013), while Facebook has more than one billion mobile users every month (Popper 2014) the first social network to surpass one billion registered accounts (Statistica 2015). To put this in perspective, if Facebook was a country it would have the third largest population in the world after China and India. There are over eight billion snapchat views sent every day (Aslam 2015). Anyone unconvinced of the influence of social media need only look at the part it played in the 2009 UK music chart's Christmas number one, Facebook's 'Rage Against the Machine' coup (Paterson 2009), the downfall of the Philippine President Joseph Estrada (Shirky 2011), or more recently in the Arab Spring (Howard et al. 2011; Dewey et al. 2012). Of course the global influence of technology should not be overestimated either; for some technologies a sizeable 'digital divide' persists. For example, the International Telecommunication Union (2015) estimates that global Internet penetration is only 43% and global mobile broadband subscription 46%, although they are as high as 78% in the developed world. Even within countries, access barriers across various social, geographic, or economic strata may lead to important cohorts being excluded, with obvious bias implications for statistics (see Struijs and Daas 2014).

The torrent of so-called 'big data' now generated as a by-product of these new digital services has been described as a "Data Deluge" (Vale 2012). But large datasets are nothing new, so what does big data mean? This is one of the challenges when discussing big data: "There is no rigorous definition of 'big data'" (Mayer-Schonberger and Cukier 2013, 6). The Wikipedia definition of big data is "a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications." Gartner, in what has become known as the three Vs, defines big data as high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. The Task Team of the United Nations Economic Commission for Europe (UNECE) High-Level Group for the Modernisation of Statistical Production and Services similarly defined big data as "data sources that can be, generally, described as high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making" (UNECE 2013). In other words, big data should be huge in terms of volume (i.e., at least terabytes), have high velocity (i.e., be created in or near real-time), and be varied in type (i.e., contain structured and unstructured data and span temporal and geographic planes).

Kitchin (2014) also argues that to be considered 'big' a dataset should be exhaustive in scope, fine-grained in resolution, relational in nature, flexible, and scalable.

The quantity of data now in existence is unknown. Here again, definitional differences complicate matters; the word 'data' means different things to different people. Consequently there are various estimates. Hilbert and Lopez (2012) estimate that 300 exabytes of data were stored in 2007 and over 90% of these were in digital format compared to only 25% in 2000. In fact the volumes of data now being generated are so great, increasing in volume at an exponential rate, that according to Lynch (2012) "90% of all information ever created was produced in the last two years alone"; Science Daily (2013) quotes a similar growth rate. According to Lillington 2013, "Stored digital content is doubling every two years, reaching one zettabyte [by 2012]". Kitchin (ibid.) states we are now in the 'Zettabyte Age' (an incomprehensibly large $2^{70}$ bytes). So while estimates and definitions differ, there is general consensus that the growth in data volume has been breathtaking and is set to continue. How much of these data are usable from a statistical perspective is not immediately clear, but there is sufficient evidence to suggest that adopting a closed mind would be unwise. Of growing interest is the blend of hard and soft, quantitative and qualitative data which may facilitate new and innovative types of analyses. Recent examples where usable statistics have been derived illustrate the point: The Billion Prices Project at the Sloan School at Massachusetts Institute of Technology (MIT) compiles price indices by web scraping (Cavallo 2012), Eggers (2011) developed an experimental sentiment index using Twitter, 'The Vibes of Ireland', pioneering work in Estonia generated transport and tourism statistics from mobile-phone data (Ahas and Tiru 2012), and the studies conducted by the SENSEable City Lab at MIT used Flickr to compile tourism statistics (Girardin at al. 2008). Furthermore, as Radermacher (quoted in Penneck 2014) has noted quite correctly, big data and social media do not respect national boundaries and thus offer the potential to generate new, internationally comparable statistics.

From the perspective of official statistics, the emergence of social media is a disruptive innovation. Many modern IT service providers are now disseminating statistics that to the untrained or undiscerning eye resemble official statistics, in order to attract media attention or attract browsers to their sites. Increasingly, official statistics are being forced to differentiate themselves from these new 'by-product' statistics by highlighting their relative quality, which comes from conceptual rigour, methodological quality, and transparency of metadata. As Borgman notes (2015), trusting other's data is crucial to reputation. Thus official statistics must ensure the quality of the data being produced is robust and fit for purpose. Furthermore, NSIs should be able to quantify the relative precision of their data. It will not be enough to say that the quality of data is fit for purpose; quality standards must be defined and supporting evidence must be provided to demonstrate proper standards have been achieved. If official statistics are to stand over bold comparisons, such as 'Michelin Star v Fast Food' quality (Radermacher 2013), when comparing official and unofficial statistics, then official statisticians must place greater emphasis on disseminating supporting evidence.

The emergence of new technologies and 'big data' are forcing many challenging questions to be asked. Mark Zuckerberg, the founder of Facebook, famously claimed that the age of privacy is over (Kirkpatrick 2010). Not everyone agrees. For example, one surmises the 39 million members of the 'Ashley Madison - Life is short, have an affair' website may be among those who disagreed when they learned the website had been hacked in 2015 with the hackers threatening to divulge their information (see Herrn 2015). There are many who have already voiced concerns over loss of privacy (see Pearson 2013; Payton and Claypoole 2015). Proposed changes to data-protection legislation in the EU to reinforce citizen's data-protection rights, including among other things the right 'to be forgotten', suggest privacy is still a real concern (European Parliament 2016). Others, such as Edward Snowden, have risked their freedom to highlight the threats to our privacy (Greenwald 2015). The increased volumes of big data being generated, and the potential to link and utilise those data, means that greater attention must be paid to data suppression techniques to ensure confidentiality can be safeguarded. This is a key tenet of the UN Fundamental Principles of Official Statistics (ibid.). But what if Zuckerberg is correct and future generations are less concerned about privacy; what are the implications for official statistics and anonymisation? If the privacy-benefit trade-off is insufficiently clear or immediate to the public, it begs the question of whether official statistics will be left in an anachronistic situation vis-à-vis the privacy of data providers.

The social media/big data space is a complex and rapidly developing area with many uncertainties, and thus an open but cautious approach is probably wise. The risk is, of course, that this is hype. Most likely it is. In 2014, Buytendijk argued that big data had passed the top of the 'Hype Cycle' and was moving towards the 'Trough of Disillusionment' and that now expectations regarding the use of big data would become more realistic. As Borgman (2015, 222) notes, big data are not always better data, stating: "Increasing the size of the haystack does not make the needle easier to find." Other practical questions centre on issues such as whether time series are still relevant as the pace of change increases (arguably this is already an issue for some areas experiencing rapid change, e.g., ICT statistics). Can big data be 'gamed' or manipulated by the public (remember the 2009 Facebook Christmas No.1 coup)? A growing proportion of Twitter accounts are actually 'social robots' used to influence public communication. "As few as 35 percent of Twitter followers may be real people, and as much as ten percent of activity in social networks may be generated by robotic accounts" (Borgman 2015, 131). Furthermore, Reich (2015) notes that in 2010, the top ten websites in the United States accounted for 75 percent of all page views. With such concentration comes power and the associated risks of abuse and manipulation. Will this inhibit the use of some big data as a source for compiling official statistics? Should new data streams generated from secondary data be made to conform to existing classifications (which are typically artefacts of compromise defined by the availability of data at a particular time), or will new contemporaneous ad-hoc classifications be generated? Strategic questions arise, too – for example, could/should official statisticians play a new role in the future, accrediting or certifying the quality of data derived from social media? Competition is naturally feared, but sometimes it results in a better product or service. But in a situation where official statistics must adhere to increasingly rigorous quality standards with all the inherent trade-offs between quality and timeliness, whereas 'marketing' statistics do not necessarily adhere to any known standards, it is not clear what the future holds for official statistics.

Official statistics must embrace and harness a range of technologies to survive in today's rapidly changing environment. These technologies and their owners have the potential to threaten the dominant role of official statistics or, more pessimistically, their existence. But if exploited, they may offer opportunities to compile or integrate and disseminate new information in new and exciting ways. In 2013, the leaders of the European Statistical System embraced the latter view in the Scheveningen Memorandum (European Statistical System Committee 2013), where they clearly saw potential for compiling official statistics from big data. For example, flow or dynamic statistics can be derived from administrative or continuous data (something that is very difficult and costly to achieve with sample-based statistics), offering the potential for more policy-relevant, outcome-based statistics. An example of where this might be very useful would be researching the real cost of innovation and testing Harford's assertion "Here's the thing about failure in innovation: it's a price worth paying" (2012, 103) by examining initial failure with later success by linking enterprise demography, Chief Executive Officer (CEO) attributes and histories with innovation survey microdata over time. New technologies and social media also offer an opportunity to move beyond traditional product innovation and to 'open innovate' or 'crowdsource', where the collective ingenuity of data consumers is tapped to become co-creators. Of course, the conundrum is how to leverage the wisdom of crowds in a productive and focused way when frequently they don't appear to know what they want (a dilemma that both Henry Ford and Steve Jobs noted. Henry Ford famously said 'If I had asked people what they wanted, they would have said faster horses' and Steve Jobs said 'People don't know what they want until you show it to them'). This is difficult to achieve, but it can be done as Lego's 'Designbyme' approach has demonstrated (Robertson and Breen 2013). But probably the single biggest challenge for official statisticians is managing expectations. No doubt in the years to come official statistics will experience a data revolution in which big data will play an increasingly important role. But those changes may not happen as quickly as some expect. Furthermore, it is highly unlikely that the costs of compilation will fall dramatically as a result.

## 3. Globalisation

Depending on the lens through which history is viewed (political, social, colonial, economic, biological) the beginnings of globalisation differ – see O'Rourke and Williamson (2002), Menzies (2003), Ferguson (2004), Osterhammel and Petersson (2005), and Mann (2011) for some varied perspectives. But all agree on its importance, with President Clinton describing it as the "central reality of our time" (Lewis, 2000). Taleb (2007) agrees, but warns against 'naive globalisation' which gives the appearance of stability but in fact creates interlocking fragility. In any event, international trade has existed in one form or another for centuries. Now Sturgeon (2013) argues that we need to distinguish between what he calls internationalisation and today's more complex and integrated globalisation. Sachs (2012) and Yergin (2012) draw similar distinctions in what they respectively term "new globalisation" or "the new age of globalisation".

Both Sachs and Friedman (2006) argue that changes in the geopolitical landscape, such as the collapse of the Berlin Wall in 1989, China joining the WTO and the emergence of the BRIC countries (Brazil, Russia, India, and China) in 2001 have effectively expanded the commercial or economic world. Yergin makes a similar case but also stresses the contribution of privatisation and deregulation in the 1980s, political reform in India and the integration of Europe. All agree that technological revolution, such as the birth of the World Wide Web in 1991 and the emergence of the digital electronic age has utterly changed the way businesses transact. Baldwin refers to this as the "second unbundling of globalisation" (2011, 4), where the Internet has dramatically reduced international communications and transactions costs, facilitating the financialisation of commodities and a blurring of the distinction of tradable and non tradable services. Financial integration depended on informational integration.

The reach of economic globalisation has had impacts on employment growth and loss, crisis contagion, trade policy, and protectionism. So much so that Mazower (2012) argues that globalisation has in fact become an ideology in and of itself. Stiglitz (2002, 2010) has highlighted the role of globalisation in rising income inequality and decline in real incomes. The preponderance of cheap goods from China and India has helped to create an artificial inflation lid across much of the western world. But for developing countries too, the benefits of economic globalisation can be mixed. Traditional measurements of the macro economy and international trade have not always made it easy to understand how economies interact or participate in global value chains. Piketty (2014) has highlighted the consequent challenges for developing countries of ensuring that their markets are not exploited or prematurely forced to liberalise. Policy makers must grapple with such issues. Official statistics must contribute to these debates and discussions in a positive rather than a limiting way. To steal a slogan commonly used about IT, statistics should be a climbing frame, not a cage.

By ignoring the impact of globalisation, official business and macroeconomic statistics may be measuring price, production, and Gross Domestic Product (GDP) incorrectly (Sturgeon 2013; Houseman et al. 2014). An analysis of Bureau of Labour Statistics price data in the US led Houseman et al. (2010, 2) to conclude that indices were biased as 'Price declines associated with the shift to low cost foreign suppliers generally are not captured'. This in turn may result in a miscalculation of Gross Value Added (GVA) or at least its attribution to the wrong country or region with implications for macroeconomic statistics and, in a European context, the distribution of structural and cohesion funds. Our understanding of how and where labour inputs interact within the production function may no longer be realistic or accurate. Consequently, the standard approach to calculating labour and multifactor productivity may be naïve, or worse – misleading.

Eurostat, in cooperation with the WTO, the Organisation for Economic Cooperation and Development (OECD), the United Nations Economic Commission for Europe (UNECE), the United Nations on Trade and Development (UNCTAD) and the United Nations Statistics Division (UNSD) have begun to address these issues. UNCTAD has been critically highlighting the growing importance of globalisation for many years (2005, 2008, 2012), and some of their more recent work (2015) raises questions regarding the measurement and our understanding of Foreign Direct Investment (FDI). In 2010, the OECD published their handbook *Measuring Globalisation* with recommendations for new indicators. The Eurostat FRIBS or Framework Regulation for Integrated Business Statistics (European Commission 2011 and 2013) and Simplification of Trade Statistics or SIMSTAT (European Commission 2012) projects, although proving contentious, are at least challenging official business and macroeconomic statisticians to question whether the existing approach is fit for purpose. UNECE established a Group of Experts to examine the impact of globalisation on national accounts. In 2012 they published their first guide on the impacts of this phenomenon. But many thorny issues remain to be unravelled. For example, MacFeely (2012) has queried whether 'nationality' is still a meaningful concept with regard to enterprise ownership as many large multinational enterprises (MNEs) are effectively super-national, defying geo-spatial classification. Alajääskö et al. (2011) and Sturgeon (2013) have illustrated the importance of global value chains and international sourcing, with implications for producer prices and classifications. For example, there is no agreed approach for the calculation and application of deflators. Other questions arise – should a 'Business Function' classification be introduced to business statistics? Should business functions replace product classifications in some instances – perhaps to codify business costs? The pros and cons must be weighed up carefully; business functions may be easier for respondents to understand, but it is not clear whether data classified to business function can provide sufficient detail to compile national accounts and supply and use tables. These questions cut right to the heart of alignment and integration. It goes without saying that properly measuring globalisation will require a coordinated global solution. For this reason, in the spring of 2015 the Statistical Commission of the UN agreed (see Decision 46/107) to establish two expert groups to look at these and other related issues, namely an Expert Group on the Handbook for a System of Extended International and Global Accounts (EG-SEIGA) and an Inter-Secretariat Working Group on International Trade and Economic Globalization Statistics (ISWG-ITEGS).

For small, open economies like Singapore or Ireland, GVA or GDP is not a very useful or informative measure of real output or standard of living and is of particularly limited use when attempting regional comparisons (MacFeely et al. 2011). Fitzgerald (2014, 34), discussing the Irish economy, is worth quoting on this point: "problems with the standard national accounts presentation arise from the exceptional openness of the Irish economy and the related globalisation of the tradable sector of the EU economy. To better understand what is happening in Ireland today significant additional data are needed, supplementary to the national accounts, which would show the contribution of each sector of the economy to GNP." These issues are not unique to Ireland and cannot be ignored if modern business and economic statistics are to be useful to policy makers grappling with the complex realities of economic globalisation.

The implications for official statistics are clear and nontrivial. This new form of globalisation is more complex and more difficult to measure than its predecessors. There is some urgency in dealing with this issue, as the research outlined above suggests that current business statistics methodology may not properly capture the complexity and impact of globalisation. This in turn has implications for core macroeconomic indicators and the policies they inform. Apart from the challenges of measuring GVA and GDP correctly, many are questioning whether GDP is an appropriate or useful measure for the 21st century (Coyle 2014).

## 4. Measuring Progress

Following World War I, the Great Depression of the 1930s and the onset of World War II, GDP emerged from these crises and the 1944 Bretton Woods conference as the pre-eminent economic indicator (Dickinson 2011; Fioramonti 2013) and the ultimate measure of a country's overall welfare. Although a purely economic measure, GDP has frequently been used as a proxy measure for welfare. Palmer in 1966 described GDP as the "chief criterion for national welfare or progress." Steve Landefeld, Director of the United States Bureau of Economic Analysis, in 2010 similarly noted the "singular focus on GDP alone as a measure of society's welfare" (Bureau of Economic Analysis, 2010). Described by Samuelson and Nordhaus (Landefeld 2000) as one of the greatest inventions of the 20th century, as Philipsen (2015, 237) notes "GDP is not just a measure of the economy. It defines the economy". But from the outset, Simon Kuznets, the economist most commonly associated with the creation of GDP, cautioned that GDP could unwittingly act as a 'statistical laundry' concealing inequality and would be an unreliable or inappropriate measure of well-being, noting "the welfare of a nation can scarcely be inferred from a measure of national income" (Kuznets 1962, 29). Stiglitz (2014) went further, saying that not only is GDP not a good measure of welfare, but "GDP is not a good measure of how well an economy is performing" and that "too much has already been sacrificed on the altar of GDP fetishism". Nevertheless, as Talberth et al. (2007, 1) note, "GDP maintains its prominent role as a catchall for our collective well being".

There have been many attempts since the 1970s to challenge the primacy of GDP as the definitive measure of progress, such as: the Measure of Economic Welfare (MEW); the Total Incomes System of Accounts (TISA) or the Index of Sustainable Economic Welfare (ISEW), which was later renamed the Genuine Progress Index (GPI), and the Human Development Index (HDI); and perhaps most famously, the unfortunately titled Gross National Happiness (GNH) proposed by the King of Bhutan. The essence or spirit of these alternatives was perhaps best encapsulated by Robert F. Kennedy's reference to GDP during a 1968 campaign speech in the University of Kansas: "it measures everything in short, except that which makes life worthwhile" (Kennedy 1968).

Considering the circumstances in which GDP achieved dominance, it is ironic that it was another economic crisis, namely the financial crash of 2008 and subsequent recession, that has arguably reinvigorated attempts to develop a more wide-ranging measure of progress. The Commission on the Measurement of Economic Performance and Social Progress (better known as the Stiglitz-Sen-Fitoussi Commission) was established by the then president of France, President Sarkozy, in 2008 to determine whether a better or more comprehensive measure of economic and social progress could be established. This commission reported in 2010 (see Stiglitz et al., 2010). In 2009, the European Commission published their roadmap "Beyond GDP", which is an amalgam of 'enlarged GDP', social and environmental indicators and other measures of well-being. The Obama administration formally established the Key National Indicators Commission in 2010 to develop a comprehensive indicator system (KNIS) for the United States, which comprises over 300 key and twelve composite indicators. The following year the OECD launched their "Better Life Index" (BLI) to try and address similar questions. UN Secretary-General Ban Ki-Moon, speaking at a High Level meeting on 'Happiness and Well-being: Defining a New Economic Paradigm' in 2012, noted the importance of establishing "a Sustainable Development Index, or a set of indicators to measure progress towards sustainable development" (United Nations 2012). Working separately, the United Nations University International Human Dimensions Programme on Global Environmental Change (UNU-IHDP) in collaboration with the United Nations Environment Programme

(UNEP) has also developed an Inclusive Wealth Index (IWI). All of these indicators – Beyond GDP, the KNIS, the BLI, the SDI, and the IWI – have adopted a dashboard approach rather than trying to develop a single aggregate index. This reflects the complexity of the modern world, and illustrates the communications and branding challenge ahead. It also reflects the wider scope of issues that now are included under the 'progress' umbrella: environmental sustainability; economic stability and sustainability; and social health, satisfaction, and general well-being.

From a cultural perspective, GDP, like $E = MC^2$, is a somewhat magical number with a celebrity status that transcends the number itself. At some superficial level these are two of the most recognised terms or concepts in the world, yet very few really understand what they mean. Chambers (1989, 128) discussing Einstein's General Theory of Relativity, neatly captured this irony: "the pathetic paradox that Einstein's discoveries, the greatest triumph of reasoning mind on record, are accepted by most people on faith". The same could arguably be said of Kuznets and GDP. The challenge for official statistics is how to supplement GDP, a number with considerable cultural authority, with another number or index that is complex enough to incorporate resource depletion, environmental degradation, well-being, social inequality, and economic performance, but is simple enough to be accepted as the new number. The public are receptive but confused; they intuitively understand they are caught in a 'Growth Trap' where progress is dependent on continual consumption and unsustainable replacement (Slade 2006). This situation prompted Richard Layard (2011) to quip "Anyone who believes in indefinite growth on a physically finite planet is either mad, or an economist". They also intuitively understand that while GDP is imperfect, it is at least, for the moment, the agreed barometer of economic progress. It has also been argued that GDP is an out of date concept, "a relic of a period dominated by manufacturing" (*The Economist* 2016, 22) struggling to capture the impact of myriad intangible innovations.

The abundance of rival indicators that have arisen in recent years to challenge the hegemony of GDP poses a problem for official statistics and potentially risks a loss of credibility. Ironically, the glut of alternatives developed to supplement GDP has only confused the public, cementing the dominant position enjoyed by GDP. So while many of these new indicators may in fact represent real technical progress, their sheer number can also be viewed as a metric of failure. There is an opportunity for official statistics to show leadership in this space and agree a definitive approach. This is of course easier said than done – many institutions and countries have invested time and energy in developing their own bespoke set. Any definitive dashboard must agree on how to address resource depletion, environmental degradation, climate change, social inequality, and sustainable economic progress. The United Nations Agenda 2030 explicitly calls for the development of "measurements of progress on sustainable development that complement GDP" (Goal 17.19), but it is noteworthy that the indicator framework agreed at the 47th United Nations Statistical Commission in March 2016 did not propose any indicators addressing this issue (United Nations Statistical Commission 2016) "Perhaps this is because there has been little consensus on a suitable replacement. Perhaps, more fundamentally, it is that there is even less consensus on how well being should really be measured and if quantitative measurements can be made at all" (Talberth et al. ibid). Thus finding an indicator (or a limited set of agreed indicators) that is not so simple that it ignores the negative externalities of production but not so complex that it is incomprehensible to users is a high-wire act. At first glance, this seems an impossible task, as the messages from dashboards are often hard to interpret, but the "Business Cycle Tracer" compiled by Statistics Netherlands provides an excellent example of how complexity can be presented in a digestible format by mapping the trajectory of fifteen indicators on to a $2 \times 2$ matrix (above/below trend and decreased/increased). There are arguments too for a single composite measure, despite the trade-offs and weighting challenges, as the media and public tend to like definitive answers (even if oversimplified). Whatever measure or measures are selected, they must be sufficiently authoritative and scientific to compete with GDP. This means moving away from terms like 'happiness' which do not translate directly or well from Eastern culture and lead many to misunderstand the concept being proposed.

## 5. Two Cross-Cutting Issues

The challenges outlined above can be broken down into many subchallenges. In this section, two particular cross-cutting issues are briefly discussed: statistical literacy and the modernisation of the compilation process. These issues have been selected as they are of relevance to everything discussed so far.

### *5.1. Statistical Literacy*

Statistics and general numeracy are increasingly becoming a necessary competency in modern life and the lingua franca of day-to-day transactions. Hence, improving statistical literacy is essential if future economies are to function efficiently and if citizens are to actively participate in and contribute to society. Producers of official statistics must take this matter seriously and consider their role in education more actively.

Arguably NSIs have an obligation to address this important challenge. If the public and the business community are insufficiently statistically literate to use statistics or distinguish between fact and comment, then everything NSIs have produced has been for nought. In a world where perhaps there are too many data (without question, too much comment is being portrayed as data or fact), users are increasingly confused and may select inappropriate data with which to inform themselves or upon which base their decisions. As societies and economies become more globalised and complex, official statistics, in order to properly measure social and economic transactions, are unavoidably becoming more complex in parallel.

To many, statistical results are not clear or intuitive. In fact, to many, results often appear improbable, impenetrable or unrealistic. In many cases this confusion arises because users do not read or understand the metadata, the background notes or the small print. Too often, users think they understand the data and their implications when clearly they do not. This situation may arise when a commonplace term means something quite different to a statistician than to the man on the street – for an example of this, consider how a typical householder and a national accountant might define or describe 'household savings'. But in other cases, NSIs must accept that the message is not always as clear as it could be. No doubt this contributes to the perception that statistics are less than useful and helps sustain the popularity of quips like Mark Twain's famous 'Lies, damned lies, and statistics'.

The day-to-day abuse and misuse of statistics may be inadvertent or deliberate, but either way, they pose a challenge for official statistics. Indiscriminate and often sensationalist media are the shop window through which official statistics are viewed. If these statistics are not presented properly, then the product and brand are tarnished. Users with little time, already confused and bewildered by the massive volume of statistics available, may find themselves drawn to statistics that are visually appealing or appear simple or straightforward. Users rarely see complex concepts or methodologies as attractive. The data deluge has created new or supplementary roles or responsibilities for national and international statistical institutes; we must repackage our products to make them more user friendly without compromising their integral quality and educate users so that they are better able to select and use appropriate data to suit their purposes. Thus brand building and education are intertwined and must become key objectives for NSIs in the future. In doing so, NSIs will provide a better service to their customers and protect their market position. On the statistical literacy side, there are already some good examples from which to learn: the New Zealand National Certificate in Official Statistics, the Irish Professional Diploma in Official Statistics for Policy Evaluation, the International Association of Statistical Education (IASE) International Statistical Literacy Project and the European Masters Programme in Official Statistics (EMOS). The UNECE (2012) has also published a guide for compilers on improving statistical literacy as part of their 'Making Data Meaningful' series. Equally, on the more promotional side, there are also some interesting examples from around the world of how this can be done: the Australian Bureau of Statistics has developed some highly creative and innovative interactive promotional tools, such as their 'Run that Town' app, as has the Central Bureau of Statistics in the Netherlands with their 'On a normal day' movie.

In tandem with promotion, NSIs must improve the accessibility of products; this means making it easier to find and use the data and as well as making it easier and more intuitive for non specialist users to understand those data. UNECE (2009a; 2009b; 2011) has published some useful guides on communicating and presenting statistics. There are also examples of how this might be done at country level: Statistics Finland and the Central Statistics Office in Ireland have used competitions to crowdsource ideas for statistical apps that encourage the use of statistics. Greater thought must also be given to ways to encourage businesses to respond and engage with statistics, perhaps through tailored or bespoke analyses to demonstrate the usefulness of the data to which they are contributing.

### *5.2. Building Interoperability*

As already noted, it would appear that official statistics can learn some valuable lessons from Lego, who deliberately set out to develop a fully integrated 'system of play' (Robertson and Breen 2013). Since 1958, over 400 billion Lego bricks have been produced, an estimated 62 pieces for every person on earth (Diaz 2008) – all of these bricks can be connected. At first glance, this seems the natural analogy for what official statistics as a discipline needs to achieve. After all, official statistics enjoys the advantage of working from more or less conceptually consistent frameworks and classification systems that have been carefully developed over time. A good example of this are the recently developed Tourism Satellite Accounts (UNSD 2010) that are conceptually consistent with the Balance of Payments and International Investment Position Manual, more commonly known as BPM6 (International Monetary Fund 2009) and UN System of National Accounts (UNSD 2009). The UN System of Environmental-Economic Accounting (United Nations 2014) is another example. This consistency across products is critical as it, like Lego, should offer the potential to develop an expandable, integrated data and analytical system. To some extent it already does, but several factors inhibit full scalability and interoperability; different IT and data infrastructures (MacFeely and Dunne 2014) and different administrative production systems. International statistical organisations promote standard approaches to production, such as the UNECE Generic Statistics Business Process Model and Generic Statistical Information Model. Considerable collaborative time and effort is also being devoted to examine how shared 'plug and play' systems can be developed (Museux 2012; Vale 2013).

From a modernisation perspective, the challenge for official statistics is how to move from existing legacy systems to a more streamlined vision of official statistics without wasting years of investment or compromising national mandates and structures. This challenge has been likened to rebuilding an airplane while in flight. Official statistics already employ a multitude of production systems around the world that unlike Lego cannot be easily joined up or connected, so it is not clear how to make the transition. Useful lessons can be taken from the evolution of freight containers. In the 1950s Malcolm McClean introduced the concept of the 'intermodal container' which eventually realised the international standardisation of shipping containers, allowing ship, road and rail transport to interconnect seamlessly (Levinson 2006). The transition from multiple freight and container systems to the ubiquitous 'twenty foot equivalent unit' container took years to implement and in the process there were winners (Singapore, New Jersey, and Felixstowe) and losers (Manhattan), but today the intermodal container has completely transformed the transport of merchandised goods and the face of international trade. Arguably the multimodal container is a better analogy for statistics than Lego – Lego was designed as an interoperable system from the very beginning, whereas as several models of official statistics have evolved independently and now require consolidation, like containers did. In many respects, current modernisation initiatives can be summarised as attempts to de-silo legacy production systems. However, in most cases, these attempts to de-silo are done within the constraints of national silos, that is, each country is attempting to de-silo independently. This should be no surprise, as this makes sense for a host of practical and strategic reasons. However, the globalised nature of many big data and the need to adopt a more integrated approach to measure globalisation suggests a more ambitious approach could be considered (for some fields at least). For some domains, the most logical and efficient approach may be to centralise statistical production in a single centre rather than replicating production many times over in individual countries. Obviously, this would not work for all domains, but where feasible, it would remove the need for 'plug and play' as the methodology would only be applied once with a better chance of international comparability.

One of great advantages enjoyed by official statistics is the high level of cooperation and, by and large, shared objectives between national and international statistical agencies around the world. Viewed as a whole entity, the capacity and global reach of the international statistical system is similar to that of a giant and powerful multinational enterprise. The challenge is to harness that power, so that the challenges of globalisation, measuring progress and levering big data are overcome without undermining valid national concerns. Moving to a continuous, expandable, fully integrated data and analytical 'system' rather than a disparate collection of statistical products or statistical compilers will take time and leadership, but if done sensibly will strengthen the official statistics brand.

## 6. Conclusion

Ever since official statistics have been compiled, they have adapted to economic, political, societal, and environmental events to remain relevant. This process of evolution continues today, albeit at an accelerating pace. This article has outlined three key issues for official statistics that will force further adaptation in the coming years. The issues of understanding and mapping globalisation, measuring progress, and harnessing big data and integrating them with more traditional sources are all complex in that they are conceptually knotty but also in that they all simultaneously present both threats and opportunities.

On the face of it, globalisation, progress or welfare and big data all seem like very disparate issues, and of course they are. Nevertheless there are some underlying themes or approaches that are common across all of them. Two cross-cutting themes were highlighted to illustrate this. For example, the importance of improved statistical literacy will be important for understanding all aspects of statistics, whether globalisation, welfare, or big data. The modernisation process too will have an impact on coordinating the improvement of quality of macroeconomic and business statistics, on environmental statistics, on well-being and on the integration of big data into day-to-day production.

The need for an improved understanding of globalisation has been recognised by the United Nations and NSIs; the UN Statistical Commission established two new working groups in 2015 to address these issues specifically. This is critical as global markets and financial integration facilitate global risks and contagion. Clear direction on how to reach a consensus approach for the measurement of progress is less evident and is now more urgent in the context of UN Agenda 2030. One of the complications around progress is that several disparate issues, including sustainable development and well-being, have been bundled under one broad umbrella. Furthermore each of these issues is individually complex and multidimensional. So as already noted, whatever the agreed indicators, they must be sophisticated enough to encompass issues of sustainability, inequality and welfare but straightforward enough to compete with the apparent simplicity of GDP.

Change always presents threats and offers opportunities. The challenge is to identify and mitigate the threats while seizing the opportunities. To do this, NSIs and producers of official statistics need to identify where they enjoy a comparative advantage and play to those strengths. Arguably, transparency, impartiality and political independence are among these strengths. These attributes are central to democracy and public accountability and their importance should not be underestimated but rather highlighted at every opportunity. Despite the abundance of competing information available today, the justification or arguments in support of official statistics have never been stronger. While the need for impartial evidence-informed policy formulation and evaluation is implicitly understood, as is the importance of statistics, compiled independently of ideology or political interference, to democratic accountability, this can easily be forgotten by the public and politicians when faced with budgetary pressures. Official statistics, as a public good, must be the bulwark against the 'asymmetries of information' that Stiglitz (2002) has railed against or the uninformed debates that Piketty (2014, p. 2–3) has described as dialogues of the deaf which are 'based on an abundance of prejudice and a paucity of fact'. The Data Revolution report "A World that Counts" has also highlighted the importance of access to data, even suggesting that uneven access to data might be considered a new frontier of inequality (Independent Expert Advisory Group on a Data Revolution for Sustainable Development 2014). Of course, the real high-wire act will be to explain our increasingly complex world in ever simpler terms.

## 7. References

Ahas, R. and M. Tiru. 2012. "Passive Anonymous Mobile Positioning Data for Tourism Statistics." Paper presented at the 11th Global Forum on Tourism Statistics, Reykjavík, 14–16 November 2012.

Alajääskö, P., P. Nielsen, S. Rikama, and J. Sisto. 2011. "Global Value Chains - International Sourcing to China and India." Available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Archive:Global_value_chains_-_international_sourcing_to_China_and_India (accessed 26 September, 2016).

Aslam, S. 2015. "Snapchat by the Numbers: Stats, Demographics and Fun Facts", Omnicore, October 7, 2015. Available at: http://www.omnicoreagency.com/snapchat-statistics/ (accessed 26 September, 2016).

Baldwin, R. 2011. "21st Century Regionalism: Filling the Gap Between 21st Century Trade and 20th Century Trade Rules." World Trade Organisation, Staff Working Paper ERSD-2011-08. Available at: https://www.wto.org/english/res_e/reser_e/ersd201108_e.pdf (accessed 25 April, 2015).

Ball, P. 2004. *Critical Mass: How One Thing Leads to Another*. London: Arrow Books.

Borgman, C.L. 2015. *Big Data, Little Data, No Data - Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Bureau of Economic Analysis. 2010. *The Economist Debates*: GDP, April 20, 2010, Available at: https://www.bea.gov/about/pdf/Newsclips_GDP_group.pdf (accessed 26 September, 2016).

Buytendijk, F. 2014. "Hype Cycle for Big Data, 2014." Gartner. Available at: https://www.gartner.com/doc/2814517/hype-cycle-big-data- (accessed 11 August, 2015).

Canny, N. 1987. *From Reformation to Resistance: Ireland, 1534 – 1660*. Dublin: Helicon Press.

Cavallo, A. 2012. "Online and Official Price Indexes: Measuring Argentina's Inflation." *Journal of Monetary Economics* 60: 152–165. Doi: http://dx.doi.org/10.1016/j.jmoneco.2012.10.002.

Chambers, W. 1989. *Ghosts on the Roof – Selected Journalism of Whitaker Chambers 1931 – 1959*. Washington, D.C: Regnery Publishing.

Coyle, D. 2014. *GDP: A Brief but Affectionate History*. Princeton: Princeton University Press.

Dewey, T., J. Kaden, M. Marks, S. Matsushima, and B. Zhu. 2012. "The Impact of Social Media on Social Unrest in the Arab Spring." Final Report prepared for the Defense Intelligence Agency, Stanford University. Available at: http://stage-ips.stanford.edu/sites/default/files/shared/2012%20Practicum%20Policy%20Brief%20SocialMedia.pdf (accessed 28 March, 2014).

Diaz, J. 2008. "Everything You Always Wanted to Know About Lego." Available at: http://lego.gizmodo.com/5019797/everything-you-always-wanted-to-know-about-lego (accessed 26 September, 2016).

Dickinson, E. 2011. "GDP: A Brief History – One Stat to Rule Them All." *Foreign Policy*, January–February 2011. Available at: http://foreignpolicy.com/2011/01/03/gdp-a-brief-history/ (accessed 26 September, 2016).

Dunning, J.H. 2005. "More – Yet More – On Globalization." *Transnational Corporations* Vol. 14, No. 2, August 2005 (UNCTAD/ITE/IIT/2005/5). Available at: http://unctad.org/en/Docs/iteiit20055a5_en.pdf (accessed 26 September, 2016).

Eggers, J. 2011. "Social Media Like Twitter and Facebook Can Influence Statistics." Paper presented at the 2nd CSO Administrative Data Seminar, Dublin Castle, 29 September 2011. Available at: http://www.cso.ie/en/newsandevents/eventsconferences-seminars/administrativedataseminars/2ndadministrativedataseminar/ (accessed 21 February, 2014).

European Commission. 2011. "Business & Price Statistics – A Strategic Approach for the Coming Years." Paper presented by Eurostat to the Business Statistics Directors Group, Doc. BSDG June 2011_4_1.

European Commission. 2012. "SIMSTAT Project for Exchange of Micro-Data on EU-intra Trade." Paper presented by Eurostat to the Business Statistics Directors Group, Doc. BSDG June 2012_16.

European Commission. 2013. "Framework Regulation for Integrating Business Statistics." Available at: http://ec.europa.eu/smart-regulation/impact/planned_ia/docs/2012_estat_011_business_statistics_en.pdf (accessed 26 September, 2016).

European Parliament. 2015. "European Parliament News - Q&A: new EU rules on data protection put the citizen back in the driving seat" Available at: http://www.europarl.europa.eu/news/en/news-room/20160413BKG22980/qa-new-eu-rules-on-data-protection-put-the-citizen-back-in-the-driving-seat (accessed 26 September, 2016).

European Statistical System Committee. 2013. "Scheveningen Memorandum - Big Data and Official Statistics", adopted 27 September 2013. Luxembourg. Available at: http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13 (accessed 31 March, 2016).

Ferguson, N. 2004. *Empire: How Britain Made the Modern World*. London: Penguin Books.

Fioramonti, L. 2013. *Gross Domestic Problem: The Politics Behind the World's Most Powerful Number*. London: Zed Books.

FitzGerald, J. 2014. "Additional Data Needed to Understand Developments in the Economy." *Quarterly Economic Commentary* 2014: 34–25.

Friedman, T. 2006. *The World is Flat: The Globalized World in the Twenty-First Century*. London: Penguin Books.

Girardin, F., F. Dal Fiore, C. Ratti, and J. Blat. 2008. "Leveraging Explicitly Disclosed Location Information to Understand Tourism Dynamics: A Case Study." *Journal of Location Based Services* 2: 41–56.

Greenwald, G. 2015. *No Place to Hide: Edward Snowden, the NSA and the Surveillance State*. UK: Penguin Books.

Harford, D. 2012. *Adapt: Why Success Always Starts with Failure*. London: Abacus.

Hern, A. 2015. "Infidelity Site Ashley Madison Hacked as Attackers Demand Total Shutdown." *The Guardian*, 20 July 2015. Available at: http://www.theguardian.com/technology/2015/jul/20/ashley-madison-hacked-cheating-site-total-shutdown (accessed 20 July, 2015).

Hilbert, M. and P. Lopez. 2012. "How to Measure the World's Technological Capacity to Store, Communicate and Compute Information." *International Journal of Communication* 6: 956–979.

Houseman, S.N., C.J. Kurz, P. Lengermann, and B.R. Mandel. 2010. "Offshoring Bias in U.S. Manufacturing: Implications for Productivity and Value Added." Board of Governors of the Federal Reserve System, International Finance Discussion Papers, Number 1007, September 2010. Available at: http://www.federalreserve.gov/pubs/ifdp/2010/1007/ifdp1007.pdf (accessed 26 September, 2016).

Houseman, S.N., T.J. Bartik, and T.J. Sturgeon. 2014. "Measuring Manufacturing: How the Computer and Semiconductor Industries Affect the Numbers and Perceptions." Upjohn Institute Working Paper 14-209. Available at: http://research.upjohn.org/cgi/viewcontent.cgi?article=1226&context=up_workingpapers (accessed 26 September, 2016).

Howard, P.N., A. Duffy, D. Freelon, M. Hussain, W. Mari, and M. Mazaid. 2011. "Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?" Working Paper No.1, Project on Information Technology & Political Islam (PITPI). Available at: http://philhoward.org/opening-closed-regimes-what-was-the-role-of-social-media-during-the-arab-spring/ (accessed 28 March, 2014).

Independent Expert Advisory Group on a Data Revolution for Sustainable Development. 2014. "A World that Counts - Mobilising the Data Revolution for Sustainable Development." Available at: http://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf (accessed 30 March, 2016).

International Monetary Fund. 2009. "Balance of Payments and International Investment Position Manual – 6th Edition." Washington, D.C. Available at: https://www.imf.org/external/pubs/ft/bop/2007/pdf/bpm6.pdf (accessed 11 February, 2014).

International Telecommunication Union. 2015. "ICT Facts & Figures: The World in 2015." Available at: http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf (accessed 6 January, 2016).

Kennedy, R. 1968. 'Remarks at the University of Kansas, March 18, 1968'. John F. Kennedy Library, Presidential Library and Museum. Available at: https://www.jfklibrary.org/Research/Research-Aids/Ready-Reference/RFK-Speeches/Remarks-of-Robert-F-Kennedy-at-the-University-of-Kansas-March-18-1968.aspx (accessed 26 September, 2016).

Kirkpatrick, M. 2010. "Facebook's Zuckerberg Says the Age of Privacy is Over." Readwrite, 9 January 2010. Available at: http://readwrite.com/2010/01/09/facebooks_zuckerberg_says_the_age_of_privacy_is_ov (accessed 9 January, 2010).
Kitchin, R. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and their consequences*. London: Sage Publications Ltd.
Krikorian, R. 2013. "New Tweets per Second Record, and How!" Engineering Blog. Available at: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how (accessed 26 September, 2016).
Kuznets, S. 1962. "How to Judge Quality." *New Republic* 1962 issue, October 2029–31. Available at: http://www.unz.org/Pub/NewRepublic-1962oct20-00029 (accessed 26 September, 2016).
Landefeld, S.J. 2000. "GDP: One of the Great Inventions of the 20th Century." BEA Survey of Current Business, January 2000. Available at: http://www.bea.gov/scb/pdf/BEAWIDE/2000/0100od.pdf (accessed 26 September, 2016).
Layard, R. 2011. 'Happiness: New Lessons' in interview with Andrew Marr. Available at: https://www.youtube.com/watch?v=4VkQsL73SgE (accessed 26 September, 2016).
Levinson, M. 2006. *The Box: How the Shipping Container Made the World Smaller and the World Bigger*. Princeton: Princeton University Press.
Lewis, N.A. 2000. "The China Trade Wrangle: The Clinton Speech; Seeing Threats Beyond Arms, President Urges Cooperation." *New York Times*, U.S. edition, 8 May 2000. Available at: http://www.nytimes.com/2000/05/18/world/china-trade-wrangle-clinton-speech-seeing-threats-beyond-arms-president-urges.html (accessed 30 March, 2016).
Lillington, K. 2013. "Big Data is Watching You." Quoted in article by K. Lillington – *Irish Times: Weekend Review*, 27 April 2013.
Lynch, M. 2012. "Data Wars: Unlocking the Information Goldmine." *BBC News*, 12 April 2012. Available at: www.bbc.co.uk/news/business-17682304 (accessed 26 September, 2016).
MacFeely, S., R. Moloney, and M. Kenneally. 2011. "A Study of the NUTS 2 Regions Using Input-Output Analysis." *Journal of the Statistical and Social Inquiry Society of Ireland* XL: 60–115.
MacFeely, S. 2012. "Reflections on Developing a Measurement Framework for Economic Globalisation." Paper presented at the Eurostat Workshop on Global Value Chains and Economic Globalisation, Luxembourg, 9 & 10 October 2012.
MacFeely, S. and J. Dunne. 2014. "Joining Up Public Sector Information: The Rationale for a National Data Infrastructure." *Administration* 61: 93–107.
Mann, C.C. 2011. *1493: How Europe's Discovery of the Americas Revolutionised Trade, Ecology and Life on Earth*. London: Granta Books.
Mayer-Schonberger, V. and K. Cukier. 2013. "Big Data: A Revolution That Will Transform How We Live, Work and Think." London: John Murray.
Mazower, M. 2012. *Governing the World - The History of the Idea*. London: Allen Lane.
Menzies, G. 2003. *1421: The Year China Discovered America*. New York: Harper Collins Publications.
Museux, J.M., N. Hilbert, and G. Pongas. 2012. "Implementation of the ESS Joint Strategy: The Development and Promotion of the Plug and Play Concept as an

Architecture Model." Paper presented at Meeting on the Management of Statistical Information Systems (MSIS 2012), Washington, DC, 21–23 May 2012. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2012/presentations/Topic_1__07__Eurostat.pdf (accessed 26 September, 2016).

Organisation for Economic Cooperation and Development. 2010. "Measuring Globalisation: OECD Economic Globalisation Indicators 2010."

Osterhammel, J. and N.P. Petersson. 2005. *Globalisation: A Short History*. Princeton: Princeton University Press.

O'Rourke, K.H. and J.G. Williamson. 2002. "When Did Globalisation Begin?" *European Review of Economic History* 6: 23–50.

Palmer, E.Z. 1966. *The Meaning and Measurement of the National Income*. Lincoln, NE: University of Nebraska Press.

Paterson, C. 2009. "Rage Against the Machine Beat X Factor Winner in Charts." *BBC News*, 20 December 2009. Available at: http://news.bbc.co.uk/2/hi/8423340.stm (accessed 15 January, 2013).

Payton, T. and T. Claypoole. 2015. *Privacy in the Age of Big Data - Recognising the Threats Defending Your Rights and Protecting Your Family*. Lanham, MD: Rowman & Littlefield.

Pearson, E. 2013. "Growing Up Digital." Paper presented at the Big Data and Statistics New Zealand Seminar, Wellington, 24 May 2013. Available at: http://statisphere.govt.nz/seminars-training-forums/official-statistics-seminar-series/archived-presentations/big-data.aspx (accessed 9 September, 2013).

Penneck, S. 2014. "New Challenges Facing Official Statistics." *Statistical Journal of the IAOS* 30: 3–16.

Philipsen, D. 2015. *The Little Big Number - How GDP Came to Rule the World and What to Do About It*. Princeton, NJ: Princeton University Press.

Piketty, T. 2014. *Capital in the Twenty-First Century*. Cambridge, MA: The Belknap Press of Harvard University Press.

Popper, B. 2014. "Facebook Now has More Than a Billion Mobile Users Every Month." *The Verge*, 23 April 2014. Available at: http://www.theverge.com/2014/4/23/5644740/facebook-q1-2014-earnings (accessed 23 April, 2014).

Radermacher, W. 2013. "The European Statistics Code of Practice as a Pillar to Strengthen Public Trust and Enhance Quality in Official Statistics." *Journal of the Statistical and Social Inquiry Society of Ireland* 43: 27–33.

Reich, R. 2015. *Saving Capitalism: For the Many, Not the Few*. London: Icon Books Ltd.

Robertson, D. and B. Breen. 2013. *Brick by Brick – How Lego Rewrote the Rules of Innovation*. London: Random House.

Sachs, J. 2012. *The Price of Civilization: Reawakening Virtue and Prosperity After the Economic Fall*. London: Vintage, Random House.

Science Daily. 2013. "Big Data, for Better or Worse: 90% of World's Data Generated over Last Two Years." 22 May 2013. Available at: https://www.sciencedaily.com/releases/2013/05/130522085217.htm (accessed 22 May, 2013).

Shirky, C. 2011. "The Political Power of Social Media: Technology, the Public Sphere, and Political Change." *Foreign Affairs* January/February Issue 2011. Available at:

https://www.foreignaffairs.com/articles/2010-12-20/political-power-social-media (accessed 11 February, 2014).

Slade, G. 2006. *Made to Break: Technology and obsolescence in America*. Cambridge, MA: Harvard University Press.

Statistica. 2015. "Leading Social Networks Worldwide as of August 2015, ranked by Number of Active Users (in Millions)." *Statistica - the Statistics Portal*. Available at: http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed 9 November, 2015).

Stiglitz, J.E. 2002. *Globalisation and its Discontents*. London: Penguin Books.

Stiglitz, J.E. 2010. *Freefall: Free Markets and the Sinking of the Global Economy*. London: Allen Lane.

Stiglitz, J.E., A. Sen, and J.P. Fitoussi. 2010. *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Available at: http://www.cfr.org/world/report-commission-measurement-economic-performance-social-progress/p22847 (accessed 7 March, 2013).

Stiglitz, J.E. 2014. "America's GDP Fetishism is a Rare Luxury in an Age of Vulnerability - Traditional Measures Seem More Abstract Than Ever as They Fail to Recognise the Struggles of Wage-Earners in the Real World." *The Guardian*, 13 October 2014. Available at: http://www.theguardian.com/business/2014/oct/13/gdp-fetishism-luxury-vulnerability-economic-measures (accessed 30 March, 2016).

Struijs, P. and P. Daas. 2014. "Quality Approaches to Big Data in Official Statistics." Available at: http://www.pietdaas.nl/beta/pubs/pubs/Q2014_session_33_paper.pdf (accessed 11 August, 2015).

Sturgeon, T.J. 2013. *Global Value Chains and Economic Globalization - Towards a New Measurement Framework*. Report to Eurostat. Available at: http://ec.europa.eu/eurostat/documents/54610/4463793/Sturgeon-report-Eurostat (accessed 6 May, 2014).

Talberth, J., C. Cobb, and N. Slattery. 2007. "The Genuine Progress Indicator 2006 - A Tool for Sustainable Development." Redefining Progress, Oakland, CA. Available at: http://rprogress.org/publications/2007/GPI%202006.pdf (accessed 11 February, 2016).

Taleb, N.N. 2007. *The Black Swan – The Impact of the Highly Improbable*. London: Allen Lane.

The Economist. 2016. "The Trouble with GDP." 30 April 2016. Available at: http://www.economist.com/news/briefing/21697845-gross-domestic-product-gdp-increasingly-poor-measure-prosperity-it-not-even (accessed 9 September, 2016).

United Nations. 2012. "Secretary-General, in Message to Meeting on 'Happiness and Well-being' Calls for 'Rio+20' Outcome that Measures More than Gross National Income." SG/SM/14204, Department of Public Information, New York, 2 April 2012. Available at: http://www.un.org/News/Press/docs/2012/sgsm14204.doc.htm (accessed 26 September, 2016).

United Nations. 2014. Resolution adopted by the General Assembly on 29 January 2014 - Fundamental Principles of Official Statistics, General Assembly, A/RES/68/261. Available at: http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf (accessed 26 September, 2016).

United Nations. 2014. System of Environmental-Economic Accounting 2012 - Experimental Ecosystem Accounting. New York: United Nations. Available at:

http://unstats.un.org/unsd/envaccounting/seeaRev/eea_final_en.pdf (accessed 26 September, 2016).

United Nations Conference on Trade and Development. 2005. 'Transnational Corporations: More - yet more - on Globalisation, by J.H. Dunning', Vol. 14, No. 2, August 2005. UNCTAD/ITE/IIT/2005/5. Available at: http://unctad.org/en/Docs/iteiit20055a5_en.pdf (accessed 26 September, 2016).

United Nations Conference on Trade and Development. 2008. *Globalisation for Development: The International Trade Perspective*. Available at: http://unctad.org/en/Docs/ditc20071_en.pdf (accessed 26 September, 2016).

United Nations Conference on Trade and Development. 2012. *The Paradox of Finance-Driven Globalisation*. Policy Brief No. 1, 2012. Available at: http://unctad.org/en/Docs/uxiiipb2012d1_en.pdf (accessed 26 September, 2016).

United Nations Conference on Trade and Development. 2015. *Trade and Development Report, 2015 - Making the international financial architecture work for development*. Available at: http://unctad.org/en/PublicationsLibrary/tdr2015_en.pdf (accessed 26 September, 2016).

United Nations Economic Commission for Europe. 2009a. "A Guide to Writing Stories about Numbers: Part 1 - Making Data Meaningful." Geneva: United Nations. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/writing/MDM_Part1_English.pdf. (accessed 26 September, 2016).

United Nations Economic Commission for Europe. 2009b. "A Guide to Presenting Statistics: Part 2 - Making Data Meaningful." Geneva: United Nations. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/writing/MDM_Part2_English.pdf (accessed 26 September, 2016).

United Nations Economic Commission for Europe. 2011. "A Guide to Communicating with the Media: Part 3 - Making Data Meaningful." Geneva: United Nations. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/writing/MDM_Part3_English_Print.pdf (accessed 26 September, 2016).

United Nations Economic Commission for Europe. 2012. "Guide on the Impact of Globalization on National Accounts." Geneva: United Nations. Available at: http://www.unece.org/fileadmin/DAM/stats/publications/Guide_on_Impact_of_globalization_on_national_accounts__web_.pdf (accessed 26 September, 2016).

United Nations Economic Commission for Europe. 2012b. "A Guide to Improving Statistical Literacy: Part 4 - Making Data Meaningful." Geneva: United Nations. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/writing/Making_Data_Meaningful_Part_4_for_Web.pdf (accessed 26 September, 2016).

United Nations Economic Commission for Europe. 2013. "What does 'Big Data' mean for Official Statistics?" Paper presented by the Task Team established by the High-Level Group for the Modernisation of Statistical Production and Services. Available at: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614 (accessed 26 September, 2016).

United Nations Statistical Commission 2016. "Report of the Inter-Agency and Expert Group on Sustainable Development Goal Indicators - Note by the Secretary-General" presented to the 47th Session of the UNSC, March 8-11, 2016, New York. Doc

E/CN.3/2016/2/Rev.1. Available at: http://unstats.un.org/unsd/statcom/47th-session/documents/2016-2-IAEG-SDGs-Rev1-E.pdf. (accessed 30 March, 2016).

United Nations Statistics Division. 2009. "System of National Accounts 2008." New York: United Nations. Available at: http://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf (accessed 26 September, 2016).

United Nations Statistics Division. 2010. "Tourism Satellite Account: Recommended Methodological Framework 2008." *Studies in Methods*, Series F, No. 80/Rev.1. New York: United Nations.

United Nations Statistical Commission. 2016. *Report of the Inter-Agency and Expert Group on Sustainable Development Goal Indicators*, 47th Session, 8–11 March, 2016. Dec Ref: E/CN.3/2016/2/Rev.1. Available at: http://unstats.un.org/unsd/statcom/47th-session/documents/2016-2-IAEG-SDGs-Rev1-E.pdf (accessed 26 September, 2016).

Vale, S. 2012. "The Data Deluge: What does it mean for official Statistics?" Paper presented at the 2nd CSO Administrative Data Seminar, Dublin Castle, 29 September 2011. Available at: http://www.cso.ie/en/newsandevents/eventsconferencesseminars/administrativedataseminars/2ndadministrativedataseminar/ (accessed 21 February, 2014).

Vale, S. 2013. "Fostering Interoperability in Official Statistics: Common Statistical Production Architecture." Posted on UNECE Statistics/MSIS Wiki on 12 February 2013. Available at: www1.unece.org/stat/platform/display/msis/2013/02/12/Plug+and+Play+project+launched (accessed 26 September, 2016).

Yergin, D. 2012. *The Quest – Energy, Security and the Remaking of the Modern World*. London: Penguin Books.

Received September 2014
Revised March 2016
Accepted August 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 811–826, http://dx.doi.org/10.1515/JOS-2016-0042

# Helping Raise the Official Statistics Capability of Government Employees

*Sharleen Forbes*[1] *and Alan Keegan*[2]

Both the production and the use of official statistics are important in the business of government. In New Zealand, concern persists about many government advisors' low level of statistical capability. One programme designed specifically to enhance capability is New Zealand's National Certificate of Official Statistics, first introduced in 2007 and originally targeted at government policy analysts and advisors. It now includes participants from many agencies, including the National Statistics Office. The competency-based 40-credit certificate comprises four taught units that aim to give students skills in basic official statistics and in critically evaluating statistical, research, policy, or media publications for their quality (of data, survey design, analysis, and conclusions) and appropriateness for some policy issue (e.g., how to reduce problem gambling), together with an 'umbrella' workplace-based statistics project. Case studies are used to embed the statistics learning into the real-world context of these students. Several surveys of students and their managers were undertaken to evaluate the effectiveness of the certificate in terms of enhancing skill levels and meeting organisational needs and also to examine barriers to completion of the certificate. The results were used to both modify the programme and extend its international applicability.

*Key words:* Raising statistical capability; certificate in official statistics; programme evaluation.

## 1. Introduction

Government benefits from having employees that are able to use, understand and generate official statistics. The New Zealand Prime Minister's personal science advisor, Dr. Peter Gluckman, drew attention to the need for clear monitoring and evaluation of key policies and programmes in New Zealand, and to a lack of capability in the state sector to achieve these tasks. He stated that "the costs and implications of inferior science or wrong data leading to policy decisions are immense" (Gluckman 2011, 15) and also emphasized that all governments want the best outcomes from their investments and need good data. Some of these data are in government-owned administrative data sets, many of which (for example in health and education) are used to generate official statistics. Although basic statistical capability is one of the skills that newly recruited government advisors are

[1] Victoria University - School of Government, P.O. Box 600 Wellington, Wellington 6011, New Zealand. Email: sharleen.forbes@gmail.com
[2] Statistics New Zealand - Statistics Education, Wellington, New Zealand. Email: alan.keegan@stats.govt.nz
**Acknowledgments:** This article is based upon, but extends, an earlier version called "Bringing the Workplace into a National Certificate in Official Statistics" that was presented at the Ninth International Conference on Teaching Statistics in Flagstaff, Arizona, in July 2014.

expected to have (Forbes 2011), consultation both by the National Statistics Office, Statistics New Zealand (Forbes 2009), and by the School of Government at Victoria University (Forbes 2011) with state-sector statisticians, policy, and senior managers (including chief executives) identified variable or insufficient statistical skills in some agencies.

As a result, using his statutory coordination role for official statistics under the Statistics Act 1975, Section 3(2) (New Zealand Government 1975), the New Zealand government statistician decided to help raise the statistical capability of state-sector employees through investment in a joint general manager (Statistics Education) position within Statistics New Zealand and an academic (adjunct professor) position in the School of Government at Victoria University. Collaborative development with academics at other universities resulted in a suite of training opportunities, including a National Certificate in Official Statistics for state-sector employees and a postgraduate (Honours) course in official statistics that uses advanced video-conferencing, allowing student interaction and input from computers or lecturers at multiple sites. The Honours course involved teaching staff from different New Zealand universities who had specialist expertise in particular areas of official statistics, such as demography, data matching, survey design, social statistics, macroeconomic statistics, and so on. In the 2014 academic year, the course had 50 students participating at six New Zealand universities (Harraway and Forbes 2013). However, the focus of this article is on the National Certificate of Official Statistics. One of the features of this certificate is that it is targeted at state-sector users of official statistics, unlike courses such as the Postgraduate Certificate/Diploma and Masters in Official Statistics developed by the University of Southampton and the Government Statistical Service in the United Kingdom (2015) or the courses offered by UNSIAP (United Nations Statistical Institute for Asia and the Pacific) that target producers of official statistics (usually employees of national statistics offices).

Several surveys of students and/or their managers were used to determine whether the certificate meets the needs of current participants and their managers and what the barriers to completion of the certificate were. Modifications have since been made to the structure and content of the certificate, some of which was designed specifically to increase its international applicability.

## 2. National Certificate in Official Statistics

The National Certificate in Official Statistics was developed in 2007 based on statistical thinking theory, but focused on official statistics as well as general statistics methods (Forbes 2009; Forbes et al. 2010). A group of academics from statistics departments at the seven main New Zealand universities were involved in its design and implementation, and several continue to be involved in teaching and assessment. The certificate is at Level 5 (approximately equivalent to the first year of university studies) on New Zealand's vocational training framework, the New Zealand Qualifications Framework (National Qualifications Framework Project Team 2005), and is competency based (no grades are awarded and students can resit questions or units until they reach the required standard). From its inception, the certificate comprised four taught units (called Unit Standards in the New Zealand context):

1. *Resolve ethical and legal issues in the collection and use of data in a public-sector context.*
   Includes understanding and interpretation of national official statistics and other relevant (e.g., privacy) legislation, internationally agreed (e.g., United Nations Statistics Commission) principles with a focus on confidentiality and security of data, and other ethical issues that should be considered (such as cultural or response-burden concerns)
2. *Interpret statistical information to form conclusions for projects in a public-sector context.*
   Covers basic descriptive statistics and graphs for univariate and bivariate combinations of categorical and numeric variables, simple time series and index numbers (including rebasing and deflation) and demographic techniques (age standardisation and odds ratios) together with the appropriate use of these to address a given policy question.
3. *Assess a sample survey and evaluate inferences in a public-sector context.*
   Includes description and assessment of administrative and survey data collections, simple estimation (calculation of 95% confidence intervals) and interpretation of simple inferential statistics (e.g., t-tests and Chi-squared tests).
4. *Evaluate and use statistical information to make policy recommendations in a public-sector context.*
   Using their learning from the previous three units, students evaluate the appropriateness of different data collections to help answer a given policy question and also describe the properties, including simple survey design, of appropriate new data collections.

The teaching for all these units takes place in one or two day blocks in a traditional classroom setting using small workshops. The units were usually spaced six to eight weeks apart to give students time to complete the assessment for one unit before doing the learning for the next. Both the learning and assessment focus on evaluating real statistical, research, policy, and media publications. Two main 'case study' publications (official statistics or other government agency releases, research reports, or media articles) are chosen for use across all four units for teaching purposes and two different publications are chosen for assessment purposes. The four units account for 24 of the total 40 credits needed to gain the certificate, with each credit representing about ten hours of work. In 2007 and 2008, students could select the remaining 16 credits from a set of level four or five units registered on the framework that were deemed to be appropriate for the public-sector context, such as management and communication skills or knowledge of public-sector processes. This was changed in 2009 to a 16-credit research project based in the student's own workplace. One reason for this was that students could then make an immediate transfer of at least some of the skills learned in the taught units to their individual workplace and demonstrate the usefulness of these skills in their day-to-day work. As Vaughan (2008) states: "*Workplace learning has a broader project and potential to link development of the individual with development of the organisation or business*, through an emphasis on sustained development and learning processes as well as learning outcomes" (p. 1). In adult education, action learning is based on the premise that "adults should have control over the content and form of their education" (Pant 2006, 95) and the workplace-based project could be viewed as "an integrated activity that combines social investigation, educational work, and action" (International Council of Adult Education, cited in Pant 2006, 97) with students being able to negotiate with their managers and work on a topic of their own choosing. From the manager's perspective, it can also be a tool through which the organisation may derive new knowledge.

Each cohort of students is designed to be small (15–20 students) and all students are in current employment. In some, but not all years, there was more than one cohort of students. By 2013 there had been ten cohorts (nine in New Zealand and one in Tonga) with 185 students attending taught units. Students came from 37 different government and local authority agencies. However, not all were formally enrolled and some of those enrolled withdrew before attempting any assessments. Only 144 of the New Zealand students had signed training agreements. In recent years, students who just wanted the learning from a taught unit, but did not want to do the assessment, were able to pay a reduced fee. Between one to seven students annually have done this, with the most popular unit enrolled in being *Evaluate and use statistical information to make policy recommendations in a public-sector context.*

It was intended that students complete the certificate within twelve to eighteen months. However, students are notoriously late in completing assessments, with a few taking several years to gain the full qualification. If we consider just the early (2007–2010) cohorts, the mean completion rate for the certificate is low (from 60–71%, mean = 65%). There are also a number of students (about ten percent) who complete all the four taught units but not the final 16-credit research component. One way of increasing completion rates could be to assist these students to finish their research projects.

As this research project is based in the student's own workplace, the level of statistics it contains is highly variable. But at a minimum, the project should contain graphs and at least one appropriate bivariate statistical analysis such as: relationships between pairs of numeric or categorical variables, confidence intervals for means or proportions, interpretation of hypothesis tests ($p$-values) or investigation of time series. Examples of successfully completed projects are:

- Factors affecting participation in a government-funded children's Internet survey
- Differences between Egyptian and New Zealand societies
- Comparing religious affiliation for Tongans in the New Zealand and Tongan Censuses
- Analysis of travel by preschool children using New Zealand Household Travel Survey data
- Comparison of local and international prices in Tonga
- Survey of SAS ECO users in Statistics New Zealand
- Survey of transport issues for residents in a new subdivision in Christchurch, New Zealand
- Differences between the CPI and Statistics New Zealand's Household Consumption Expenditure Implicit Price Deflator.

As reported in Forbes et al. (2010), the first cohort of students participating in the certificate was viewed, in part, as an introductory pilot. These students were surveyed in

2007, then in 2009 this survey was updated to include all those students (and their managers) who had enrolled in the certificate by mid-2009 (four cohorts of students). Another survey of the New Zealand students that started the certificate after the research project was made a compulsory part of the certificate (2009-2012 cohorts) was undertaken in late 2013. This was followed by a second survey of these students' managers to investigate their perceptions of the value of the certificate and barriers to its completion. These are important issues as all the students have their course costs approved by their managers and paid by their organisations.

## 3. Survey Method

Overall, the above three surveys (each comprising several student cohorts) of students and their managers were used to investigate three research questions:

1. Does the certificate programme meet the needs of students?
2. Does the certificate programme meet the needs of students' managers?
3. What are the barriers to students completing the certificate?

In all three surveys, only students formally doing the certificate (taking part in assessments) and their managers were asked to respond. Statistics New Zealand's Questionnaire Methodology and Development (QMD) team provided some assistance with questionnaire creation and pretesting of questions.

### *3.1. Pilot and 2009 Surveys*

The views of the 13 students in the first (2007) cohort that had completed at least one taught unit were surveyed using a structured questionnaire with open-ended responses. Students were asked about their reasons for enrolling in the certificate, their prior statistics learning, perceived barriers to completing the certificate, and whether they would recommend the certificate to others. Each student's manager was also surveyed for their expectations of the programme, support that they gave to students, perception of the statistical skills of students prior to enrolment and whether they would recommend the certificate to others. Managers were also asked about what improvements in staff performance or confidence they expected as a result of the certificate and whether they had any evidence of these being achieved. The course assessor was also asked to comment on possible improvements, and the level of complexity of the assessment questions in each taught unit was analysed using a method designed by Black and described in detail in Forbes et al. (2008), where each question received a score according to the following increasing level of complexity:

1 = Idiosyncratic,
2 = Verbal,
3 = Transitional,
4 = Procedural,
5 = Integrated Process.

This pilot survey was updated to include all students (and their managers) who had enrolled in the certificate by mid-2009. The major difference was that the students in the

2007 pilot were interviewed, whereas the 2009 survey used self-completed written questionnaires that could either be filled in online or downloaded and completed by pen and paper.

As both the pilot and update surveys used the same questions, the results were combined to give comparable results for all 58 students. The overall response rate of the combined evaluation was 62% for both students (36/58) and managers (21/34 managers responding for 27 students). However, it was not necessarily the case that the responding managers were managers of responding students.

### 3.2. 2013 Surveys

The online software SurveyGizmo was used to manage the 2013 surveys. The questionnaire (see Supplemental material online at: http://dx.doi.org/10.1515/jos-2016-0042) contained a mixture of four-point Likert scale, simple multiple-choice (usually yes/no or range bands such as 'none', 'at least once', '2–4 times', etc.), and open-ended questions. Some questions were adapted from previous course surveys. Features of the questionnaire included: automatic skipping of questions, breaking up the survey into sections with headings, progress bars, greying out of choices when the "none of these" option is chosen, and soft compulsion where respondents are reminded once about unanswered questions. Soft compulsion was used more often at the start of the questionnaire, where questions that were of higher priority were placed. It was not possible to automatically link students and managers nor determine whether nonresponding students had moved jobs or had completed the certificate.

#### 3.2.1. Student Survey

Given that some students complete all the taught units but not the research project, the primary motivation for this survey was to identify potential barriers to completion of this project. Email addresses from students' training agreements were used to send a link to an online questionnaire. These were mainly workplace email addresses. Any alternative addresses given in automated responses were also used, but only cursory Internet searches were made to find other new email addresses. The first email invitation only yielded a response rate of ten percent. Two further reminder emails increased this to 30% overall (20 students). A further eight people answered about half the questions in the survey (that is, there can be up to 28 responses on any given question). The low response rate and out-of-date sample frame mean there are a number of potential sources of bias. Students with stronger opinions or those with time available to answer the survey may have been more likely to respond. There was also lower representation of earlier cohorts (as shown in Table 1, which demonstrates that the response rates were related to the time delay between the cohort and the survey).

#### 3.2.2. Manager Survey

The design of the managers' survey was very similar to the students' survey. The same sample frame and method to select students was used, as all students had been asked to supply their managers' details and some managers had also attended workshops about the project component of the certificate. Response rates by cohort year for the managers were eight percent for 2009 cohorts, 60% for 2010, 27% for 2011, and 47% for 2012, with one

Table 1. Students and managers response rates by year.

| Year | Student cohort size | Students | | Managers | |
|---|---|---|---|---|---|
| | | Respondents | Response rate (%) | Respondents | Response rate (%) |
| 2012 | 23 | 10 | 43.5 | 9 | 47 |
| 2011 | 16 | 6 | 37.5 | 4 | 27 |
| 2010 | 14 | 2 | 14.3 | 6 | 60 |
| 2009 | 15 | 2 | 13.3 | 1 | 8 |

additional respondent whose year was unknown. As shown in Table 1, the managers' response rates were not as directly related to the time delay between the cohort and survey as the students' response rates. The overall response rate for managers was 42% (21 respondents), slightly greater than that for the students. Just under a quarter (5/21) of responding managers had more than one staff member do the certificate. In most cases this would have been two staff, so overall the managers were reporting on their experiences with a total of at least 26 students.

Identical questions about barriers to completion were used in both surveys to enable comparison of the perceptions of managers and their staff. Given the sources of potential bias, in the main, results have been reported for the responding students and managers only. If this were to be treated as a random sample of all the students and their managers, the margins of error would be substantial. For example, for a proportion of 20 students it is roughly (with finite population correction) plus or minus 18 percentage points. The results given below should therefore be viewed as indicative only.

## 4. Survey Results

### *4.1. Pilot and 2009 Surveys*

The most common reasons for enrolment given by students from the combined 2007 and 2009 surveys were to improve their statistical knowledge and promotion prospects, or as a refresher course. Students that enrolled to increase their statistics knowledge or assist with career advancement were classed as 'high' motivation, and those who wanted either a refresher, to "contribute to the pilot" or because their manager suggested it were classed as 'low' motivation. Almost two thirds of students (65%) were classified as having high motivation. Unsurprisingly, students with little or no prior knowledge of statistics found the statistical content difficult compared to those with some prior knowledge, but there was no obvious relationship between students' motivation and their prior statistics knowledge.

Only 60% of students completed the certificate within the suggested time (one year). The time it took for a student to complete a taught unit was found to be related to the student's motivation to do the qualification, with 92% of high-motivation and 34% of low-motivation students completing units. However, prior statistical knowledge did not seem to be related to completion time. About half the students reported that one of the barriers to completion was balancing doing assessments with commitments in their work and personal life.

Seven people moved into a new role or organisation during or after the certificate. Only one was explicit that the certificate contributed to their move, stating: "Changed job after completing certificate. Knowledge from certificate is very useful in new job, and encouraged me to apply for new job." All seven of those who changed workplaces midway through the certificate completed it (contrary to the findings of Curson (2004) that this is often a barrier to course completion).

All but one student stated that they would recommend the certificate to others. All the responding managers also said that they would recommend the certificate. While most managers indicated that it was too soon to determine whether the certificate had met their expectations and had a 'wait and see' attitude regarding its impact in the workplace, two thirds (14/21) also stated that there was a noticeable increase in the confidence of staff enrolled in the certificate.

### *4.2. 2013 Student Survey*

Responding students came into the certificate with highly variable backgrounds in statistics: 21% (6/28) had never studied statistics, 54% (15/28) studied statistics either in senior secondary school or introductory university courses, 14% (4/28) at advanced tertiary level and eleven percent (3/28) as professional development for their job. For a number of students, their prior learning had taken place more than four years ago. Multiple responses were possible to the question on why students had enrolled in the course. Of the 28 students responding to this question, 46% (13/28) did so because they wanted to be more confident using official statistics and 43% (12/28) because they wanted to learn about official statistics. Only just over 14% (4/28) of the responding students enrolled in the certificate because their manager told them to (Figure 1).

As with the taught units in the 2009 evaluation, completion of the project was related to its perceived priority. Of the 20 students responding to this section, almost half (9/20) did not start working on the project until after finishing the four taught units. A quarter had not yet started doing the project. However, of the twelve that had not yet started or completed the project, 75% (9/12) still intended to complete it. A quarter (of the 20 respondents) indicated that they had changed role or organisation since beginning the certificate. It is possible that this may have a bearing on future completion of the research project.

A high proportion (over 90%, 20/22 or 21/22 respondents for each unit) of the students agreed or strongly agreed that each of the taught units was relevant to the research project As one student stated, "I'm trying to incorporate all the topics from the course into my research project in some aspect. In that case, all topics are relevant to my research project".

Of the students that had started their project, 27% (4/15) had already applied the skills and knowledge gained in the research project to a moderate extent in their workplace. Students' comments about how they had applied these skills in the workplace included:

- "Applied sampling theory from the course in scoping an observational cell phone use survey." not
- "By developing collection reports with some analysis and recommendations from survey metadata."
- "Interpreting and explaining margins of errors, odds ratios etc. for research papers."
- "To harp on to others about data matching and information privacy."

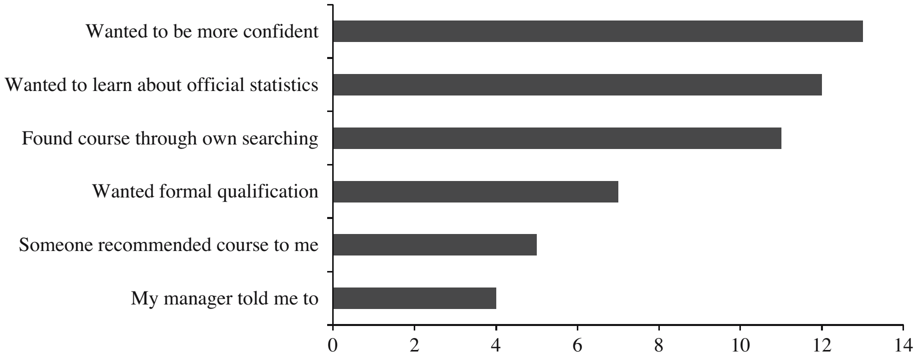


*Fig. 1. Numbers of students giving reasons for enrolling in the Certificate of Official Statistics.*

- "I am now able to analyse data to a deeper extent and have more of an understanding of the information that I send to others. It has given me more confidence in what I do and I can inform others with confidence."
- "I started a mini-project (currently on hold) as suggested by my new manager to check if a certain source of administrative data is good enough to be used as quality check for current data."

All the responding students (100%) felt that they were well informed about the research project and 86% (13/15) also considered their managers to be well informed.

### *4.3. 2013 Manager Survey*

While only 14% of students reported enrolling in the certificate because their manager requested them to, 19% (4/21) of managers reported requesting students to take part. However, it should be noted that the responding managers may not necessarily be the managers of the responding students, as these groups were surveyed separately. Given the small numbers, and different time periods at which the surveys were carried out, no pairing of managers with their students was undertaken.

All managers reported having expectations of their staff prior to staff doing the certificate, with most managers (86%, 18/21) expecting increased skills and two thirds (67%, 14/21) expecting their staff to have increased awareness. But only just under half (48%, 10/21) expected their staff to be able to think and/or work statistically as a result of completing the certificate. One manager reported that they expected an "increased ability to connect knowledge of stats with what is happening in our business".

The students had all completed some units in the certificate when managers were surveyed. At this point, about two thirds of the managers (67% (14/21), 62% (13/21) and 71% (15/21) respectively) reported that the course had improved their staff's confidence in statistics, their basic statistical skills and knowledge or their awareness of official statistics. In addition, 52% (11/21) reported that their staff's ability to think and/or work analytically had increased and 29% (6/21) stated that staff had improved in other areas. Less than ten percent (2/21) of managers thought that their students needed more support with the statistical content in the taught courses and only 14% (3/21) that they needed more support with the statistical content of the research project. Open-ended comments made by the managers included that:

- "Completing the certificate was a valuable experience for my staff member. I think it increased her personal confidence with working with statistical information and whetted her appetite for further learning."
- "Found it a useful course that I would recommend to a new analyst with limited statistical knowledge and/or knowledge of the OSS (Official Statistics System)."

One manager also commented on the need for follow-up after the certificate was completed, asking "where to after the research project?", and suggesting that discussions could take place "with other parties about the data to better understand it".

### *4.4. Comparison of Common Questions*

Although some of the responding managers might not have managed responding students, and similarly some of the responding students might not have been managed by responding managers, the responses for common questions in both the students' and managers' surveys were compared.

The course organisers expected that students would have discussions with their manager about the workplace-based research project in particular. Both the students and the managers were asked what discussions had taken place (Table 2).

Most of the students had discussed the research topic and getting data with their manager, but only half discussed choosing methods, creating statistical outputs, their first draft, or submission of the assignment with their manager. Some students also disagreed that their discussions with their manager were productive (40%, 8/20). Just under two thirds (62%, 13/21) of managers reported reviewing the project with their staff. The proportions of managers who reported having had a discussion were lower than the students in all cases except writing a first draft. Managers were more likely to respond N/A (not applicable) in each of these categories than students were. Some managers surveyed had inherited staff doing the course, which would have increased the number of N/A responses. Chi-squared and Fisher's exact test for odds ratios indicated no statistically significant differences between any of the responding managers' and students' responses.

Identical questions were also asked about perceived barriers to completion of the research project. As Figure 2 shows, there was a difference between what managers and students reported. Figure 2(a) gives the managers' perceptions of barriers to completion, and Figure 2(b) gives the students' perceptions.

*Table 2. Percentage of respondents reporting discussion about aspects of the research project.*

| Number of respondents | | Choosing a topic (%) | Getting data (%) | Choosing statistical methods (%) | Creating statistical outputs (%) | Writing a first draft (%) |
|---|---|---|---|---|---|---|
| Students | 20 | 80 | 80 | 50 | 50 | 50 |
| Managers | 21 | 71 | 52 | 29 | 24 | 57 |

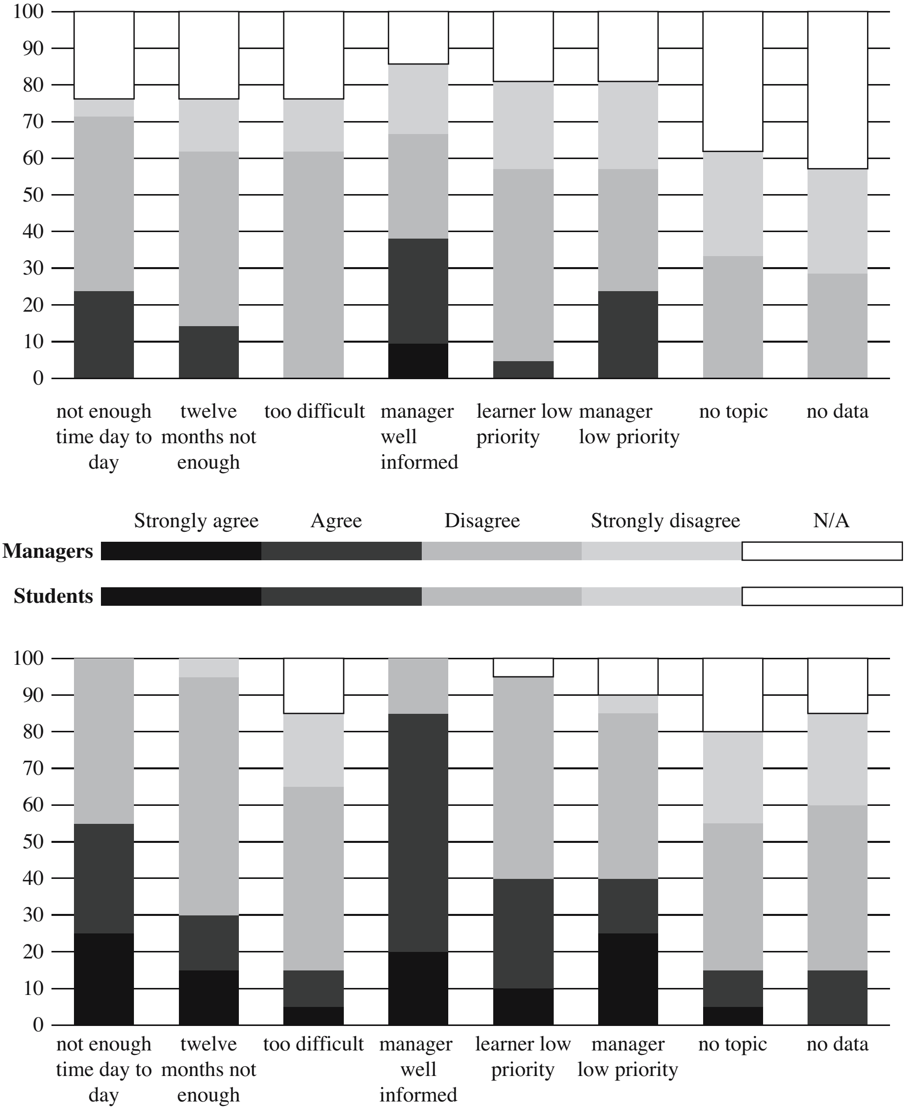


*Fig. 2. Perceptions of barriers to completion of the research project: (a) Managers (b) Students.*

The 95% confidence interval for the odds ratios are given in Table 3. An odds ratio of less than 1 implies that the students were more likely than the managers to select a barrier and an odds ratio of more than 1 that they were less likely. Confidence intervals that contain 1 are not statistically significant. Managers were again more likely than students to tick the N/A response. The results indicate that the students were more likely than the managers to say that the manager was well informed about the project or that the research project was a low priority for the student. That is, managers and students had similar perceptions on whether the project was a low priority for the manager, but not on whether it was a low priority for the student. Neither managers nor students thought that the certificate was too difficult or that twelve months was not long enough.

*Table 3. 95% confidence interval for odds ratios of likelihood to perceive a barrier.*

| Perceived barrier | 95% CI for odds ratio |
|---|---|
| Not enough time day to day | (0.05, 1.15) |
| Twelve months not long enough | (0.05, 2.27) |
| Too difficult | (0.00, 2.23) |
| Manager well informed about project | (0.02, 0.58) |
| Low priority for the student | (0.00, 0.71) |
| Low priority for the manager | (0.10, 2.16) |
| No topic was available | (0.00, 2.23) |
| No data was available | (0.00, 2.23) |

## 5. Lessons Learned and Modifications Made to the Certificate

### *5.1. Original Research Questions*

#### 5.1.1. Meeting Students' And Managers' Needs

Despite the students' variable statistical background, very few thought that the content of the certificate was too difficult. Feedback from both the 2009 and 2013 surveys indicated that students felt well informed, that almost all would recommend the certificate to others, and that some had started applying their new skills in the workplace while still partway through the certificate, suggesting that the certificate did meet students' needs. Feedback from students also reinforced the relevance of the material in the taught units to their workplace-based research projects. Managers also indicated that at least some of their needs were being met with about two thirds in both surveys reporting increased staff confidence. Managers thought the content was not too difficult and all would recommend the certificate to others. Two thirds of the managers surveyed in 2013 also reported improved basic statistics skills, and almost three quarters reported increased knowledge and awareness of official statistics. That almost a quarter of these managers had more than one student doing the certificate may also imply some satisfaction with the certificate.

#### 5.1.2. Barriers to Completion

Some barriers to completion of the certificate were easy to identify, including the time constraints imposed by other commitments, delaying the start to the research project until after all the taught units were completed (about half the students), and the project being a low priority for students and/or their managers. High enrolment motivation was also related to completion of the certificate. Some factors, such as the students' prior level of statistics or movement between jobs, did not appear to be related to course completion. However, there is clearly room for improvement in terms of increased completion rates.

### *5.2. Modifications Made to the Certificate*

The feedback obtained from the 2007 pilot, together with concerns expressed by the then assessor and the analysis of the level of complexity of the assessment questions, resulted in a number of changes to the certificate. These included: the order of delivery of the taught units, the teaching style of one teacher, assessment questions (number of questions reduced and questions ordered by difficulty); worked examples (these step-by-step descriptions of how to answer questions were fine-tuned to show students what was required for a pass); and tutoring and mentoring systems (extended).

Following the 2009 update, a 16-credit 'umbrella' workplace-based statistics project was included so that there was direct transference of the learning from the taught units back to the respective student's workplace.

As one of the major lessons learned was that students often delayed the start of their research project until they had completed all the taught units, in the first cohort after the 2013 survey a programme to encourage students to begin work on their research project early was introduced. This included: providing information about the project at the very beginning of the certificate and reiterating this in each taught unit; bringing managers together in short (1–2 hour) workshops about the project and requiring students to submit research plans in the first few months of starting the certificate. Early indications suggest that these initiatives are already having a positive impact, in particular the last, with all but one of the students in the latest cohort completing a research plan by the required date. In a number of cases, the submission of one-page plans for the research made it clear to the course coordinator where students needed guidance, especially in terms of the type and level of statistical analysis. The impact of these initiatives on completion rates will continue to be monitored. Both the course material and the assessment questions were also reviewed following the 2013 survey.

### *5.3. International Extensions*

Following the 2013 survey, two new initiatives involving the Certificate of Official Statistics were developed.

The first was the result of the Tongan cohort that took place early in 2012. Initially 20 students from six Tongan Government agencies (eight from policy ministries or the Ministry of Finance and National Planning, and twelve from the Statistics Department) enrolled, but only 18 submitted assessment material. These students were not included in the New Zealand surveys but also had a completion rate for the assessments that dropped off substantially over time. At the end of 2013, 17 students had achieved one or more of the taught units and six were working on or had passed the final research project unit (and were therefore heading towards completing the certificate). One outstanding student subsequently went on to do postgraduate studies. Unlike the New Zealand students, a number of the students who had partially completed the certificate withdrew because they had changed jobs. Several others also partially completed because they moved to New Zealand to live. Although this cohort may be viewed as only moderately successful overall, for many students it was their first post-school qualification, and it was used as the basis of a proposal to the New Zealand Ministry of Foreign Affairs that provides aid to Pacific Island countries. Funding was subsequently approved in mid-2014 to develop and deliver the certificate, adapted for the Pacific context, to up to 100 students in four Pacific Island countries. This is now underway.

Not all students participating in the taught units were seeking formal qualifications. Students also found it hard to balance learning with work and family commitments. One possible way of addressing these issues is to have learning available in different formats. A second joint endeavour involved New Zealand academics, John Harraway from the

University of Otago and Sharleen Forbes from Victoria University, and staff from the Royal Statistical Society Centre for Statistical Education at Plymouth University. This project aimed to develop problem-focused (rather than technique-focused) e-learning material to enhance and replace some of the written material on the certificate. The resulting web apps are freely available, accessible on a variety of IT platforms (e.g., desktop computers and laptops, tablets, and smartphones) and have interactive content, making them less like static e-books and more like a miniature learning environment including questions, quizzes, animation, videos and interactive tables and graphs. The first three web apps developed were:

1. Measuring Price Change (focusing on the CPI, working with price indices, change of base, time series in connection with the CPI, moving averages, trends, seasonality, and policy uses of price indices)
   http://iase-web.org/islp/apps/gov_stats_priceindices
2. Comparing populations (over time, between countries and between groups within countries including aspects of demography such as fertility, mortality, migration, life tables, population pyramids, age standardisation, and odds ratios)
   http://iase-web.org/islp/apps/gov_stats_populations
3. Graph It in Excel (discussing good and misleading data presentation and giving instructions for the creation of simple graphs including boxplots and population pyramids). This web app was developed because many state-sector policy advisors only have access to Microsoft Office products in their workplace.
   http://iase-web.org/islp/apps/gov_stats_graphing

The three web apps were officially launched at the 2015 World Statistics Congress and are now hosted on the international Statistical Literacy Project website. They can be accessed using the URLs given above. National Statistics Offices in several African countries (including Mozambique, Angola, Malawi, Gambia, Kenya, and South Africa) have already expressed interest in using the web apps. UNSIAP also plans to use them in a workshop in Bangkok in 2015 for statisticians from National Statistics Offices in Asia and Africa.

## 6. Concluding Comments

In its current form, the New Zealand Certificate of Official Statistics is a mix of externally taught and assessed units and a workplace-based, but externally assessed, project. As Vaughan (2008, 19) states, *"Workplace learning is not just a one-way process then. It is an interaction between workplace, learning, and the learner"*. That is, in this certificate students need to interact with their managers when choosing the research project and throughout its development. Possible reasons for the perceived managers' low priority for the project could be they see it as additional to their negotiated annual work plans, or they feel that they have contracted out the students' learning in the certificate through payment of their enrolment fees. An ongoing issue for the workplace-based research component of the certificate is getting managers to accept that "Workplace learning projects require substantial investments of resources, commitment and energy from both employers and employees" (Alcántara 2006, 1). Extra effort has been required to remind managers to allocate sufficient work time for both themselves and their students in order to gain maximum return from their investment in this training.

While a number of initiatives have been introduced to help improve the completion rate in the certificate, what has not been measured yet is whether or not participation in this national certificate has a positive effect on organisational performance, decision-making processes, or statistical outputs in the agencies where students work. This is an area for further research.

## 7. References

Alcántara, L. 2006. "Expanding Effectiveness of Workplace Learning Projects Through Multi-Level Application of Action Technologies." In Conference papers of the Seventh International Conference on HRD Research and Practice across Europe, May 2006. Tilburg University, Department of HR Studies, Netherlands. Available at: http://www.ufhrd.co.uk/wordpress/wp-content/uploads/2008/06/10-2_alcantara.PDF (accessed October 2013).

Curson, R. 2004. *Completion Issues in Industry Training*. Wellington, New Zealand: Industry Training Federation.

Forbes, S. 2009. "Creation and Evaluation of a Workplace Based Certificate in Official Statistics for Government Policy Makers." In Proceedings of the International Association of Statistics Education (IASE) Satellite Conference, Next Steps in Statistics Education, 14-15 August 2009. Durban, South Africa. Available at: http://iase-web.org/documents/papers/sat2009/9_2.pdf (accessed November 2015).

Forbes, S. 2011. "Collaboration and Cooperation: the Key to Reaching Out." In Proceedings of the International Association of Statistics Education (IASE) Satellite Conference, Statistics Education and Outreach, 18-19 August 2011. Malahide, Dublin. Available at http://iase-web.org/documents/papers/sat2011/IASE2011Paper6.1Forbes.pdf (accessed November 2015).

Forbes, S., M. Black, K. Moses, N. Pihama, N. Bechervaise, and K. Cullen. 2008. *Towards Retention: the Role of Assessment, Motivation and Prior Knowledge in Learner Completion. Report for Ako Aotearoa*. Wellington, New Zealand: Ako Aotearoa.

Forbes, S., P. Bucknall, and N. Pihama. 2010. "Helping Make Government Policy Analysts Statistically Literate." In Proceedings of the 8th International Conference On Teaching Statistics (ICOTS8), July 2010. Ljubljana, Slovenia. Available at: http://iase-web.org/documents/papers/icots8/ICOTS8_6E3_FORBES.pdf (accessed September 2016).

Gluckman, P. 2011. *Towards Better Use of Evidence in Policy Formation: A Discussion Paper*. Office of the Prime Minister's Science Advisory Committee. Wellington: New Zealand Government.

Harraway, J.A. and S.D. Forbes. 2013. "Partnership Between National Statistics Offices and Academics to Increase Official Statistics Literacy." *Statistical Journal of the IAOS* 29: 31–40. Available at: http://content.iospress.com/articles/statistical-journal-of-the-iaos/sji00761. Doi: http://dx.doi.org/10.3233?SJI-130761.

National Qualifications Framework Project Team. 2005. *The New Zealand National Qualifications Framework, Revised paper*. Wellington: New Zealand Qualifications Authority.

New Zealand Government. 1975. *Statistics Act 1975*. Wellington: New Zealand Government. Available at: http://www.legislation.govt.nz/act/public/1975/0001/latest/DLM430705.html (accessed September 2016).

Pant, M. 2006. "Unit 1. Introduction to Participatory Research and Unit 2. Conceptual Understanding Participatory Research." In *Certificate in International Perspective in Participatory Research*. New Delhi, PRIA (Society for Participatory Research in Asia).

University of Southampton and Government Statistical Service in the United Kingdom, 2015. http://www.southampton.ac.uk/demography/postgraduate/taught_courses/msc_official_statistics.page (accessed September 2016).

Vaughan, K. 2008. *Workplace Learning: a Literature Review: Report Prepared for Competenz*. Wellington: New Zealand Council for Educational Research.

Received September 2014
Revised March 2016
Accepted August 2016

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 1–7, http://dx.doi.org/10.1515/JOS-2016-0042

# Certificate of Official Statistics - Student Survey

## Prior to enrolment

hidden=false&rec

**1. Why did you enrol in this course? Check all that apply.** ***This question is required.**

- [ ] My manager told me to
- [ ] I heard about this course from someone else
- [ ] I found this course through my own searching
- [ ] I wanted to be more confident using official statistics
- [ ] I wanted to learn about official statistics
- [ ] I wanted a formal qualification
- [ ] Other reasons. Please state: Please enter an 'other' value for this selection. *** This question is required.**

hidden=false&rec

**2. Prior to enrolling for this certificate, where have you previously studied statistics? Check all that apply.** ***This question is required.**

- [ ] At school, up to Year 11 (5th form)
- [ ] At school, up to Year 12 or 13 (6th or 7th form)
- [ ] Introductory statistics course at a tertiary institution (university or polytech)
- [ ] Advanced (degree level) at a tertiary institution (university or polytech)
- [ ] Professional development course
- [ ] Other place. Please specify: Please enter an 'other' value for this selection. *** This question is required.**
- [ ] I have not previously studied statistics

**Before the certificate, when did you most recently study statistics?** ***This question is required.**

- ○ 2012
- ○ 2011
- ○ 2010

2009

2008 or earlier

## Working on the Research project

hidden=false&req

**3. How relevant was each unit to your research project?** ***This question is required.***

| 3. How relevant was each unit to your research project? *This question is required. | Not relevant | A little relevant | Moderately relevant | Very relevant |
|---|---|---|---|---|
| Ethical and legal issues in official statistics (23271) | | | | |
| Interpreting descriptive official statistics (23268) | | | | |
| Making inferences with official statistics (23270) | | | | |
| Making recommendations with official statistics (23269) | | | | |

hidden=false&req

0

**4. Please describe at least one topic, covered in any of the units above, that was relevant to your research project.**

hidden=false&req

0

**5. Please describe at least one topic, covered in any of the units above, that was not relevant to your research project.**

hidden=false&req

0

**6. When did you decide on a research topic?** ***This question is required.**

- ( ) Haven't started my research project
- ( ) Before the workshops began
- (•) After the first or second workshop
- ( ) After the third or fourth workshop

**How did you choose a research project topic?**

**7. When did you start working on the research project?** ***This question is required.**

- ( ) Haven't started my research project
- ( ) Before the workshops began
- ( ) After the first or second workshop
- ( ) After the third or fourth workshop

## Research project - intention

hidden=false&req

0

**8. Have you changed roles at your employer, or left your employer, since starting the course?**

- ( ) Yes
- ( ) No

hidden=false&req

0

**9. To what extent do you agree with the following statements?**

| 9. To what extent do you agree with the following statements? | Strongly agree | Agree | Disagree | Strongly disagree |
|---|---|---|---|---|
| I don't have enough time day to day to work on the research project. | ( ) | ( ) | ( ) | ( ) |
| Twelve months is not enough time to complete all assessments. | ( ) | ( ) | ( ) | ( ) |

| 9. To what extent do you agree with the following statements? | Strongly agree | Agree | Disagree | Strongly disagree |
| --- | --- | --- | --- | --- |
| The research project is a low priority for me. *This question is required | | | | |
| The research project is a low priority for my manager. *This question is required | | | | |
| The content for the research project is too difficult. *This question is required | | | | |
| I couldn't pick a topic for the research project. *This question is required | | | | |
| No suitable data were available for my research project. *This question is required | | | | |
| I didn't get enough support from my manager. *This question is required | | | | |
| I didn't get enough support with the statistical content. *This question is required | | | | |

hidden=false&req

0

**10. What other issues, aside those in the previous question, made it difficult for you to work on the research project?**

**11. How could these issues be resolved to help you to complete the project?**

## Being Informed About the Research Project

hidden=false&req

0

**12. How well informed do you think you were about the research project?**

Not at all informed

A little informed

Moderately Informed

Very informed

hidden=false&req

0

**13. How well informed do you think your manager was about the research project?**

Not at all informed

A little informed

Moderately Informed

Very informed

hidden=false&req

0

**14. What unanswered questions do you have about the research project?**

hidden=false&req

0

**15. Did you intend to complete the research project?**

Yes

No

hidden=false&req

0

**16. Would you still like to complete the research project?**

Yes

No

## Discussions

hidden=false&req

0

**17. How much did you discuss the research project with your manager?**

Not at all

A little amount

A moderate amount

A lot

hidden=false&req

**18. In any of your discussions, did you talk about any of the following? Yes No**

**Choosing a topic**

**Getting access to or collecting data**

**Choosing statistical methods**

**18. In any of your discussions, did you talk about any of the following? Yes No**

**Creating statistical outputs ready for reporting**

**Writing a first draft**

**Submitting the project**

hidden=false&req

0

**19. As a whole, how productive do you think your discussions were?**

Not productive at all

A little productive

Moderately productive

Very productive

hidden=false&req

0

**20. Which discussions, if any, were not productive and why?**

## After the course

hidden=false&req

0

**21. What did you value about the research project?**

asdf

hidden=false&req

0

**22. How much have you applied the skills and knowledge, gained on the research project, to other aspects of you work?**

Not at all

A little amount

- A moderate amout
- A lot

hidden=false&rec

0

**How did you apply this skill and knowledge to your work?**

**Are there any other comments you would like to make about the research project?**

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 827–848, http://dx.doi.org/10.1515/JOS-2016-0043

# Statistical Capacity Building of Official Statisticians in Practice: Case of the Consumer Price Index

*Tomi Deutsch*[1]

This article focuses on the issue of statistical capacity building of official statisticians using the case of the consumer price index (CPI) as an illustrative example. Although used for indexation of salaries, pensions, and social welfare benefits, but also as an approximation of the general inflation rate, there are several unresolved methodological issues associated with CPI's calculation. Apart from the choice among two alternative concepts, the challenge of how to include owner-occupied housing (OOH) in CPI has also not been adequately resolved yet. Analysis in the article is based on Slovenian data. The results show that accuracy of the CPI significantly improves if it is calculated using one of the superlative and symmetric formulas, and that it makes sense to include OOH in CPI using the total acquisitions approach. The analysis further indicates that the choice of the index formula for calculating CPI has a much greater impact on the CPI value than inclusion of OOH. Academic research findings such as these should not remain unknown to the wide professional community of official statisticians. Formal channels for knowledge transfer from academia to official statistics providers should be established to facilitate continuous statistical capacity building of official statisticians.

*Key words:* Statistical capacity building; consumer price index; superlative index formulas; owner-occupied housing; total acquisitions approach.

## 1. Introduction

The World Bank (2016) defines statistical capacity as "the ability of countries to meet user needs for good quality statistics". Usually, the statistics this definition refers to are those we tend to label 'official': produced by national and international statistics providers to help local, regional, and national governments, as well as unions of countries, make informed policy decisions and monitor their impact.

Following from the definition of statistical capacity, statistical capacity building can be defined as all efforts and activities towards raising the ability of countries to meet user needs for good quality statistics, or, put more simply, to establish a statistical system capable of providing users with timely and reliable official statistics.

The statistical capacity building activities of the United Nations, the World Bank, International Statistical Institute and its Committee on Statistical Capacity Building, and many other organizations, are usually focused on training of official statisticians in developing countries. However, even in countries with well-established statistical systems, statistical capacity building should be an ongoing process. Dilemmas of official

[1] University of Ljubljana, Faculty of Economics, Kardeljeva ploščad 17, 1000 Ljubljana, Slovenia. Email: tomi.deutsch@ef.uni-lj.si

statisticians in developed countries might be different, but they are by no means less numerous than dilemmas faced by their colleagues from developing countries.

This article focuses on the issue of statistical capacity building of official statisticians in practice, using the case of the consumer price index (CPI) as an illustrative example. Although utilized for indexation of salaries, pensions, and social welfare benefits, but also as an approximation of the general inflation rate, there are presently several unresolved methodological issues associated with the calculation of this indicator. Apart from the choice among two alternative concepts, the challenge of how to include owner-occupied housing (OOH) in CPI has also not been adequately resolved yet. Two important CPI-related methodological questions are therefore addressed in this article:

1. Do any of the existing index formulas help to significantly improve the accuracy of CPI?
2. Which among the available approaches to inclusion of OOH in CPI should be used in practice?

Following the introduction, we first provide an overview of formulas available for CPI calculations. We then discuss the pros and cons of existing approaches for inclusion of OOH in CPI. After establishing the theoretical framework we put theory to practice using Slovenia, a small open economy with a register-based system of official statistics, as a showcase. A discussion of possibilities to systematically disseminate academic research findings in the field of official statistics among the wide professional community of official statisticians concludes the article.

## 2. Formulas Available for CPI Calculations: An Overview

When deciding how to approach the CPI calculations, official statisticians face the choice between two alternative concepts: the fixed basket concept and the cost of living concept (ILO 2004). In practice, both concepts usually yield a different final CPI value, although convergences are possible (and feasible).

Availability of adequate weights (or quantities of goods) makes use of formulas for weighted price indices in CPI calculations possible. The choice among several alternatives depicted in Table 1 largely depends on the adopted concept and data availability required by the specific index formula.

According to the fixed basket concept, the CPI is calculated using a formula in which weights remain fixed and do not change between the compared periods. There are several options for choosing weights. By selecting the weights from one of the two compared periods, that is, the base or the current period, the Laspeyres ($P_L$), or the Paasche price index ($P_P$) is obtained.

The Laspeyres price index is calculated using the weights from the base period of prices, independent of the structure of population consumption in the current period. Assuming that changes in consumption are largely caused by replacement of relatively expensive goods with relatively cheaper ones, this index yields values which overestimate actual price movements between two periods.

The Paasche price index is calculated by taking into account the structure of consumption in the current period. Assuming that changes in consumption are largely

*Table 1. Alternative index formulas used in CPI calculations.*

| Description | Weights reference period | Index formula |
|---|---|---|
| Fixed basket index (Target index) | $t = 1$ | $P_P = \frac{\sum_{i=1}^n p_i^t q_i^t}{\sum_{i=1}^n p_i^0 q_i^t} = \left\{ \sum_{i=1}^n \left( \frac{p_i^t}{p_i^0} \right)^{-1} s_i^t \right\}^{-1}; \; s_i^t = \frac{p_i^t q_i^t}{\sum_{i=1}^n p_i^t q_i^t}$ |
| | $t = 0$ | $P_L = \frac{\sum_{i=1}^n p_i^t q_i^0}{\sum_{i=1}^n p_i^0 q_i^0} = \sum_{i=1}^n \frac{p_i^t}{p_i^0} s_i^0; \; s_i^0 = \frac{p_i^0 q_i^0}{\sum_{i=1}^n p_i^0 q_i^0}$ |
| Cost of living index (Target index) | $t = 1, t = 0$ | $P_F = \sqrt{P_L P_P}$ |
| Fixed basket index (Approximation) | $t \leq -1$ | $P_{Lo} = \frac{\sum_{i=1}^n p_i^t q_i^b}{\sum_{i=1}^n p_i^0 q_i^b} = \sum_{i=1}^n \frac{p_i^t}{p_i^0} s_i^{0b}; \; s_i^{0b} = \frac{p_i^0 q_i^b}{\sum_{i=1}^n p_i^0 q_i^b}$ <br> $P_{Yo} = \sum_{i=1}^n \frac{p_i^t}{p_i^0} s_i^b; \; s_i^b = \frac{p_i^b q_i^b}{\sum_{i=1}^n p_i^b q_i^b}$ |
| Cost of living index (Approximation) | $t = 0$ | $P_{GL} = \prod_{i=1}^n \left( \frac{p_i^t}{p_i^0} \right)^{s_i^0}; \; s_i^0 = \frac{p_i^0 q_i^0}{\sum_{i=1}^n p_i^0 q_i^0}$ |
| | $t \leq -1$ | $P_{GYo} = \prod_{i=1}^n \left( \frac{p_i^t}{p_i^0} \right)^{s_i^b}; \; s_i^b = \frac{p_i^b q_i^b}{\sum_{i=1}^n p_i^b q_i^b}$ |

Source: ILO, 2004.

caused by replacement of relatively expensive goods with relatively cheaper ones, this index yields values that underestimate the actual movement of prices between two periods.

From the theoretical viewpoint, it is impossible to determine which index formula, the Laspeyres or the Paasche one, is more suitable for CPI calculations. While both seem equally probable, they generally give different answers to the same question (ILO 2004, par. 15.17). For practical reasons, the Laspeyres index formula has been more frequently used in practice, because the weights from the base period are relatively easily obtainable in comparison to the weights from the current period.

Neither the Laspeyres nor the Paasche price index is suitable for measurement of inflation faced by households. An established inflation rate should not have a probable alternative value and should account for the structure of the household consumption in both the base and the current period. Diewert (1999, 31–32) reports that Walsh therefore (already in 1921) suggested to combine both indices.

Combining the quantities from both periods with the arithmetic mean formula results in the so-called Marshall-Edgeworth index. This index is rather problematic in international comparisons and therefore rarely used.

Combining the quantities from both periods with the geometric mean formula yields the Walsh price index. This index, together with the Törnqvist-Theil (or equivalent Törnqvist) and Fisher price index ($P_F$), forms a group of symmetric and superlative formulas which under certain assumptions (constant utility function over time, homothetic preferences – see Breuer 2007, 3) represent a good approximation of the real cost of living index (Diewert 1976).

The weights in the Walsh index are calculated as a combination of weights from the base and the current period and are fixed. Consequently, this index reflects price changes only. This characterizes it as an unbiased pure price index (Hill 1999, 10).

For the normal time series of data, Fisher, Törnqvist-Theil, and Walsh indices are only approximately consistent in aggregation (ILO 2004, par. 6.17), but yield similar values. Consequently, it does not matter which of these three index formulas are chosen as the target index for the CPI (ILO 2004, par. 17.5). With the Walsh index formula declared suitable for the fixed basket concept, and given the negligible differences between superlative formulas, we can assume that the other two remaining superlative formulas, that is, Fisher and Törnqvist-Theil, are also suitable for CPI calculations according to the fixed basket concept. This has been proven by Diewert (2002, 590). He also showed that among these three, the Fisher formula yields the best results both in terms of the economic and test approach (Diewert, 1999, pp. 7 and 18).

When calculating CPI in real time according to the concept of the fixed basket, it makes sense to use one of the alternative formulas which result in an approximation of the Laspeyres index. The most appropriate and most commonly used in practice are the Young index ($P_{Yo}$) and the Lowe index with hybrid weights ($P_{Lo}$). The Young index uses available weights from the period prior to the base period. If we multiply weights used for the calculation of the Young index with the base period prices, we obtain the Lowe index with hybrid weights (but have to take into account issues related to the elasticity of substitution – see UN 2009, par. 4.83).

Given that superlative indices also cover substitution between goods, geometric formulas assuming this substitution can be used. Such CPI approximations are calculated using the geometric forms of the Young ($P_{GYo}$) and/or Laspeyres ($P_{GL}$) index. These require the same data as their arithmetic equivalents (Armknecht and Silver 2012, 4; ILO 2004, par. 1.40). However, due to the delayed availability of weights, only the geometric Young index can be calculated without any major problems in real time.

Practical limitations to the use of these two indices include avoidance of prices that are equal to zero. Attention should be paid to the significant drops in prices. Elasticity of substitution is an issue as well: Armknecht and Silver (2012, 4) note that the geometric Young index is biased downwards when inelastic substitution is present. Similar bias can also be expected when the geometric Laspeyres index is used.

## 3. Approaches for Inclusion of the Owner-Occupied Housing in Consumer Price Index: An Overview

Five possible approaches for inclusion of OOH in CPI are reviewed next:

- the user cost approach,
- the rental equivalence approach,
- the payments approach,
- the net acquisitions approach,
- the total acquisitions approach.

The review helps us round the theoretical framework for the subsequent empirical showcase.

### *3.1. The User Cost Approach*

This approach builds on the cost of living concept which means that it considers the cost incurred by the use of durable goods instead of their prices (Diewert 2003, 3; UN 2009, par. 9.1.9). Consequently, according to ILO (ILO 2004, par. 10.8) we attempt to measure changes in the cost to owner-occupiers for using the dwelling. ILO (ILO 2004, par. 10.10) also provides a typical formula for the estimation of user costs: $UC = rM + iE + D + RC - K$.

The equation for the calculation of user costs (UC) can be interpreted as the difference between the cost of the acquisition of housing, later use of the housing and selling of the housing at the end of the period (Spiteri 2008, 52): $M$ and $E$ represent mortgage debt and equity in the home, and $r$ and $i$ represent mortgage interest rates and the rate of return available on alternative assets, respectively. $D$ is depreciation, $RC$ other recurring costs and capital gains $K$ (with a negative prefix).

The user costs approach has a number of theoretical and practical issues which do not make it popular among statisticians (Kurz and Hoffmann 2004, 3). Two of the most pressing ones include:

- Reduced consistency of CPI: the user cost approach should be applied to all durable goods (i.e., all goods used longer than one month). In reality, the approach is only used for OOH inclusion.
- User costs cannot be calculated only on the basis of actually observed data (irrespective of the degree of difficulty in data collection processes). Certain parameters also need to be estimated and imputed. However, simplified forms of user costs approach are used in practice, even though they do not eliminate the problems of setting the parameter values (Guðnason and Jónsdóttir 2008).

### *3.2. The Rental Equivalence Approach*

As a consequence of the user cost approach weaknesses, the rental equivalence approach was developed. Apart from its perceived simplicity in comparison with the user cost approach, its main advantage lies in its connection to both the cost of living and the fixed basket concepts.

The rental equivalence approach attempts to measure changes in the prices of housing services consumed by users of OOH by estimating the market value of these services or by estimating how much users would have to pay to rent their own housing (ILO 2004, par. 10.14). Imputed rents for OOH are equivalents of actual rents, which, besides the payment for the rent of the building or part of the building, also include the rent for the associated land. The amount of the imputed rent – as in actual rents – thus depends on the physical characteristics and the location of OOH.

Certain requirements need to be fulfilled in order to use this approach (Spiteri 2008, 53):

- there is a market for rented dwellings and a significant amount of rented dwellings with characteristics similar to those of owner-occupied dwellings,
- the rental market is competitive and not heavily regulated,
- rented and owner-occupied housing are not subject to different tax regimes,
- rental statistics do not include services like electricity and insurance, but represent pure rental cost.

Further, measured rent that was actually paid may deviate from the pure rental value due to potential inclusion of running maintenance costs of housing in the rent, equipment of rental housing and other reasons (Baldwin et al. 2006, 15).

The rental equivalence approach assumes that the rental housing services and OOH are perfect substitutes, but a number of countries have in the past experienced rapid growth in housing prices that was accompanied by moderate growth in rents (Heath 2007, 6). This creates a gap between the estimate of the costs of maintaining a certain standard of living and the actual costs faced by the households in order to provide for this standard. Consequently, it results in a biased CPI.

### *3.3. The Payments Approach*

According to the payments approach, besides the real prices of housing (and other incurred costs), interest payments for the mortgage loans are also included in the CPI. The average mortgage debt is defined as the ‘good’, and the interest rate as its price (McCarthy 2007, 87). Consequently, the CPI includes the following housing expenditure (ILO 2004, par. 10.20):

- down payments or deposits on newly purchased dwellings,
- legal and real estate agency fees payable on property transfers,
- repayments of mortgage principal,
- mortgage interest payments,
- alterations and additions to the dwelling,
- insurance of the dwelling,
- repair and maintenance of the dwelling,
- property rates and taxes.

A great advantage of the payments approach in comparison to the user costs approach is that the index includes only those categories that can be actually observed. However, it is difficult to collect all the necessary data, so in practice this approach is rather difficult to implement without giving estimates on certain values, and is therefore rarely used. The need to include a large amount of data will probably remain the biggest obstacle for the implementation of this method in the CPI calculations (UN 2009, par. 9.1.11).

### *3.4. Net Acquisitions Approach*

This approach is based on the inclusion of the actually observed prices and quantities (volume of expenditure), with both the prices and the volume of expenditure for products and services in the CPI attributable to the period in which the acquisition occurred. Consequently, the CPI covers only new constructions and housing new to the household sector.

The costs related to the OOH acquisition according to the net acquisitions approach, include (ILO 2004, par. 10.40):

- net purchases of dwellings (i.e., purchases less sales by the reference population),
- direct construction of new dwellings,
- alterations and additions to existing dwellings,

- legal and real estate agency fees payable on property transfers,
- repair and maintenance of dwellings,
- insurance of dwellings,
- property rates and taxes.

The net acquisitions approach sets limitations on the inclusion of OOH expenditure only for the purchase of housing, while the rest of the OOH expenditure according to the net approach is not affected (deviation occurs only if the associated land is treated separately). Net purchases of housing include (ILO 2004, par. 10.42):

- dwellings purchased from businesses (newly constructed dwellings, company houses, or rental dwellings),
- dwellings purchased or transferred from the government sector,
- any purchases, for owner-occupation, of rental dwellings from the reference population households.

If the CPI is prepared for a subpopulation, purchases of housing from other types of households are included in the net purchase of housing (ILO 2004, par. 10.42).

Determination of the net volume of expenditure, which according to the net acquisitions approach drops in the CPI, opens quite a few dilemmas:

- Definitions of new construction purchases and existing housing sold to the households by companies, local and national authorities as well as other organizations outside of the household sector are quite clear and unproblematic in terms of data collection. However, the mere inclusion of this range of housing sold from the standpoint of the net acquisitions approach is not satisfactory. If we want to operate with the net purchase value of the reference population, we need to deduct all sales by the households to the remaining sectors from the weights these housing purchases have. That means a certain shift in the data collection; CPI is calculated on data on purchases of goods by households, while this additionally requested information relates to the sales by households.
- Another difficulty arises because of the inclusion of the existing housing sold by the households from the reference population. This housing is excluded from the net approach, with the exception of housing that has been functioning as rental housing before the purchase. By the analogy with the logic of including former private rental housing, all housing new to the household sector, which will be rented out after the acquisition, has to be excluded from the CPI. Furthermore, according to the same logic, the value of housing, which was rented after the owner moved away, should be deducted from the value of housing new to the household sector. On the other hand, the CPI should include all housing that was rented, but became owner-occupied because the owner moved in without a transaction carried out on the market.
- A further dilemma is related to the status of purchasing housing through a real estate agent. According to the net acquisitions approach the CPI includes all purchases of existing housing, which occur through a real estate agent, which means the difference between the purchase price and the selling price, wherein this difference represents the service of a real estate agent to the buyer (UN 2009, par. 9.6.4). All taxes (VAT, sales tax, and other mandatory charges on transactions of existing

housing among the reference households) must be treated similarly, given that taxes are an integral part of the CPI (they represents the difference between the expenditures and incomes of households). For successful inclusion of housing in the CPI according to the net acquisitions approach, data related to the transactions with the existing housing is needed. Apart from the list of transactions with housing and transaction prices, we also need to acquire data on the percentage of expenditure that falls to taxes, which is included in the official records, and data on the percentage of the expenditure that goes to the real estate agent, which is usually not shown in the official records. In practice therefore, deviations from the framework of the net acquisitions approach are likely to occur. The official records on the sales of residential properties also do not show the household type (irrespective of the role of the household in the transaction), which prevents the possibility of including purchases of existing housing that were sold to the reference subpopulation by other types of households or households outside the reference subpopulation. This limitation prevents the calculation of the subsidiary CPI according to the net acquisitions approach, which includes expenditure of reference households for OOH, for more accurate indexations of various household incomes.

– Probably the largest difficulty with the implementation of the net acquisitions approach is the treatment of the land associated with the OOH. Given that land is an integral part of actual rents in the CPI, the net acquisitions approach excludes land associated with the OOH. Arguments for excluding land prices are – similar to the user cost approach – based on the investment theory instead of the theory of consumption, disregarding the fact that the land prices will affect the cost of living in the same range as the cost of buildings (Cournède 2005, 7). In the areas where the housing prices will rise due to lack of building land or increased demand on housing because of the favourable location of the area (e.g., due to the increase of available jobs), households will be faced with a rise in housing prices, which will in no way be reflected in the values of the CPI. The exclusion of land from the CPI is simple only when land for construction is bought. For the land related to OOH, the exclusion from the CPI could be carried out only by separating transaction price of the housing into the parts, which represent the value of the land and the value of the building. Such a division is neither realistic nor feasible without a certain error. Inability to separate land prices from the prices of buildings is a serious deficiency in the use of the net transactions approach (Johnson 2015, 121). Because of the difficulties with separating the land prices from the prices of buildings, alternative models of including land prices according to the net acquisitions approach are emerging. This path has also been chosen by Eurostat in the foreseen inclusion of OOH in the Harmonised Index of Consumer Prices (HICP) from 2018 onwards (Eurostat 2013b, 22).
– The net acquisitions approach is further criticized because of insufficient weighting of OOH in the CPI. The advantage that this approach has over other methods of including OOH in the CPI is thus nullified because OOH is not attributed with the weight it has in reality.
– Finally, the net acquisitions approach is also not satisfactory when measuring inflation, since inflation is not independent of price movements of the existing

> housing. The increase of the general price level also includes the increase in the prices of existing housing on the market. By increasing prices of existing housing, the ratio between money and existing housing changes; with the same amount of money less existing housing can be bought. In a situation like that, we cannot claim that inflation equals zero.

All these difficulties are extremely serious arguments against the use of the net acquisitions approach in practice.

### *3.5. Total Acquisitions Approach*

The difficulties with the net acquisitions approach can be avoided by including all acquisitions or household purchases, regardless of who the seller is, in the CPI. The so-called total acquisitions approach can be applied to the housing as most important durable good as well as to other durable goods with secondary markets, without any significant changes in the process of the CPI preparation.

Although the sale of used goods in the CPI should be excluded because ". . . a higher turnover rate (number of transactions) gives a higher total expenditure" (ILO 2004, par. 3.127), exactly this increase in the expenditure of these goods in the CPI is the biggest practical advantage of this approach. The size of weights, which are based on all housing purchases, is in fact the best possible transmitter of the importance that housing has for the reference population. The increase in the volume of transactions with housing (irrespective of the type of transaction) means an increase in the importance of housing in household consumption, which is not inconsistent with the definition of the weights in the CPI. Underrepresented housing (and by this the main disadvantage of the acquisitions approach compared to the user costs approach and to the rental equivalence approach) is thereby nullified. Moreover, having the adequate size of the weights for OOH, the weights according to the total acquisitions approach are determined on the basis of actual measurements, rather than on the basis of any imputations. According to the total acquisitions approach, the CPI includes all housing units, which in the observed period were purchased by the reference population, so this approach does not bring impracticable division of housing into those units which the index should cover, and those units which have to be eliminated. All difficulties with the separation of the prices of buildings from the prices of land are also eliminated, since the prices of land are also included in the CPI.

The inclusion of even the existing housing and the purchases of other second-hand goods is not in contrast with the international resolution on the CPI, which states that "the expenditure weights for second-hand goods should be based either on the net expenditure of the reference population on such goods, or the gross expenditure, depending on the purpose of the index" (ILO 2003, par. 31). The total acquisitions approach is thus consistent with the international resolution for the indexation purposes as well as for the measurement of inflation, yet, paradoxically, not recognized as a possible method for OOH inclusion. Nevertheless it resolves numerous methodological issues inherent to the other four approaches.

According to the total acquisitions approach, OOH can be included in the CPI in a relatively simple way. In the first step we use available data and calculate an OOH price index, which contains all purchases of housing by the reference households. This index is then assigned an adequate weight (which is the same as the volume of all housing purchases by these households) for inclusion in the total CPI. The household expenditures for OOH, which are included in the CPI in the preparation of prices and weights process are:

- market purchases of housing together with the associated land,
- purchase of land for construction and direct construction of new housing,
- adaptions and extensions of existing housing along with the repair and maintenance,
- legal costs and the costs of real estate brokerage in housing transactions (these are at least partially included in the selling price of housing),
- housing insurance, property rates, and taxes.

Given that according to the total acquisitions approach the OOH price index includes the purchases of housing on behalf of the reference population (regardless of the type of use of this housing), the purchases of housing which may become rental housing in the future are also covered. Yet eventual transition of the acquired housing to rental housing is not problematic within the total acquisitions approach, which further speaks for its practical implementation.

Certain limitations in the total acquisitions approach are represented solely by the transactions with shares of housing which generally occur among members of households and in other specific circumstances. Since these transactions typically do not reflect the real market prices of housing, their inclusion in the CPI is not justified.

## 4. Putting Theory to Practice: The Case of Slovenia

Let us take a closer look at how the total acquisitions approach could be implemented in practice using the data from the Slovenian statistical system.

### *4.1. Data Sources and Limitations*

Our analysis is based on microdata collected and edited by the Statistical Office of the Republic of Slovenia (SORS 2012a, 2012b) for the purposes of calculating the official CPI and HICP, and collected from the Real Estate Market Register regulated by the Surveying and Mapping Authority of the Republic of Slovenia (SMARS 2012). Annual data from the Household Budget Survey (HBS) and the average monthly retail prices of products and services calculated with the arithmetic mean (AM) and geometric mean (GM) formulas was also obtained from the SORS. SORS provided access to all data to the level of products and services for researching purposes. Data on carried out transactions for all types of housing was obtained from the Real Estate Market Register.

Data from the HBS was used for the preparation of annual weights. Annual average prices used in the analysis were calculated from the monthly average prices using the formula for unweighted arithmetic mean. 285 weights were obtained for the period from 2004 to 2011 and 285 average prices of products and services were obtained for the period from January 2006 to December 2011. The analysis included a smaller than usual (for CPI calculations) number of average prices of products and services in order to simplify the analysis and eliminate specific impacts on the results and because of certain limitations in the database obtained from the SORS.

Prior to the analysis, we eliminated average prices for the products and services with missing values, products and services unrecorded in the HBS (products and services that were not covered with an adequate weight), products and services with the same weight (in case of a larger number of products and services at one weight, one adequate representative for the elementary aggregate was selected for the analysis) as well as products and services where changes in data collection occurred in the period analysed. The fact that products and services without an adequate weight appear in the dataset is mainly due to the mismatch in the classification of products and services; Classification of Individual Consumption According to Purpose (COICOP) classification is not used in the average price database at the level of products and services.

SORS calculates the acquired average prices of products and services using the AM formula (input data for the CPI) and GM formula (input data for the HICP). Although this is an advantage that enables the analysis of the substitution at the elementary level, these data have several limitations. Regardless of the different methodology of calculating the average prices, their values in certain products and services do not differ. Of the 285 products and services, 16 per cent have the same value in both databases. Among the products and services with the same average price in both databases, there are a number of substitutes. Due to these data limitations, the analysis of the differences that occur because of substitutions at the elementary level is slightly curtailed.

Data obtained from the HBS is not optimal. The survey is troubled with quite a large nonresponse rate, which may somewhat distort the recorded consumption of households. In the 2012 survey of 7,002 households in the initial sample, response was only received from 3,663 households (a 46 per cent nonresponse rate). The survey and logbook together were filled out by 2,647 households (a 61 per cent nonresponse rate) (Vrabič Kek et al. 2014).

From 2009 to 2011 the response rate was slightly higher than in 2012, suggesting a decrease in willingness of households to participate in the survey. From 2009 to 2010 there were 6,288 households in the initial sample, and the survey was filled out by 3,924 (a 35 per cent nonresponse rate); the survey and the logbook together were filled out by 2,970 households (a 51 per cent nonresponse rate) (Vrabič Kek et al. 2014).

The annual data from the HBS is based on the responses of about 1,300 households. A rather low response rate in the Slovenian HBS is not an exception. Other countries are also facing similar or even lower response rates (Tršelič Selan 2006, 33) with a detectable trend towards an increasing nonresponse rate (Johnson 2015, 18). In the future we cannot realistically expect to obtain better results only by improving methods of collecting the data by HBS. It is more likely that the trend of increasing nonresponse rate and thus inadequate coverage of household consumption will continue. Sooner or later statistical offices will be forced to replace the HBS with other, especially administrative sources.

Elementary CPI was updated with the data on housing transactions from the Real Estate Market Register for the period from January 2007 to December 2011. The Real Estate Market Register is the first and only systematic source of data for systematic monitoring of the contract prices of the real estate in Slovenia (Perovšek 2009) and as such the basis for other officially published data on the real estate prices in Slovenia. Given that this is a

Table 2. Data on housing in apartment buildings.

| Year | Number of transactions | Contract price (in EUR) | Area (in square meters) | Price of square meter (in EUR) |
|---|---|---|---|---|
| 2007 | 6,557 | 82,467.2 | 56.46 | 1,521.9 |
| 2008 | 5,231 | 96,932.1 | 58.02 | 1,724.1 |
| 2009 | 4,662 | 88,615.4 | 55.35 | 1,642.1 |
| 2010 | 7,116 | 88,480.7 | 52.25 | 1,734.5 |
| 2011 | 7,242 | 85,329.9 | 51.43 | 1,696.6 |

relatively recently created register, the quality of data is not optimal and stable yet (SMARS 2008 and 2013). For example, in the period relevant for our analysis, the reporting on new constructions was not reliable; we also needed to do some updating and/or cleaning of data on sales of new and existing houses which might have impacted our results.

Data cleaning eliminated only the transactions that are not adequate for further analysis. Setting limits on the value of the transaction or the size of the housing was done arbitrarily, but these limits were set in such a way as to only eliminate the extreme values, which more often than not indicated an input error or at least excessive specifics of these transactions. For housing in apartment buildings, all transactions for which the contract value was below EUR 5,000 or above EUR 500,000 were eliminated. When considering the size of housing, all transactions with housing that have area of less than 15 or more than 200 square meters were eliminated. The same goes for all transactions with the value per square meter below EUR 100. For houses, all transactions with the contract value below 5,000 or above EUR one million were eliminated. Finally, all transactions with a price less than EUR 100 per square meter or higher than EUR 9,500 were also eliminated.

After completion of data cleaning procedures, 30,808 transactions with housing in apartment buildings and 9,131 transactions with single-family houses that relate to the period from 2007 to 2011 remained in the database. Annual data on housing in apartment buildings are shown in Table 2 and on single-family houses in Table 3.

There are significant differences in all parameters shown between the housing in apartment buildings and single-family houses. For the main part they stem from different recording mode in the Real Estate Market Register: for single-family houses the net floor area of a building is entered and for housing in apartment buildings usable area of the building. The net floor area is the sum of all rooms and the usable area is the sum of all places for residing not including the technical and common areas (SMARS 2011).

Table 3. Data on single-family houses.

| Year | Number of transactions | Contract price (in EUR) | Area (in square meters) | Price of square meter (in EUR) |
|---|---|---|---|---|
| 2007 | 2,450 | 127,276.5 | 127.01 | 1,088.1 |
| 2008 | 949 | 134,900.2 | 133.56 | 1,143.4 |
| 2009 | 1,207 | 125,271.1 | 135.43 | 1,028.2 |
| 2010 | 2,190 | 121,942.2 | 141.47 | 936.9 |
| 2011 | 2,335 | 120,024.3 | 141.40 | 938.6 |

Housing (new and used) is sold on the real estate market by different legal entities. In the period analysed there were 25,396 housing units in apartment buildings and 8,158 single-family houses sold on behalf of the households. Other legal entities in the period analysed sold 5,412 housing units in apartment buildings and 973 single-family houses. Among the housing in apartment buildings, 17.6 per cent was thus new to the household sector; for single-family houses this percentage amounts to 10.7 per cent.

Although new construction is slightly under-represented in the Real Estate Market Register, these initial difficulties in setting up the database probably did not significantly affect the share of transactions in which the seller comes from outside the household sector. A small number of these transactions are mainly the result of unsatisfactory renewal of the housing stock with new construction in Slovenia and a result of reduced sale of housing because of the onset of the economic crisis. The latter also significantly impacted the sale of existing housing in the analysed period.

### *4.2. Weights and Formulas*

In our analysis, the annual average prices of goods and services are calculated on the basis of the monthly average prices provided by SORS, using the AM formula. The average housing prices, which are inputs into the OOH price index and the updated CPI, are calculated from data on the contractual values of housing transactions, also using the AM formula. The average prices per square meter are calculated from the individual values on an annual basis.

We base our analysis on the average prices per square meter and not the average prices of housing units as a whole. This approach guarantees a higher comparability of housing from different periods as it mostly eliminates the housing size impact.

The implemented stratification (the method recognized as one of the possibilities to control for changes in the quality of housing – see Eurostat 2013a) is carried out using characteristics such as age, number of rooms, and location. Two main strata (one for housing in apartment buildings and the other for single-family houses) are created and an average price per square meter of housing calculated for both.

Annual weights for OOH price index (for housing purchases) are calculated from the volume of housing transactions, with all transactions included in the calculation of the average prices also included in the calculation of the volume. Due to the preliminary data cleaning (and due to the assumption that new housing transactions are partially underrepresented), the weights are slightly underestimated. However, based on the assessment and review of all available data, it is safe to assume that this underestimation is not large. Weights for the remaining 285 products and services in the CPI as well as for major works and renovations are calculated from the HBS data.

The OOH price index itself is prepared in a slightly simplified version (for most of the analysis only focusing on the housing purchases). Other costs of households with owner-occupied housing are not covered due to the unavailability of sufficiently detailed data although their CPI inclusion is not problematic. However, it is also not essential for demonstrating the total acquisitions impact on the CPI value. The same goes for purchases of land for construction: the omission of this data does not significantly impact the results of the analysis either because the impact of the land price is already partially included in the OOH price index via the prices of new and existing houses with associated land. Finally, the impact of land for construction on the final value of the OOH price index and the CPI is, due to the relatively small weight for this land, correspondingly small.

In the next step, annual indices using the Laspeyres, Fisher and Young index formulas, along with the formulas for the Lowe index with updated weights and the geometric versions of the Young as well as the Laspeyres index were used. Despite the fact that superlative indices are consistent with the fixed basket concept, our analysis is mainly carried out as comparison between the cost of living index (using the Fisher price index formula) and the fixed basket index (using the Laspeyres price index formula).

### 4.3. *Calculation of the Basic Consumer Price Index*

Let us now take a look at how the use of formulas allows substitution to impact the CPI values.

We analyse the values of the Laspeyres index using the average prices calculated with the arithmetic and geometric mean formulas, and values of the Fisher index using the average prices that are calculated with the geometric mean formula. Table 4 shows the values of annual indices on the previous year for the period from 2006 to 2011, and the values of the annual chained indices in 2011 (2005 = 100) in the last column of the table.

In the analysed period, the highest annual inflation was detected in 2008 and the lowest in 2010. The chained indices for the analysed period show an accumulation of a rather large difference; its minor part stems from alternative formulas used for calculation of the average prices, but most of it results from alternative weights used in different basic CPI formulas.

It follows from Table 4 that index which does not account for any kind of substitution has the highest value; it is followed by the index that takes into account the substitution at the level of goods (which is a lower level substitution) and the index which accounts both for substitution at the level of goods and substitution among goods.

The identified difference between two Laspeyres indices in Table 4 is probably slightly smaller than the actual difference would be, had we had more adequate data at our disposal (as already indicated, and contrary to expectations, some arithmetic and geometric averages for prices of goods in our two databases have the same value). Nevertheless, the results still indicate a large overestimation of inflation when the latter is calculated using formulas that do not consider substitution.

One of the theoretical arguments against calculating the CPI as a cost of living index using a superlative and symmetric index formula is that to calculate such an index, weights from both the base and the current period are needed. Therefore, it seems more appropriate

*Table 4. Values of the basic CPI according to the use of different formulas.*

| | Previous year = 100 | | | | | | 2005 = 100 |
|---|---|---|---|---|---|---|---|
| Index formula | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2011 |
| Laspeyres (AM) | 102.86 | 103.34 | 105.62 | 101.67 | 101.10 | 102.61 | 118.42 |
| Laspeyres (GM) | 102.76 | 103.24 | 105.46 | 101.56 | 100.67 | 102.51 | 117.26 |
| Fisher (GM) | 102.41 | 102.95 | 105.15 | 101.27 | 100.09 | 102.26 | 114.91 |

to calculate the CPI as a fixed-basket index, for which only the weights from the base period are needed.

In practice, the limitations in weights will occur in calculation of both types of indices, because of the time lag which makes it impossible to calculate the CPI index with weights that would coincide in time with the base period. Therefore, the CPI as a fixed basket index is usually calculated with weights derived from the period prior to the base period. To avoid major divergence between the available data and the actual household consumption, in practice updating of weights is used, these price-updated expenditure weights should represent the household expenditure in the base period better.

According to the fixed-basket concept, the Laspeyres formula is most suitable for the CPI calculation. We already know that if weights from the period before the base period are available, the Young index formula can be used. When using the updated weights for the CPI calculation, the Lowe index formula with hybrid weights or the updated Lowe index is used. These indices are shown in Table 5 (all of them are based on average product prices calculated with the AM formula).

With the exception of 2007, differences between the annual Laspeyres and Young index are negligible, that is, less than 0.1 index points. Applying weights from the year prior to the base year, that is, from 2008 to 2011, does not significantly impact the annual CPI value calculated as a fixed-basket index. Updating the weights for this period is therefore unnecessary or moreover, as the results of our analysis show, the updating procedure even slightly reduces the accuracy of the annual CPI in 2008 and 2010. On the other hand, updating has proven to be quite appropriate in 2007, when the difference between the Young and the Laspeyres index is the highest.

Current values of the CPI as a cost of living index can be obtained using the formula for the geometric Young index. Using the formula for the geometric Laspeyres index we obtain CPI values with a certain time lag, but still faster than in case of the symmetric indices. This finding applies to the use of both geometric index formulas, as they can be calculated in the same time and with the same data as their more widely used arithmetic counterparts. All indices in Table 6 are based on average product prices calculated with the GM formula.

The geometric Laspeyres index values are quite close to the Fisher index. In the annual comparisons, difference between these two indices is typically up to 0.1 index points, with the exception of 2010 when there is a difference of 0.17 index points. Divergence between the annual Fisher and the geometric Young indices is slightly higher, especially in the period until 2008, when the difference approaches the value of 0.2 index points. When using the annual chained indices, the difference between geometric Young and Fisher

*Table 5. Values of the basic CPI using outdated weights and target Laspeyres index.*

| | Previous year = 100 | | | | | 2006 = 100 |
|---|---|---|---|---|---|---|
| Index formula | 2007 | 2008 | 2009 | 2010 | 2011 | 2011 |
| Young (AM) | 103.11 | 105.57 | 101.75 | 101.16 | 102.54 | 114.89 |
| U. Lowe (AM) | 103.32 | 105.69 | 101.60 | 100.97 | 102.66 | 115.01 |
| Laspeyres (AM) | 103.34 | 105.62 | 101.67 | 101.10 | 102.61 | 115.13 |

*Table 6. Values of the basic CPI using outdated weights and target Fisher index.*

| | Previous year = 100 | | | | | | 2005 = 100 |
|---|---|---|---|---|---|---|---|
| Index formula | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2011 |
| G. Young (GM) | 102.25 | 102.76 | 104.97 | 101.31 | 100.03 | 102.14 | 114.16 |
| G. Laspeyres (GM) | 102.42 | 102.98 | 105.05 | 101.22 | 99.92 | 102.21 | 114.55 |
| Fisher (GM) | 102.41 | 102.95 | 105.15 | 101.27 | 100.09 | 102.26 | 114.91 |

indices amounts to 0.75 index points from 2005 to 2011 and the difference between the Fisher and geometric Laspeyres indices to 0.36 index points.

In comparison with the differences that occur in the final values of the CPI, the deviation is typically negligible – if it occurs – due to the use of weights from the period (year) before the base period. However, because from time to time, larger deviations can occur, all approximations to the target indices should be updated retrospectively.

By calculating the CPI according to the fixed basket concept, final values would be obtained with a one-year lag and by calculating the CPI according to the cost of living concept with a two-year lag. When using the cost of living concept, the value of the CPI could be checked in two steps – after one year by using the formula for the geometric Laspeyres index and after two years by using the formula for the Fisher index. A very similar correction is performed in the United States, where their chained CPI calculated as cost of living index is corrected twice (Bureau of Labor Statistics 2011). At the aggregate level they carry out the first and the second index calculation by using adjusted formulas for geometric mean. The third and final calculation is carried out by using the Törnqvist-Theil index formula (Cage et al. 2003).

### *4.4. Calculation of the Simplified Owner-Occupied Housing Price Index*

To adequately update the CPI with OOH, in the first step we need to calculate the OOH price index. Due to unavailability of all data we simplify and reduce the OOH price index in our analysis to the housing purchases only. Consequently, the analysed CPI is composed solely of purchases of housing in apartment buildings and purchases of single-family houses. The weights for individual type of housing are shown in Table 7. We obtain them by calculating the share of individual type of housing in the overall sum of contract prices.

Since the contract prices for single-family houses are on average higher than those for housing in apartment buildings, the weight for this type of housing is larger than their share in the total number of transactions. In 2008, when in Slovenia the share of transactions with single-family houses fell to 15 per cent of all housing transactions, the

*Table 7. Annual weights for individual type of housing in the OOH price index.*

| | Year | | | | |
|---|---|---|---|---|---|
| Type of housing | 2007 | 2008 | 2009 | 2010 | 2011 |
| Housing in apartment buildings (%) | 63.42 | 79.84 | 73.21 | 70.22 | 68.80 |
| Single-family houses (%) | 36.58 | 20.16 | 26.79 | 29.78 | 31.20 |

*Table 8. Values of simplified OOH price index (housing purchases) using various index formulas.*

| | Previous year = 100 | | | | 2007 = 100 |
|---|---|---|---|---|---|
| Index formula | 2008 | 2009 | 2010 | 2011 | 2011 |
| Laspeyres (AM) | 110.28 | 94.17 | 101.74 | 98.52 | 104.10 |
| Fisher (AM) | 110.90 | 93.96 | 101.30 | 98.53 | 104.00 |

weight for single-family houses was still 20 per cent. After the normalization of the number of transactions in 2010 and 2011, the weight for single-family houses stabilized at about 30 per cent.

Indices presented in Table 8 are based on the average product prices calculated with the arithmetic mean formula.

Comparison shows no major differences between the values of both indices. Contrary to expectations, in 2008 a slightly higher growth in housing prices was recorded by the Fisher index. This is mainly due to the household behaviour – in this year the households mostly purchased housing in apartment buildings, prices of which were consequently increasing above average. A similar situation occurs again in 2011 (in this year households to a slightly larger extent purchased single-family houses, prices of which were consequently increasing above average), but in general the values of both indices are very similar and the differences between them negligible.

### 4.5. *Calculation of the Updated Consumer Price Index*

The updated CPI contains the OOH price index with the weight, which – due to simplifications described in the previous sections – represents only that part of the actual OOH price index, which relates to housing purchases.

The weight for the Slovenian housing purchases can be determined either from the HBS data or the Real Estate Market Register data (from the sum of all contracts). Although the data in the Real Estate Market Register is, as already stated, slightly condensed due to the database cleaning and due to the certain lack of coverage with new construction in the analysed period, the volume of housing transactions from this source is larger in all years than the volume that was recorded by the HBS. With the exception of 2008 and 2009, the differences are quite large. Since the Real Estate Market Register is an administrative source, we deem it much more relevant for the preparation of weights anyway. Consequently, the weights for the OOH price index in the CPI are prepared on the basis of this source. The weights for other goods in the updated CPI are calculated from the volume of expenditure on these goods in the HBS. Table 9 contains the time series of weights for the analysed period.

Apart from both crisis years 2008 and 2009, the weight for housing purchases in the CPI amounts to about six per cent. Although slightly underestimated for reasons already explained, we do not expect the size of this weight to significantly increase by total coverage of housing purchases. The potential inclusion of transactions with shares of housing might have a slightly larger impact on the weight size, yet does not stand to debate because typically, resales within households cannot be equated with transactions in the real estate market.

*Table 9. Annual weights for OOH price index (housing purchases) in the CPI.*

| Description of expenditures | Year 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|
| Other consumer goods (HBS) (%) | 93.47 | 95.54 | 96.18 | 93.69 | 94.06 |
| Housing purchases (Real Estate Market Register) (%) | 6.53 | 4.46 | 3.82 | 6.31 | 5.94 |
| Total (%) | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

Annual weights for the OOH price index that also includes major works and renovations are shown in Table 10. For reasons of transparency, weights for major works and renovations are shown separately from the weights for the housing purchases, although both components belong in the OOH price index. Weights for major works and renovations are calculated from the volume of expenditure in the HBS. Weights that are shown do not cover other expenditures that fall within the OOH price index and the CPI (land for construction and other OOH costs) and other expenditures that are not included in the CPI.

The size of the weights for the OOH price index, which includes the housing purchases as well as major works and renovations, is only slightly larger than the weights which OOH has in the Slovenian national accounts (the size of weights for imputed rents). We can therefore conclude that the importance/relevance of OOH according to the total acquisitions approach is similar to the importance/relevance, which OOH would have, if we included the estimated costs of housing use (user costs or imputed rents) in the CPI. Implementation of the total acquisitions approach that includes transactions between households therefore does not cause inflation of weights for the OOH.

Based on the calculated weights for the OOH price index in the CPI, update of the CPI is carried out. Values of the annual updated CPI are shown in Table 11, and the CPI was updated with the OOH price index which only includes the housing purchases (with average prices per square meter of housing calculated using the arithmetic mean formula).

At the annual level, the inclusion of the housing purchases in the CPI did not lead to major changes in the value of the CPI; there are no significant differences among the values of the updated annual indices and indices calculated only for the 285 products and services remaining in the database after data cleaning (compare with the basic CPI shown in Table 4). Values of the OOH price index were largely neutralized by an adequate

*Table 10. Annual weights for OOH price index in the CPI (in percentages).*

| Description of expenditures | Year 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|
| Other consumer goods (HBS) (%) | 87.01 | 88.46 | 88.29 | 86.73 | 88.05 |
| Major works and renovations (HBS) (%) | 6.91 | 7.41 | 8.20 | 7.43 | 6.39 |
| Housing purchases (Real Estate Market Register) (%) | 6.08 | 4.13 | 3.51 | 5.84 | 5.56 |
| Total (%) | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

*Table 11. Annual updated CPI using various index formulas.*

| | Previous year = 100 | | | | 2007 = 100 |
|---|---|---|---|---|---|
| Index formula | 2008 | 2009 | 2010 | 2011 | 2011 |
| Laspeyres (AM) | 105.83 | 101.39 | 101.14 | 102.36 | 111.09 |
| Laspeyres (GM) | 105.67 | 101.28 | 100.74 | 102.28 | 110.27 |
| Fisher (GM) | 105.40 | 100.99 | 100.16 | 102.04 | 108.80 |

weighting of the index in the overall CPI. The impact of updating the CPI with the OOH, compared to the impact that formula selection on a higher level and partially formula selection on the elementary level have on the value of the CPI, is almost negligible.

On the basis of our analysis we can therefore safely conclude that the use of the total acquisitions approach increases the validity of the CPI as an indicator of inflation while at the same time does not noticeably change the overall CPI.

## 5. Conclusions

Our main theoretical and empirical findings can be summarized as follows:

1. Accuracy of the CPI significantly improves if it is calculated using one of the superlative and symmetric formulas. For the fairly widespread fixed basket indices, which do not predict adjustment of consumption to market conditions, it has been shown that they overestimate the actual inflation faced by households.
2. It makes sense to include OOH in the CPI using the total acquisitions approach. With this approach the difficulties that are present in the currently established approaches of including OOH in the CPI are eliminated. Total acquisitions approach can be implemented in the CPI without major interventions in the calculation method and the structure of the CPI, since the differences to the currently applied approach of net acquisitions are minimal. Due to the consistency of the CPI, this approach, without any major difficulties, also applies to the purchases of other used products if they have a significant enough share in the total household consumption.
3. For those housing markets which are relatively small and subject to the influence of modest building cycles (that lead to quite uneven and unsatisfactory renewal of housing stock) the total acquisitions approach is the only possible way of including OOH in the CPI. The unproblematic nature of the total acquisitions approach's implementation in this specific case also indicates the possibility of its universal applicability in the CPI calculations.
4. Our analysis further indicates that the choice of the index formula for calculating CPI has a much greater impact on the CPI value than inclusion of OOH.

These and similar academic research findings on indicators with such enormous economic and social policy implications as those inherent to the CPI should not remain unknown to the wide professional community of official statisticians. Formal channels for knowledge transfer from academia to official statistics providers should facilitate continuous statistical capacity building of official statisticians.

Joint conferences organized by the likes of International Statistical Institute and its seven associations (one of which is the International Association for Official Statistics) are one established channel, scientific journals such as Statistical Journal of the IAOS and Journal of Official Statistics another. The third channel offers itself via the recently launched European Master in Official Statistics study programme, which promises not only statistical capacity building of official statisticians via formal pedagogical activities, but also by focused joint research efforts (formal and informal ones) of participating academic institutions and official statistics providers included in the European Statistical System and beyond.

## 6. References


Armknecht, P. and M. Silver. 2012. *Post-Laspeyres: The Case for a New Formula for Compiling Consumer Price Indexes*. IMF Working Paper, WP/12/105. Available at: https://www.imf.org/external/pubs/ft/wp/2012/wp12105.pdf. (accessed December 15, 2013).

Baldwin, A., M. Prud'homme, and A. Nakamura. 2006. *An empirical analysis of the different concepts for owned accommodation in the Canadian CPI: the case of Ottawa, 1996 – 2005*. Paris: Organisation for Economic Co-operation and Development.

Breuer, C.C. 2007. *Cost-of-Living Indexes for Germany*. Ottawa: 10th Ottawa Group Meeting.

Bureau of Labor Statistics. 2011. *Consumer Price Index*. Available at: http://www.bls.gov/cpi/superlink.htm (accessed April 3, 2014).

Cage, R., J.S. Greenlees, and P. Jackman. 2003. *Introducing the Chained Consumer Price Index*. Paris: Seventh meeting of the International Working Group on Price Indices. Available at: http://www.bls.gov/cpi/super_paris.pdf. (accessed March 23, 2011).

Cournède, B. 2005. *House Prices and Inflation in the Euro Area: Economics Department Working Paper No. 450*. Paris: Organisation for Economic Co-operation and Development. Doi: http://dx.doi.org/10.1787/037842735074

Diewert, W.E. 1976. "Exact and Superlative Index Numbers." *Journal of Econometrics* 4: 115–145. Doi: http://dx.doi.org/10.1016/0304-4076(76)90009-9

Diewert, W.E. 1999. *The Consumer price index and index number purpose*. Reykjavik: Fifth meeting of the International Working Group on Price Indices. Available at: http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/home/Meeting+5/$file/1999%205th%20Meeting%20-%20Diewert%20Erwin%20-%20The%20Consumer%20Price%20Index%20and%20Index%20Number%20Purpose%202nd%20Edition.pdf. (accessed March 30, 2010).

Diewert, W.E. 2002. "Harmonized Indexes of Consumer Prices: Their Conceptual Foundations." *Zeitschrift für Volkswirtschaft und Statistik let.* 138: 547–637.

Diewert, W.E. 2003. *The Treatment of Owner Occupied Housing and Other Durables in a Consumer Price Index*. Available at: http://webstat.stat.fi/og2004/diewertpaper.pdf (accessed February 19, 2008).

Eurostat. 2013a. *Handbook on Residential Property Prices Indices (RPPIs)*. Luxembourg: Publications Office of the European Union. Doi: http://dx.doi.org/10.2785/34007

Eurostat. 2013b. *Methodological Manual Referred to in Commission Regulation (EU) No 93/2013*. Luxembourg: Eurostat. Available at: http://ec.europa.eu/eurostat/documents/272892/272983/Methodological_manual_referred_in_Reg_93_2013.pdf/922f0abf-339d-4215-9c88-16c502557b87. (accessed, July 30, 2015).

Guðnason, R. and G.R. Jónsdóttir. 2008. *House Price Indexes, Approaches and Methods*. Washington DC: The 2008 World Congress on National Accounts and Economic Performance Measures for Nations. Available at: https://hagstofa.is/media/49627/house_prices_indexes_mai_2008.pdf. (accessed January 30, 2010).

Heath, A. 2007. *CPI Measurement Issues with a Special Focus on Owner-Occupied Housing*. Basel: Bank for International Settlements. Available at: http://www.nbrm.mk/WBStorage/Files/Macedonia%20paper_Alex%20Heath0.pdf. (accessed February 19, 2008).

Hill, P. 1999. *Inflation, the Cost of Living and the Domain of a Consumer Price Index*. Joint UNECE/ILO Meeting on Consumer Price Indices. Available at: http://hdl.handle.net/11176/360226. (accessed January 14, 2010).

ILO. 2003. *Resolution Concerning Consumer Price Indices*. Geneva: International Labour Organization. Available at: http://www.ilo.org/public/english/bureau/stat/download/res/cpi2.pdf. (accessed December 12, 2012).

ILO. 2004. *Consumer Price Index Manual*. Geneva: International Labour Organization. Available at: http://www.ilo.org/public/english/bureau/stat/guides/cpi/#manual. (accessed December 12, 2012).

Johnson, P. 2015. *UK Consumer Price Statistics: A Review*. London: UK Statistics Authority. Available at: https://www.statisticsauthority.gov.uk/archive/reports—correspondence/current-reviews/uk-consumer-price-statistics—a-review.pdf. (accessed September 22, 2015).

Kurz, C. and J. Hoffmann. 2004. *A Rental-Equivalence Index for Owner-Occupied Housing in West Germany, 1985 to 1998*. Vancouver: SSHRC International Conference on Index Number Theory and the Measurement of Prices and Productivity. Available at: https://www.bundesbank.de/Redaktion/EN/Downloads/Publications/Discussion_Paper_1/2004/2004_03_18_dkp_08.pdf?__blob=publicationFile. (accessed February 19, 2008).

McCarthy, C. 2007. "Owner-Occupied Housing Costs and Bias in the Irish Consumer Price Index." *Quarterly Economic Commentary*, Autumn 2007: 83–98. Available at: https://www.esri.ie/publications/owner-occupied-housing-costs-and-bias-in-the-consumer-price-index/ (accessed February 26, 2008).

Perovšek, A. 2009. *Evidenca Trga Nepremičnin (ETN): Prispevek za Sosvet SURS*. Ljubljana: Surveying and Mapping Authority of the Republic of Slovenia.

SMARS. 2008. *Poročilo o Slovenskem Trgu Nepremičnin za Leto 2007*. Ljubljana: Surveying and Mapping Authority of the Republic of Slovenia. Available at: http://e-prostor.gov.si/fileadmin/etn/POROCILO_2007_GESLO.pdf. (accessed March 24, 2015).

SMARS. 2011. *Tehnična Navodila in Primeri za Izračun Površin Prostorov in Delov Stavb*. Ljubljana: Surveying and Mapping Authority of the Republic of Slovenia. Available at: http://www.gu.gov.si/fileadmin/gu.gov.si/pageuploads/zakonodaja/ZEN_am/Tehnicna_navodila_07_11_11.pdf. (accessed March 24, 2015).

SMARS. 2012. *Evidenca Trga Nepremičnin (Microdata)*. Ljubljana: Surveying and Mapping Authority of the Republic of Slovenia.

SMARS. 2013. *Metodološka Pojasnila k Periodičnim Poročilom (Različica 3.1)*. Ljubljana: Surveying and Mapping Authority of the Republic of Slovenia. Available at: http://www.e-prostor.gov.si/fileadmin/etn/Porocila/Metodoloska_pojasnila_k_periodicnim_porocilom_3_1.pdf. (accessed March 23, 2015).

SORS. 2012a. *Anketa o Porabi v Gopodinjstvih, 2004 – 2011 (Microdata)*. Ljubljana: Statistical Office of the Republic of Slovenia.

SORS. 2012b. *Povprečne Mesečne Cene, 2005 – 2011 (Microdata)*. Ljubljana: Statistical Office of the Republic of Slovenia.

Spiteri, A. 2008. "Owner-Occupied Housing and its Effect on the Measurement of Inflation for Malta." *Bank of Valletta Review* 38: 47–72. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.577.3110&rep=rep1&type=pdf (accessed March 30, 2010).

Tršelič Selan, A. 2006. *Metodološke Značilnosti Ankete o Porabi Gospodinjstev v Sloveniji in Evropski Uniji*. Ljubljana: Institute of Macroeconomic Analysis and Development of the Republic of Slovenia. Available at: http://www.umar.gov.si/fileadmin/user_upload/publikacije/dz/2006/dz07-06.pdf. (accessed January 11, 2008).

UN. 2009. *Practical Guide to Producing Consumer Price Indices*. New York and Geneva: United Nations. Available at: http://www.unece.org/fileadmin/DAM/stats/publications/Practical_Guide_to_Producing_CPI.pdf. (accessed May 23, 2014).

Vrabič Kek, B., M. Arnež, and P. Blažič. 2014. *Standardno Poročilo o Kakovosti za Raziskovanje Poraba v Gospodinjstvih za Leto 2012*. Ljubljana: Statistical Office of the Republic of Slovenia. Available at: http://www.stat.si/doc/metodologija/kakovost/SPK_APG_2012.pdf. (accessed May 14, 2014).

World Bank. (2016). *Statistical Capacity Building*. Available at: http://www.worldbank.org/en/data/statistical-capacity-building/overview (accessed September 30, 2016).

Received May 19, 2014
Revised October 11, 2016
Accepted October 13, 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 849–865, http://dx.doi.org/10.1515/JOS-2016-0044

# Data-Mining Opportunities for Small and Medium Enterprises with Official Statistics in the UK

*Shirley Y. Coleman*[1]

There is a growing interest in data amongst small and medium enterprises (SMEs). This article looks at ways in which SMEs can combine their internal company data with open data, such as official statistics, and thereby enhance their business opportunities. Case studies are given as illustrations of the statistical and data-mining methods involved in such integrated data analytics. The article considers the barriers that prevent more SMEs from benefitting in this field and appraises some of the initiatives that are aimed at helping to overcome them. The discussion emphasizes the importance of bringing people together from the business, IT, and statistical worlds and suggests ways for statisticians to make a greater impact.

*Key words:* Entrepreneurs; company data; open data; analytics; impact.

## 1. Introduction

There is a growing trend within the business world for utilizing big and small data of all types and there are some excellent examples of enterprising companies creating added value by data mining with official statistics. Small and medium enterprises (SMEs) are relatively slow to adopt such a data-focused approach, but competitive pressures on businesses are creating awareness and interest in this area. It is timely therefore for statisticians to consider the issues involved and to help SMEs come to terms with the new ways of thinking about data.

In spite of the vast importance of official statistics for policy making at all levels of society, there is comparatively little use of such data in the business world. Large and small companies and industries in all sectors can benefit from census and business data collected by National Statistics Institutes (NSIs). Some large companies, such as Google and Amazon, have led the way in data analytics, for example using customer data to build recommendation systems that predict what the customer wants or needs next. In quality improvement initiatives such as Six Sigma (see, for example, Snee and Hoerl 2004), large organizations are earlier adopters than SMEs, and this is the case for integrated data-analysis initiatives. A study by the European Union (EU) funded BLUE-ETS project found that "small businesses use data modestly" and "the main obstacles preventing businesses from using NSI statistics (more intensively) include. . . lack of interest" (Bavdaž 2011).

This article takes a closer look at how statisticians can help SMEs use improved data analytics. Specifically the article focusses on the benefits of integrating internal company

[1] Industrial Statistics Research Unit, School of Mathematics and Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK. Email: shirley.coleman@newcastle.ac.uk

data with publically available open data. Such integration provides added impetus to SMEs showing interest in improving their data capability to their business advantage.

The contextual characteristics of SMEs affect their willingness to embark on data-analysis initiatives. Statisticians need to understand these and work with SMEs to raise awareness of the opportunities. Case studies of successful applications of statistics with company and official data are a vital way to provide both motivation and a roadmap. Other mechanisms are needed, however, and some potential routes to reach out to SMEs are discussed in the present article.

Expanding the utilization of official statistics is timely from the official statistics point of view as governments become increasingly more careful with their money, looking for evidence of return on investment. If business usage is increased and widened, then this helps to justify the costs of compulsory collection of business and other data.

The objectives of the article are to:

– showcase examples of SMEs realizing business advantage from working with integrated data and highlight the appeal of such projects for statisticians;
– review the barriers specific to SMEs that dissuade them from utilizing integrated data analytics;
– explore and appraise some activities that aim to overcome the barriers;
– make recommendations for statisticians to help SMEs take advantage of the increased availability of data.

The article is structured as follows: Section 2 provides a brief background including the contextual characteristics of SMEs that influence the way they relate to data, the wider definition of official statistics used in the article, the meaning of terms such as data analytics, and the methods of statistical analysis and data mining that are particularly useful. This is followed by illustrative examples with increasing depth of analysis, ranging from descriptive analytics of official statistics to prescriptive analytics of integrated internal company data and official statistics used for calibration. Section 3 reflects on the issues raised by the examples and considers barriers to SMEs embarking on integrated data analytics. Various engagement activities such as networks, data hacks, and institutional initiatives are appraised in the final Section 4, and the article concludes by suggesting areas for increased involvement by statisticians, particularly in terms of business impact and positive benefit to society.

## 2. SMEs and Integrated Data Analytics in Theory and Practice

### 2.1. Background and Definitions

SMEs make up 99% of an estimated 19.3 million enterprises in the EU and provide around 65 million jobs representing two-thirds of all employment (World Bank 2015). The average European business provides employment for four people, including the owner or manager. SMEs are characterized by an innovative idea or product, usually created by the owner. There are unlikely to be many data-competent staff. Case studies from the EU-funded BYTE project show mixed feelings about using data, including worries about security and a wariness of investing resources and time in developing the necessary data-analytic skills (BYTE 2015b). SME business needs, however, are to reach appropriate markets and to make an impact by showing the importance of their product. Identifying areas where their specialist skills can add value to a free resource should be a major focus. It is therefore likely that if SMEs can see the benefits of data mining and using different sources of data, then they will be motivated to get started.

The term “official statistics” usually refers to data collected by governmental NSIs. The data are characterized by their high quality, their standardized definitions and thorough descriptions, and considerable harmonization between departments and countries. Official statistics include extensive demographic data from census and government department sources as well as data from compulsory business surveys. Although official statistics are generally available in most countries, this article refers mainly to the UK because the UK is particularly interested in promoting the use of official statistics. For example, the Royal Statistical Society (RSS) gave ten recommendations in its Data Manifesto (RSS 2014) for widening the use of data and official statistics in particular; the National Statistician, John Pullinger, is currently reviewing the Approved Researcher process by which access to government data are selectively allowed, with the aim of widening usage of official statistics and making access more straightforward. In addition, the UK Open Research Data Forum produced a Draft Concordat on Open Research Data (Research Councils UK 2015) which aims to ensure wherever possible that research data are made openly available for others to use.

Alongside official statistics, it is sensible also to consider the large quantities of data made available under the umbrella of open data. The Open Data Institute (ODI) in the UK was founded in 2012 by Sir Tim Berners-Lee and Professor Nigel Shadbolt as an independent, nonprofit, nonpartisan company limited by guarantee. ODI aims include unlocking the supply of data, including data from the private sector, and communicating its value to potential users; catalyzing the evolution of open data culture to create economic, environmental, and social value, generating demand, creating, and disseminating knowledge to address local and global issues (ODI 2015).

Open data includes data collected in projects receiving public funding, although clearly the published data may differ in depth, granularity, and coverage from the actual research results. Official statistics are particularly useful for SMEs because they are consistently and reliably available. However, both open and official statistics data are free and are a potentially useful resource for businesses to utilize. In the rest of the article we will distinguish between open data, including official statistics, and internal company data, which is generally closed and for company use only.

Data science is big business. It is a combination of data awareness, skills in data storage and manipulation, and development and application of analytical methods. Considerable insight can be obtained from descriptive analytics in which data are summarized, sliced and diced to produce tables applicable for business consumption. Descriptive analytics looks at historic data to find out what has happened. Predictive analytics goes a step further and uses analytical methods to extract patterns in the data and make predictions about the future. Predictive analytics involves data-mining techniques such as decision trees and cluster analysis, and statistical methods such as regression modeling and principal components. Real data are often very noisy, but predictive analytics makes it possible to compose rules for achieving targets and to detect relationships that are useful, for example in customer segmentation (Ahlemeyer-Stubbe and Coleman 2014). Prescriptive analytics not only predicts what will happen in the future but also considers what decisions to make as a result of the predictions. It also looks at the explanations for and implications of the predictions. The application of data-mining techniques and statistical methods is referred to as data analytics.

Internal company data include sales records with details of customer demand and location, quantities and values of transactions, time stamps, and feedback. It is a rich source of information even on its own, as exemplified in Ahlemeyer-Stubbe and Coleman (2014, chap. 10.3). In conjunction with official statistics they become even more powerful, as is shown in the examples below. A major issue is the integration of the datasets; they need to be matched, the data records may have different key identifiers, they may differ in temporal and spatial granulation and in completeness. Kenett and Shmueli (2015, chap. 10) address these issues and also note the strengths of the resulting integrated data. Applying data-mining techniques and statistical methods to integrated data is referred to as integrated data analytics.

'Big data' is an increasingly popular phrase that can apply to some SME data. Big data are characterized by having large VVV – because it may include a great *Variety* of data types, including counts, measured data, 2D and 3D spatial and image data; it may arrive with great *Velocity* and with great *Volume*. 'Great' is defined as requiring more than commonly available storage, manipulation, and analytical facilities. Big data has entered into everyday usage; for example, it was mentioned in a recent UK government annual budget statement. Such increased exposure has aroused curiosity and people, especially those working in SMEs, are becoming increasingly interested in hearing what it is all about. Hence now is a good time for statisticians to work with SMEs and encourage them to think more about analyzing their data.

EU funding recognizes the importance of data; for example, the European Data Forum project (European Data Forum 2015) arranges an annual event where anyone (organizations, charity, policy makers, and researchers) can discuss the challenges of analyzing big data and the development of emerging data economies. The special focus is on SMEs as they are the main players in the emerging data economies. The forum discusses the challenges of changing business models and issues of legality and privacy in big data; it helps organizations understand the risks and benefits of data analysis. Conference proceedings include a list of relevant EU-funded projects and other groups that showed an interest in the discussions (European Data Forum 2014). EU projects such as BYTE (2015a) and BIG (2015) concern data but neither of these focusses specifically on official statistics and the opportunities for SMEs.

### 2.2. *Illustrative Examples*

#### 2.2.1. Section Overview

In this section, the aim is to show a range of applications of official statistics to help businesses. The first example looks at companies that have used descriptive analytics to good effect without having to apply any specialist statistical skills. The only requirements are that the companies are aware of the availability of the official statistics, know how to access them and are motivated to use them because they are confident of their business impact. The second example looks at a company that is just starting to develop their data-analytic capabilities. It shows how official statistics can be integrated with more specific company data and make a distinct improvement in the value of their data. It shows the need for data-manipulation skills and demonstrates why SMEs may find it difficult to embark on this sort of work. The third example looks at a company that has a specific goal in mind and has access to statistical and data-mining expertise. It shows how official statistics can be used in a calibration capacity to check and corroborate data obtained in other ways.

### 2.2.2. Official Statistics as Descriptors

There are great opportunities for SMEs to use official statistics in a specific way and sell on the user-friendly information. Publically available data on house sales from the UK land registry, for example, have been transformed by www.zoopla.co.uk into a fascinating insight into house prices and valuations. Using demographic data from census relating to local areas, Zoopla presents a whole range of interesting facts about any chosen housing area. The data are fundamentally freely available and the value added by the company derives from the automatic data accumulation, accessibility and presentation designed by them. The business model is to attract visitors to the website and sell advertising.

Official statistics can be used to summarize the current situation and predict the likely trajectory of market sectors. For example, consider the announcement:

> *The UK maritime industry contributes up to £14 billion pa; expected to rise significantly. Direct employees include around 260,000 people.*

> Source: UK Government (2014).

These figures are highly valued by an SME engaged in the maritime equipment industry. In fact, this SME is so enthusiastic about data that they are currently engaged in a two-year Knowledge Transfer Partnership (KTP 2015) which aims: "to use statistical and data-mining techniques to facilitate the further development and enhancement of the company's *enginei* monitoring system".

The project includes data mining company data combined with using official statistics to develop a new product and reach out to new areas of interest to its customers. Sensors provide big data on fuel consumption, speed over ground, and geographical location, and these data are combined to detect the mode of operation of a ship automatically. An important application area is emissions control, which is becoming increasingly more highly regulated (Coleman et al. 2015b). Emissions can be calculated for specific engines by equations based on engine characteristics. However, the company needs to demonstrate that their calculations are correct. Conversion factors from international standards are used to verify the calculations. The conversion factors apply to general classes of engines, whereas the calculation relates to specific engines. Once verified, the emissions calculation can be used to give tailored responses for specific engines. In this application, the expected emissions for different modes can be calculated and used to inform operational decisions.

Many SMEs market a niche product appropriate for a subset of the population. Official statistics can clearly be used to identify and prioritize areas for business expansion. For example, an SME designed an expert system to help people who are unable to carry out one or more activities of daily living. There are a large variety of products available to help people walk or wash or whatever they need, but people do not know which product to choose. The expert system identifies suitable products that match the person's needs exactly. The company has many millions of records of searches dating back over the last ten years and is engaged in a two-year Knowledge Transfer Partnership (KTP 2015): "to exploit existing big data arising from decision making, by appropriate mining, monetization and marketing and to extend company offering from public to new markets including the private sector."

Using official statistics the company can find, for example, the percentage of people with disabilities in a particular geographical area and how the percentage is changing with time. This information gives insight into likely demand and a means to estimate their market penetration (Coleman et al. 2015a).

### 2.2.3. Segmentation Using Company Data and Business Statistics

Company data are usually private and closed for company use only. In conjunction with official statistics, the full breadth of demographic information collected in census and surveys can be used by the company to increase their understanding of the customer base. A project was carried out in a Knowledge Transfer Partnership (KTP 2015) with the aim: "to develop analytical skills applicable to gas flow data to improve the efficiency of gas distribution at minimum cost to the system."

Energy companies have big data from operations and system control, including measurements such as flows and leakage, and records relating to asset condition. In addition, they have copious data on customer demand. It is very important to be able to predict demand as this ensures minimum stress on supply systems, less storage requirements and more efficient and reliable energy delivery. Customer demand varies according to season, day of week, time of day, alternative energy options and prices, domestic or business/industrial use and type of customer. Demand forecasting based on seasonality and customer type is well established; however, less work has been carried out on the effect of customer characteristics on demand. Official statistics provides an opportunity to develop this area in greater detail.

As part of the KTP, company data on gas demand are plotted geographically to demonstrate their variability. There are distinct patterns, and interest focusses on how these patterns relate to demographic characteristics. A literature review suggests which characteristics are most likely to be important and relevant indicators are extracted from official statistics. Income and an index of deprivation in housing, for example, are then superimposed onto spatial maps of the demand data (Coleman and Yabsley 2015). Further official statistics are collected and prescriptive analytics are carried out. Official statistics tend to be correlated and techniques such as partial least-squares analysis can be used to find powerful predictive models for demand. By integrating customer data and publically available statistics, customers can be clustered into segments; for example, Fontdebaca et al. (2012) identified six segments of water users. Such segmentation helps the utility company to improve its planning and the reliability of its supply. Such analysis is very useful in providing the company with insights and motivating future interventions to give a better service to their customers.

This sort of analysis involves many steps and different skills. It is not surprising therefore if SMEs are reluctant to embark on such exercises. For example, in the KTP, customer data are logged in terms of postcode; postcodes need to be mapped to government-defined local area codes and data need to be extracted from official statistics for those codes. All the information is available from Office for National Statistics (ONS) websites, but the files tend to be very large and require preprocessing to be accessible on a personal computer. Mapping postcodes to local area codes requires data manipulation, for example using lookup tables in Excel. The modeling requires statistical software, either a package such as SPSS or R algorithms, the use of which has to be learned. There is also an overhead of dealing with messy data in formats that may need converting; for example, the customer-demand data were in pdf files that needed to be converted to spreadsheets before the figures could be extracted. However, the effort involved is well repaid by the interesting results.

### 2.2.4. Business Statistics as Calibration in Job Vacancies

Innovantage (see www.innovantage.co.uk) is an SME which searches the web for job advertisements and collates and digests them, selling the information on to recruitment agencies and others interested in labor-market trends. The SME collates nearly all online job adverts using a proprietary web search system. This results in a database consisting of approximately 1.5 million job adverts from nearly 200 job boards every month, which makes it an extremely rich source of intelligence. This example was presented as a case study at an RSS seminar entitled "Statistics: About Businesses, for Businesses" described in the discussion below (ENBES 2013).

A major barrier to providing high-quality information is that a single job advert can be posted on multiple job boards and by multiple recruitment agencies; furthermore, it can be reposted multiple times by individual job boards to drive up their apparent traffic. Stripping out all of this duplication in order to arrive at the true number of job vacancies is a significant statistical exercise. In the project, job vacancy information derived from online job advertisements was compared with official estimates of job vacancies supplied by the ONS using their Vacancy Survey and Labour Force Survey. There were encouraging similarities between them, but the comparison revealed considerable challenges in de-duplicating and classifying vacancy information derived from Internet job sites. The official statistics data were used to develop new methods of de-duplication, identify unexpected data-quality issues, and to inspire a radical way of overcoming a barrier to labor-market intelligence that was thought to be insurmountable by all participants in the industry, namely the issue of commercial relevance and timeliness. After developing the methodology for establishing a high-quality database of job vacancies, the company was able to carry out data-mining activities, including segmentation and modeling, to clarify the underlying trends in the market, identify patterns and predict shortfalls and demand, all of great business value.

In this example, official statistics are used to calibrate a dataset derived from data from other sources. Other examples of calibration include Dalla Valle (2014), in which a nonparametric Bayesian belief network is built using data from an association that collects information about companies. Information about companies from an official statistics source is then used to see what characteristics a set of firms should have in order to perform similarly to the firms described in the official statistics. In this way the association's data can be calibrated and exceptional companies can be investigated.

Company data are most powerfully used in conjunction with official statistics, as discussed above. However, company data have usually been collected for other purposes, such as immediate operational control, and whilst the quality may be good enough for a specific task it often suffers from difficult issues such as the overall quality of the data, the use of operational systems that were never designed for analysis, legacy systems that do not fit well with modern software, missing or incomplete records and so on. Dalla Valle and Kenett (2015) consider the difficulties of calibrating official statistics with such variable datasets. In addition, they propose ways of improving the information quality (InfoQ) of the official statistics used in this context. Their analyses show important benefits for the decision makers in the companies involved. The case study of Innovantage was selected as an example, however, because it shows how an SME is building its business specifically on the data it manipulates.

## 3. Barriers to Adoption of Integrated Data Analytics in SMEs

### *3.1. Lack of IT Skills*

The examples in Section 2 require a range of analytical skills and knowledge, including knowing how to access and manipulate suitable data and apply appropriate statistical and data-mining skills. Many large organizations have moved into data analytics by partnering with an IT company that can provide these capabilities. For example, the supermarket Tesco is now well known for its extensive use of customer data, but Tesco's first role as a company is as a supermarket. Realizing the potential of data mining, Tesco started working with Dunnhumby in 1995; Dunnhumby are specialists in data collection and analysis and they helped Tesco to roll out the loyalty-card program. The loyalty cards encouraged customers to shop at Tesco but also they were the means of recording the vital data that Tesco needed to explore and understand their customers, create target advertising around the items they bought, utilize clever stocking which would increase the number of sales in items and so on. Tesco eventually bought Dunnhumby and they would not have been so successful without the help of an outside organization (Mirani 2015).

Another example is Spotify. The company offers users an ad-free music streaming service on a subscription basis. Their main business is music; however, they use data-mining techniques to link songs together and they bought Echo Nest, a company involved in music analytics, to help them do this. One of the main reasons for the company having a turnover of £131.4 million in 2013, a 42% increase from the previous year, was because of their effective utilization of data analysis (Duedil 2015). In 2013, 4.5 billion hours of music were listened to; Spotify was able to offer a better experience to users, with recommendations of similar music in addition to the music they are currently listening to (Shah 2014). Spotify can also predict what songs are likely to make top of the charts; Spotify could not have achieved this success if they had worked alone.

The option of buying or partnering a bespoke IT company is not usually open to SMEs and they have to grow their own expertise. Making sense of data has always required appropriate skills and knowledge. There are particular problems associated with trying to explore large datasets, and visualization is critical – both for examining structures and identifying data problems. Errors in a large dataset may not be trivial to correct; indeed, the larger the dataset, the harder the task. One solution, in the UK at least, is to consider government-supported schemes, such as Knowledge Transfer Partnerships (KTPs). A government agency, called Innovate UK, part funds a graduate for one to three years to work on a well-defined project. The graduate is employed by a university and supervised by an academic but works full time at the company. KTPs are particularly suited to developing a new way of working as the KTP associate provides an independent, external viewpoint with a limited timescale. KTPs provide a low-risk way for a company to try a new venture such as integrated data analytics and develop new expertise in technical skills (Knowledge Transfer Partnerships 2015).

Another way to grow data-analysis skills is to nurture enthusiasm in young people, inspiring them to work in this area. An example is the Young Rewired State initiative in UK schools (http://yrs.io). This is now well established but was slow to get started. In 2009, Rewired State ran an event called "Young Rewired State", a weekend hosted by Google in their London offices, intended to introduce open government data to the coding youth of the UK (Rewired State 2015). Their website notes:

> "With great excitement and anticipation of meeting these young programmers we flung open the doors with a limited capacity of 50, due to the restrictions at Google London offices.
>
> Three young people signed up.
>
> As we called schools and scoured the internet we realized that there was a far bigger problem than young people not engaging with open government data. That was the lack of young programmers in the country, and the fact that we were still left with isolated kids, teaching themselves how to code in their bedrooms – terrifying their parents that they were up to no good. Schools, would often identify a lone individual who might be interested – but beyond that they could not help as they had long since stopped teaching programming.
>
> We then spent three months focused on finding the founding fifty, and with huge relief and even more anticipation, we brought them together. We ran a weekend, with mentors and government data experts on hand to help, and watched as they collaborated and created a blistering array of apps and websites, all using open government data.
>
> In front of our eyes a community was born, something that was so needed – never again would these young geniuses be coding alone, from now on they had their peers and mentors to be a part of their education and maturation into engaged civic programmers."

Traditional Information Technology (IT) school teaching does not offer students the chance to really excel in innovation. Coding workshops, however, are a more appropriate way of engaging with young people. Each year Young Rewired State organizes a festival of code somewhere in the UK; entries to the final competition display a very high level of complexity and innovation.

The Digital Agenda of the EU, managed by the European Commission Directorate-General for Communications Networks, Content & Technology supports the Europe Code Week. This is a grassroots initiative which aims to bring coding and digital literacy to everybody in a fun and engaging way (http://codeweek.eu/). In October 2015, millions of children, parents, teachers, entrepreneurs, and policy makers came together in events and classrooms to learn programming and related skills. During the workshop, young people are asked to help test out state-of-the-art equipment such as 360° cameras and wrist-worn sensors, as well as learning about computational thinking by using LEGO.

### 3.2. *Lack of Statistical Skills*

The examples in Subsections 2.2.3 and 2.2.4 demonstrate the benefits of statistical analysis and data-mining techniques applied to integrated data. Basic descriptive analytics, as in Example 2.2.2, can also be very useful; however, these still require a certain level of confidence in handling numbers. A major reason for the lack of use of official statistics by business is poor numeracy and statistical competence within the business community. SMEs show little awareness of business statistics, as found in the BLUE-ETS study (Bavdaž 2011). To develop statistical literacy in the UK, the RSS launched *getstats* in 2010 to improve how the practical numbers of daily life, business and policy are handled (RSS 2015). Initiatives that support the ten-year campaign focus on three areas: the media, politicians and policymakers, and education, including higher education. The RSS is active in all these areas and has reached a broad spectrum of the population. Improvements in national statistical competence are difficult to quantify, but it is clear that acceptance of the importance of statistics has improved. For example, on World Statistics Day (October 20, 2015) the RSS delivered a training session to elected Members of Parliament in the London Parliament Building focusing on the use and interpretation of statistics in public life.

The increasing availability of Massive Open Online Courses (MOOCs) may help ease the difficulties encountered by SME staff wanting to acquire new skills and develop new learning. A wide range of courses is offered worldwide and staff have more choice about when they learn; see, for example www.mooc-list.com. Nevertheless, learning is not the same as doing, and it is important for staff to be allowed to experiment with company and official statistics data to see for themselves what benefits can be achieved.

Eurostat is the Statistical Office of the European Communities and as a Directorate-General of the European Commission is responsible for gathering data from NSIs throughout the European Community. Eurostat's mission is to provide the European Union with a high-quality statistical information service and to generate statistics at European level that enable comparisons between countries and regions. Eurostat does not collect data but consolidates it and has a large, harmonized cache of data accessible via its website (http.//ec.europa.eu/eurostat/). Eurostat recently showed its willingness to engage with users by providing a (free) training event to local government officers with sessions aimed at enabling businesses to access and utilize Eurostat data. The full attendance at the event indicates growing confidence in handling data and interest in data mining both at the descriptive and the predictive levels.

The wealth of official statistics is a source of fascination to academic statisticians, but few of them refer to such data in their lectures or use them in examples (Bavdaž 2011).

Clearly official statistics can provide excellent examples for data manipulation, tabulation, graphics, and statistical analysis. So, in addition to poor levels of numeracy, SME managers have little awareness of the free resources that would considerably help their businesses, and when they recall the statistics they have learned they do not immediately think of official statistics.

### 3.3. *Lack of Interest*

The examples in Section 2 derive from companies whose interest in data has been stimulated by personal contact with statisticians, by their susceptibility to articles in their trade press, and by their entrepreneurial spirit. This motivation has enabled them to overcome the typical challenges for SMEs when faced with the prospect of making use of business statistics. Many SMEs are discouraged by the lack of technically skilled staff, and by staff who are expected to multitask with little time available for proactively developing new expertise. They also worry about confidentiality issues and are wary of jeopardizing relationships with their customers. The main issue, however, is a general lack of interest. Nevertheless, as we found in an earlier research project focusing on helping SMEs adopt Six Sigma practices, there can be some sea changing improvements in performance in all parts of the business if SMEs are motivated by expert support, example, and peer-group activities bringing people together and giving them the chance to help each other (Stewardson and Coleman 2003). Becoming more data driven makes good business sense. As found in a survey reported in the Harvard Business Review (McAfee and Brynjolfsson 2012):

> "The more companies characterized themselves as data-driven, the better they performed on objective measures of financial and operational results. In particular, companies in the top third of their industry in the use of data-driven decision making were, on average, 5% more productive and 6% more profitable than their competitors."

It is essential to advertise the success of data analytics to inspire others, and suitable market places for such exchanges are beginning to emerge. The examples in Section 2 show some opportunities for SMEs gained by using official statistics. It is hoped that more case studies will become available and will be instrumental in encouraging other applications in this area. Joint projects often emerge when disparate groups come together. One of the reasons for the lack of use of official statistics is that data providers, data users, and entrepreneurs tend to be isolated from each other and do not often meet. Yet there have been some excellent examples of mutual benefit when these different groups cooperate and it is evidently important to bring people together to discuss their ideas.

Hacks are events in which data suppliers, data owners, designers, and IT specialists come together to analyze data using their combined expertise. The rationale is that mixing diverse skills will prompt new ways of thinking and create innovative solutions. For example, a Culture Code Hack held in Newcastle, UK, included librarians presenting data from lending libraries, arts managers with data on location and demographics of theatre goers, and musicians with songs from different parts of Northumberland. The author presented a large set of retail sales data from the ENBIS challenge (ENBIS 2015). Computer scientist hackers then spent the next 24 hours finding innovative ways to illustrate and interpret the datasets that particularly appealed to them. Groups then presented their work, which included interactive maps showing where and when people borrow computer software library books, what type of people go to which theatre performances and how far they travel, and what melodies to expect in different Northumberland villages. The outputs were fascinating and informative, but the lack of statistical thinking was disappointing. Statistical multivariate analysis and predictive modeling could have added substantially to the insight derived from the nevertheless excellent visualizations and user-friendly interfaces.

## 4. Discussion, Conclusions, and Recommendations

Businesses are becoming more interested in using their data more effectively, and amongst NSIs there is a determination to make their official statistics more widely appreciated in the business world. It is timely, therefore, to focus on how integrated data analytics can become more mainstream.

Networks and forums are an important way to help members share new ideas, and three examples are briefly described. The Association of Public Data Users (APDU 2015) in the US is a network that connects users, producers, and disseminators of government statistical data. The network members share interest in the collection, distribution, safeguarding, and explanation of public data. Organizations align themselves with APDU to gain a better understanding of public data, and potentially gain access to the public data for further analysis.

A recent cooperation between the ONS and the RSS resulted in the RSS web-based portal called StatsUserNet, which is aimed at general user engagement (StatsUserNet 2015). The interactive website provides a forum for all communities interested in official statistics. For example, the business and trade statistics community promotes dialogue, shares information and maintains close liaison between the producers and users of official business and trade statistics. Membership of individual communities of StatsUserNet has risen to over 3,000 people.

Since its initiation in 2008, the European Network for Better Establishment Statistics (ENBES) has worked on advancing exchange between practitioners, methodologists, and academics on matters relating to business statistics. In parallel with – amongst others – the European Commission's MEETS program and the BLUE-ETS research project that were initiated at approximately the same time, ENBES is endeavoring to bring business statistics closer to its users, as well as helping to gain a better understanding of user needs for business statistics. A seminar entitled "Statistics: About Businesses, for Businesses" was co-organized by ENBES, the RSS Official Statistics Section (OSS), and the RSS Quality Improvement Section (QIS) and held at the RSS in London. The OSS is interested in issues around collecting high-quality data and presenting them clearly and accurately, the QIS is interested in supporting statisticians using data to improve the effectiveness and profitability of enterprises in all sectors. This was the first time that the two sections had held a joint meeting and this reflects their growing interest in stimulating more effective use of official statistics, especially by SMEs. The seminar brought together practitioners, users, and methodologists, and addressed entrepreneurs, researchers, and policy makers. The example described in Subsection 2.2.4 above was presented as one of the case studies

(ENBES 2013). Such seminars are very useful in raising the profile of official statistics, showing their range and generating ideas on how they could be better communicated and utilized. A follow-up meeting was held at the RSS conference in Sheffield (ENBES 2014).

The initiatives for improving access to official statistics, such as Eurostat training, and making more data openly available are welcome mechanisms that encourage greater use of data. If it is made a requirement to publish reports from such open data analysis, there will soon be a copious supply of case studies from which businesses can learn. Statisticians should watch out for interesting case studies that either show evidence of statistical thinking or could be improved by statistical thinking and make sure that SMEs hear about them.

In conclusion, many companies are aware that they have valuable internal data but they do not maximize their use of them. They are receptive to being part of research initiatives aimed at helping them, providing they can see the direct benefits. Case studies are an important tool for increasing awareness. As the aim is to reach a wide audience, authors may be well advised to focus on publication in trade journals and the popular press, such as the RSS journal *Significance* and website "Stats Life", as well as the website "Statistics Views" which aims to be a "one-stop shop" for the statistics community. This suggests the following recommendation:

- Statisticians need to publish case studies.

Future development of integrated data analytics requires that statisticians take part in entrepreneurial activities such as hacks. Coding sessions and events aimed at young people encourage the involvement of the next generation. These experiences help to overcome the barriers preventing fluent use of data and official statistics. Parents and teachers may not feel that they have time to spare to get involved in these activities, so it is very important for universities and businesses to offer as much support as possible, and statisticians are recommended to take an active part:

- Statisticians need to support hacks and coding events for young people.

The full attendance at Eurostat's local government training indicates interest and confidence in handling data amongst regional policy makers and planners. What is now needed is a mechanism for cascading that enthusiasm from local government to the businesses they deal with. Statisticians should ensure that their business contacts, colleagues, and students are aware of networks such as ENBES and forums such as *StatsUserNet* which link government and business, and they are recommended to contribute case studies to encourage more extensive utilization of data:

- Statisticians need to help to cascade interest from government to businesses.

Researchers engaging in projects funded by the EU can help to promote the growth of integrated data analytics by SMEs. Projects can build on current initiatives such as the European Data Forum. Statisticians should be encouraged to include an exploration of official statistics in work packages in EU projects and KTP projects, providing a definite focus on the opportunities that they can bring. Academics should be encouraged to include projects that exploit the benefits of official statistics in Master's-degree-level dissertations and doctoral training wherever possible. For example, Coleman (2014) described group project work which encourages integrated data analytics within the center for doctoral training in "Cloud computing for big data" at Newcastle University:

- Statisticians need to include integrated data analytics in EU projects, KTPs, and research training programs.

There are some excellent examples of open data integration; for example, Hans Rosling combines data on wealth and life expectancy in his video presentation "200 years that changed the world" (Gapminder 2015) and Stotesbury and Dorling (2015) integrate datasets on inequality and social outcomes, including environment, education, and health, and find some intriguing associations. However, neither of these examples focusses on the integration of open data and internal company data which is of particular promise for SMEs. Although some SMEs are making productive use of their data, carrying out data mining and making use of official statistics, it is unclear how prevalent this is. It is therefore recommended that:

- A survey of SME attitudes, activities, and needs should be carried out.

The interest in integrated data analytics shown by the companies in the examples in Section 2 was stimulated by contact with statisticians, by articles in their trade press and by their entrepreneurial spirit. Entrepreneurs are seeing the opportunities for adding value to open data and making business use of company data. Statisticians need to make a bigger impact both by active involvement in data communities and by writing wide-reaching articles. They need to keep pace with the changes taking place in their profession; they need to maintain contact with experts in other fields such as computer science, operations research, and data science to ensure that statistics continues to make a valuable contribution in the increasingly data-driven business world. Working alongside data owners, designers, hackers, and programmers, statisticians need to make sure that statistics as a subject is firmly established as an intrinsic component in helping businesses exploit the increasing availability of data.

## 5. References

Ahlemeyer-Stubbe, A. and S. Coleman. 2014. *A Practical Guide to Data Mining in Business and Industry*. London: Wiley.

APDU. 2015. The Association of Public Data Users. Available at: http://apdu.org/ (accessed August 2015).

Bavdaž, M. 2011. "Business Use of NSI Statistics Based on External Sources (NSIs, Publications, Expert Opinions)". Edited by M. Bavdaž. Available at: www.blue-ets.istat.it/fileadmin/deliverables/Deliverable3.1.pdf (accessed October 2015).

BIG. 2015. "BIG – Big Data Public Private Forum." Available at: http://www.big-project.eu/ (accessed August 2015).

BYTE. 2015a. "The Big Data Roadmap and Cross-DisciplinarY Community for Addressing SocieTal Externalities." Available at: http://byte-project.eu/ (accessed August 2015).

BYTE. 2015b. "Case Studies from Project." Available at: http://byte-project.eu/wp-content/uploads/2015/06/FINAL_BYTE-D3-2-Case-studies-report-.pdf (accessed October 2015).

Coleman, S.Y. 2014. "Business and Academic Synergy in the World of Big Data." In Proceedings of ENBIS-14 Conference of European Network of Business and Industrial Statistics, September 21–24, Linz, Austria. Abstract available at: http://www.enbis.org/activities/events/current/253_ENBIS_14_in_Linz//abstracts (accessed October 2016).

Coleman, S., S. Whitfield, J. Berry, G. Johnson, and P. Gore. 2015a. "Monetising Company Big Data." In Proceedings of European Network of Business and Industrial Statistics (ENBIS) Spring Meeting on Predictive Analytics with Big and Complex Data, 2–5 June, Barcelona, Spain. Abstract available at: http://www.enbis.org/activities/events/current/379_ENBIS_Spring_Meeting_2015//abstracts. Project details at http://info.ktponline.org.uk/action/details/partnership.aspx?id=9671 (accessed October 2016).

Coleman, S., I. Zaman, R. Norman, and K. Pazouki. 2015b. "Business Strategy Embraces Data Analytics to Meet New Challenges in the Shipping Industry." In Proceedings of ENBIS-15 Conference of European Network of Business and Industrial Statistics, 6–10 September, Prague, Czech Republic. Abstract available at: http://www.enbis.org/activities/events/current/380_ENBIS_15_in_Prague//abstracts. Project details at: http://info.ktponline.org.uk/action/details/partnership.aspx?id=9738 (accessed October 2015).

Coleman, S. and W. Yabsley. 2015. "Monetising Data in the Energy and Utility Sectors." In Proceedings of ENBIS-15 Conference of European Network of Business and Industrial Statistics, 6–10 September, Prague, Czech Republic. Abstract available at: http://www.enbis.org/activities/events/current/380_ENBIS_15_in_Prague//abstracts (accessed October 2015).

Dalla Valle, L. 2014. "Official Statistics Data Integration Using Copulas." *Quality Technology and Quantitative Management* 11: 111–131. Doi: http://dx.doi.org/10.1080/16843703.2014.11673329.

Dalla Valle, L. and R.S. Kenett. 2015. "Official Statistics Data Integration for Enhanced Information Quality." *Quality and Reliability Engineering International* (advance online publication). Doi: http://dx.doi.org/10.1002/qre.1859.

Duedil. 2015. "Spotify Limited." Available at: https://www.duedil.com/company/06436047/spotify-limited (accessed August 2015).

ENBES. 2013. Seminar announced at http://www.statisticsviews.com/details/event/4325891/Statistics-For-Businesses-About-Businesses-seminar.html (accessed October 2016).

ENBES. 2014. ENBES seminar presentations. Available at: http://www1.unece.org/stat/platform/display/ENBES/Session+on+Data+Collection+Challenges+in+Establishment+Surveys (accessed October 2016).

ENBIS. 2015. "ENBIS Challenge by JMP." Available at: http://www.enbis.org/activities/jmp_challenge/index?_ts=85163 (accessed October 2015).

European Data Forum. 2015. *European Data Forum Homepage*. Available at: http://www.data-forum.eu/ (accessed August 2015).

European Data Forum. 2014. *European Data Forum (EDF)*. Available at: http://www.data-forum.eu/sites/default/files/EDF2014-Report.pdf (accessed August 2015).

Fontdecaba, S., P. Grima, L. Marco, L. Rodero, J.A. Sanchez-Espigares, I. Sole, X. Tort-Martorell, D. Demessence, V.M. de Pablo, and J. Zubelsu. 2012. "A Methodology to Model Water Demand Based on the Identification of Homogeneous Client Segments. Application to the City of Barcelona." *Water Resources Management* 26: 499–516. Doi: http://dx.doi.org/10.1007/s11269-011-9928-5.

Gapminder. 2015. *200 Years that Changed the World*. Video. Available at: http://www.gapminder.org/videos/200-years-that-changed-the-world/ (accessed October 2015).

Kenett, R.S. and G. Shmueli. 2016. Official Statistics. Chapter 10 in *Information Quality – The Potential of Data and Analytics to Generate Knowledge*, by R.S. Kenett and G. Shmueli: Wiley.

Knowledge Transfer Partnerships. 2015. *Innovate UK Homepage*. Available at: http://ktp.innovateuk.org/ (accessed October 2015).

McAfee, A. and E. Brynjolfsson. 2012. "Big Data: The Management Revolution." *Harvard Business Review*, October 2012. Available at: https://hbr.org/2012/10/big-data-the-management-revolution (accessed October 2015).

Mirani, L. 2015. "Why an Obscure British Data-Mining Company Is Worth $3 billion." Available at: http://qz.com/323944/why-an-obscure-british-data-mining-company-is-worth-3-billion/ (accessed August 2015).

ODI. 2015. *Open Data Institute Homepage*. Available at: http://www.theodi.org (accessed October 2015).

Research Councils UK. 2015. "Draft Concordat on Open Research Data." Available at: http://www.rcuk.ac.uk/research/opendata/ (accessed October 2015).

Rewired State. 2015. *Rewired State Homepage*. Available at: http://rewiredstate.org (accessed October 2015).

RSS. 2014. Data Manifesto. Available at: http://www.rss.org.uk/Images/PDF/influencing-change/rss-data-manifesto-2014.pdf (accessed October 2016).

RSS. 2015. "Statistical Literacy." Available at: http://www.rss.org.uk/RSS/Influencing_Change/Statistical_literacy/RSS/Influencing_Change/Statistical_literacy.aspx?hkey=821bf2f4-8a09-413c-8d22-290e2209a92a (accessed October 2016).

Shah, F. 2014. "How Spotify and Shazam Predict Music's Next Big Talent." Available at: http://dataconomy.com/predicting-musics-next-big-talent-big-data-case-spotify-shazam/ (accessed August 2015).

Snee, R.D. and R.W. Hoerl. 2004. *Six Sigma Beyond the Factory Floor: Deployment Strategies for Financial Services, Health Care, and the Rest of the Real Economy*. New Jersey: Prentice Hall.

StatsUserNet. 2015. "Business and Trade Statistics." Available at: http://www.statsusernet.org.uk/communities/viewcommunities/groupdetails/?CommunityKey=36dd28ed-e10a-440e-b7fb-86650b746c43 (accessed October 2015).

Stewardson, D.J. and S.Y. Coleman. 2003. "Success and Failure in Helping SMEs: A Three-Year Observational Study." *Industry and Higher Education* 17: 125–130.

Stotesbury, N. and D. Dorling. 2015. *Understanding Income Inequality and its Implications: Why Better Statistics Are Needed*. Statistics Views, October. Available at:

http://www.statisticsviews.com/details/feature/8493411/Understanding-Income-Inequality-and-its-Implications-Why-Better-Statistics-are-N.html (accessed October 2016).
UK Government. 2014. Press release: Government launches maritime study to boost growth. https://www.gov.uk/government/news/government-launches-maritime-study-to-boost-growth (accessed October 2016).
World Bank. 2015. "SME Statistics." Available at: http://siteresources.worldbank.org/CGCSRLP/Resources/SME_statistics.pdf (accessed October 2015).

Received March 2014
Revised October 2015
Accepted March 2016

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 867–885, http://dx.doi.org/10.1515/JOS-2016-0045

# From Quality to Information Quality in Official Statistics

*Ron S. Kenett*[1] *and Galit Shmueli*[2]

The term *quality of statistical data,* developed and used in official statistics and international organizations such as the International Monetary Fund (IMF) and the Organisation for Economic Co-operation and Development (OECD), refers to the usefulness of summary statistics generated by producers of official statistics. Similarly, in the context of survey quality, official agencies such as Eurostat, National Center for Science and Engineering Statistics (NCSES), and Statistics Canada have created dimensions for evaluating the quality of a survey and its ability to report 'accurate survey data'.

The concept of Information Quality, or InfoQ provides a general framework applicable to data analysis in a broader sense than summary statistics: InfoQ is defined as "the potential of a data set to achieve a specific (scientific or practical) goal by using a given empirical analysis method." It relies on identifying and examining the relationships between four components: the analysis goal, the data, the data analysis, and the utility. The InfoQ framework relies on deconstructing the InfoQ concept into eight dimensions used for InfoQ assessment.

In this article, we compare and contrast the InfoQ framework and dimensions with those typically used by statistical agencies. We discuss how the InfoQ approach can support the use of official statistics not only by governments for policy decision making, but also by other stakeholders, such as industry, by integrating official and organizational data.

*Key words:* InfoQ; survey data; decision making; industry; government.

## 1. Introduction

Official statistics are produced by a variety of organizations including central bureaus of statistics, regulatory health care agencies, educational systems, and national banks. A common trend is the integration of official statistics and organizational data to derive insights at local and global levels.

One example is provided by the Intesa Sanpaolo Bank in Italy, which maintains an integrated database to support analytic research requests by management and various decision makers (Foresti et al. 2012). The bank uses regression models applied to internal data integrated with data from a range of official statistics providers, such as:

- Financial statements (CEBI),
- EPO patents (Thomson Scientific),
- Foreign direct investment (Reprint),
- ISO certificates (Accredia),

[1] KPA Ltd., Box 2525, Raanana 43100, Israel and University of Turin, Turin, Italy. Email: ron@kpa-group.com
[2] National Tsing Hua University, Institute of Service Science, 2, Hsinchu, 30013 Taiwan. Email: galit.shmueli@iss.nthu.edu.tw
**Acknowledgment:** The second author is partially funded by grant 104-2410-H-007-001-MY2 by Ministry of Science and Technology in Taiwan.

- Trademarks (UIBM, OIHM, USPTO, WIPO),
- Credit ratings (CEBI, Intesa Sanpaolo), and
- Corporate group charts (Intesa Sanpaolo).

In another example, data from surveys of companies in the north of Italy are combined with official data from the Italian stock exchange to calibrate the survey data (Dalla Valle 2014; Dalla Valle and Kenett 2015). The survey data is from Assolombarda, an Italian association of about 5,000 manufacturing and service companies located in the north of Italy. Assolombarda periodically collects data through questionnaires sent to the associated firms in order to gather information about the economic climate, firms' activity and production, and the number and types of employees. The FTSE-MIB is the official Italian stock market index for the Italian national stock exchange and consists of the 40 most-traded stock classes on the exchange. Dalla Valle and Kenett (2015) apply Bayesian networks to a dataset that integrates the Assolombarda surveys with information from the balance sheets of the 40 largest Italian firms on the Italian stock market. Proper data integration is key to this study.

In this article, we focus on information derived from an analysis of official statistics data with or without integration with other data sets. The objective is to provide decision makers with high-quality information. We use the Information Quality concept and framework by Kenett and Shmueli (2014; 2016) to evaluate the quality of such information to decision makers or other stakeholders.

Information Quality, or InfoQ, provides a general framework applicable to data analysis in a broader sense than summary statistics. InfoQ relies on identifying and examining the relationships between four components: the analysis goal, the data, the data analysis, and the utility. The first and last components, analysis goal and utility, require a translation between the particular domain and the world of statistics, data mining or another data-analytic field.

Official statistics "need to be used to be useful" (Forbes and Brown 2012, 91), and utility is one of the overarching concepts in official statistics. An issue that can lead to misconceptions and therefore a challenge to the translation of domain to statistics and back is that many of the terms used in official statistics have specific meanings in this context which are based on, but not identical to, their meaning in everyday usage. Forbes and Brown (2012, 91) state: "All staff producing statistics must understand that the conceptual frameworks underlying their work translate the real world into models that interpret reality and make it measurable for statistical purposes. . . . The first step in conceptual framework development is to define the issue or question(s) that statistical information is needed to inform. That is, to define the objectives for the framework, and then work through those to create its structure and definitions. An important element of conceptual thinking is understanding the relationship between the issues and questions to be informed and the definitions themselves."

In an interview-based study of 58 educators and policy makers, Hambleton (2002) found that the majority misinterpreted the official statistics proficiency reports that compare results across grades and across years. This finding was particularly distressing as policy makers rely on such reports for the appropriation of funds and other key decisions. In terms of information quality, the quality of the information provided by the reports was low. The translation from statistics to domain-specific insights was faulty.

The US Environmental Protection Agency (EPA), together with the Department of Defense and Department of Energy, launched the Quality Assurance Project Plan (see EPA 2005, 35), which presents "steps . . . to ensure that environmental data collected are of the correct type and quality required for a specific decision or use." They used the term Data Quality Objectives to describe "statements that express the project objectives (or decisions) that the data will be expected to inform or support." These statements relate to descriptive goals, such as "Determine with greater than 95% confidence that contaminated surface soil will not pose a human exposure hazard." These statements are used to guide the data-collection process. They are also used to assess the resulting data quality.

Central bureaus of statistics are now combining surveys with administrative data in dynamically updated studies that have replaced the traditional census approach, so that the proper integration of data sources is becoming a critical requirement. We suggest that evaluating InfoQ can significantly contribute to the range of examples described above. The article proceeds as follows: Section 2 reviews the InfoQ dimensions proposed by Kenett and Shmueli (2014), putting them in the context of official statistics research studies. Section 3 presents quality standards applicable to official statistics and their relationship with InfoQ dimensions, and Section 4 describes standards used in customer surveys and their relationship to InfoQ. We conclude with a discussion and suggest directions for further work in Section 5.

## 2. Information Quality (InfoQ) and Official Statistics

InfoQ is defined as "the potential of a data set to achieve a specific (scientific or practical) goal by using a given empirical analysis method" (Kenett and Shmueli 2014, 3). InfoQ is determined by the data ($X$), the data analysis ($f$) and the analysis goal ($g$), as well as by the relationships between them. Utility is measured using specific metric(s) ($U$). By examining each of these components and their relationships, we can learn about the contribution of a given study as a source of knowledge and insight. The components of InfoQ have been mapped onto eight dimensions that represent a deconstruction of the concept. Here, we present the eight InfoQ dimensions and provide some guiding questions that can be used in planning, designing, and evaluating official statistics reports.

### *2.1. Data Resolution*

Data resolution refers to the measurement scale and aggregation level of the data. The data's measurement scale should be carefully evaluated in terms of its suitability to the goal, the analysis methods used, and the required resolution of the utility $U$. Questions one could ask to figure out the strength of this dimension include:

- Is the data scale used aligned with the stated goal of the study?
- How reliable and precise are the data sources and data-collection instruments used in the study?
- Is the data analysis suitable for the data aggregation level?

A low rating on data resolution can be indicative of low trust in the usefulness of the study's findings.

An example of data resolution is provided by Google's ability to predict the prevalence of flu based on the type and extent of Internet search queries (www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2y6ASfagk). These predictions match the official figures published by the Centers for Disease Control and Prevention (CDC) quite well. The point is that Google's tracking has only a day's delay, compared to the week or more it takes for the CDC to assemble a picture based on reports from doctors' surgeries. Google is faster because it is tracking the outbreak by finding a correlation between what people search for online and whether they have flu symptoms. An application aiming to achieving an online prediction of flu prevalence might combine the weekly updated CDC official statistics with the dynamically updated Google estimates.

### 2.2. *Data Structure*

Data structure relates to the type(s) of data and data characteristics such as corrupted and missing values due to the study design or data-collection mechanism. Data types include structured numerical data in different forms (e.g., cross-sectional, time series, network data) as well as unstructured, non-numerical data (e.g., text, text with hyperlinks, audio, video, and semantic data). The InfoQ level of a certain data type depends on the goal at hand. Questions to ask to figure out the strength of this dimension include:

- Is the type of data used aligned with the stated goal of the study?
- Are data-integrity details (corrupted/missing values) described and handled appropriately?
- Are the analysis methods suitable for the data structure?

A low rating on data structure can be indicative of poor data coverage in terms of the project goals. For example, using a cross-sectional analysis method to analyze a time series warrants special attention when the goal is parameter inference, but is of less concern if the goal is forecasting future values. Another example is removing records with missing data when missingness might not be random. A paper analyzing online transactions with the objective of evaluating actual behavior versus declared behavior also needs data on declared behavior through focused queries or questionnaires. Without this, the structure of the data will not provide adequate information quality.

### 2.3. *Data Integration*

With the variety of data sources and data types available today, studies sometimes integrate data from multiple sources and/or types to create new knowledge regarding the goal at hand. Such integration can increase InfoQ, but in other cases it can reduce InfoQ, for example by creating privacy breaches (for a video of methodologies discussed during the 2011 international privacy data day, see https://www.youtube.com/watch?v=QES3-X0U1Q_Q). Questions to ask to figure out the strength of this dimension include:

- Are the data integrated from multiple sources? If so, what is the credibility of each source?
- How is the integration performed? Are there linkage issues that lead to dropping crucial information?
- Does the data integration add value in terms of the stated goal?
- Does the data integration cause any privacy or confidentiality concerns?

A low rating on data integration can be indicative of missed potential in data analysis.

A prime example of data integration is the fusion feature in Google (https://support.google.com/fusiontables/answer/2571232). In the risk in open source software (RISCOSS) FP7 project, a methodology was developed that aggregates quantitative data captured from OSS communities with qualitative expert opinion through an assessment of risk scenarios to derive risk indicators using Bayesian networks (see www.riscoss.eu). Another example, focused on analyzing semantic data for risk assessment based on the MUSING FP6 project, is presented in Kenett and Raanan (2010). Another example of data integration is the combination of structured and unstructured semantic data (Figini et al. 2010). See also Penny and Reale (2004) and Vicard and Scanu (2012).

### 2.4. *Temporal Relevance*

The process of deriving knowledge from data can be placed on a timeline that includes the periods of data collection, data analysis, and usage of results as well as the temporal gaps between these three stages. The different durations and gaps can each affect InfoQ. The data-collection duration can increase or decrease InfoQ, depending on the study goal, for example studying longitudinal effects versus a cross-sectional goal. Similarly, if the collection period includes uncontrollable transitions, this can be useful or disruptive, depending on the study goal. Questions to ask to figure out the strength of this dimension include:

- Considering the data collection, data analysis and deployment stages, are any of them time-sensitive?
- Does the time gap between data collection and analysis cause any concern?
- Is the time gap between the data collection and analysis and the intended use of the model (e.g., in terms of policy recommendations) of any concern?

A low rating on temporal relevance can be indicative of an analysis with low relevance to decision makers due to data collected in a different contextual condition. This can happen in economic studies with policy implications that are based on old data.

### 2.5. *Chronology of Data and Goal*

The choice of variables to collect, the temporal relationship between them, and their meaning in the context of the goal at hand affects InfoQ. Questions to ask to figure out the strength of this dimension include:

- If the stated goal is predictive, are all the predictor variables expected to be available at the time of prediction?
- If the stated goal is causal, do the causal variables precede the effects?
- In a causal study, are there issues of endogeneity (reverse-causation)?

A low rating on chronology of data and goal can be indicative of low relevance of a specific data analysis due to misaligned timing. A customer-satisfaction survey that was designed to be used as input to the annual budget planning cycle becomes irrelevant if its results are communicated after the annual budget is finalized (Kenett and Salini 2012).

### 2.6. Generalizability

The utility of $f(X| g)$ is dependent on the ability to generalize $f$ to the appropriate population. Two types of generalizability are statistical generalizability and scientific generalizability. Statistical generalizability refers to inferring from a sample to a target population. Scientific generalizability refers to applying a model based on a particular target population to other populations. This can mean either generalizing an estimated population pattern/model $f$ to other populations, or applying $f$ estimated from one population to predict individual observations in other populations. Determining the level of generalizability requires careful characterization of $g$. Generalizability is related to the concepts of reproducibility, repeatability, and replicability (Drummond 2009; Banks 2011; Kenett and Zacks 2014; McNutt 2014; Kenett and Shmueli 2015). Reproducibility represents insights that are replicable (but not necessarily identical), while repeatability is about achieving the same results in a repeated experiment. Replicability is used most often in genome wide association studies where a follow up experiment is conducted to identify a subset of genes as active, following a large study investigating thousands of genes. Repeatability relates to data quality and analysis quality, while reproducibility relates to InfoQ. Questions to ask to figure out the strength of this dimension include:

- Is the stated goal statistical or scientific generalizability?
- For statistical generalizability in the case of inference, does the study under review answer the question "What population does the sample represent?"
- For generalizability in the case of a stated predictive goal (predicting the values of new observations; forecasting future values), are the results generalizable to the data to be predicted?

In the context of item response studies, Georg Rasch used the term *specific objectivity* to describe that case essential to measurement in which "comparisons between individuals become independent of which particular instruments – tests or items or other stimuli – have been used. Symmetrically, it is thought to be possible to compare stimuli belonging to the same class – measuring the same thing – independent of which particular individuals, within a class considered, were instrumental for comparison" (Rasch 1977, 58). The term *general objectivity* is reserved for the case in which absolute measures (i.e., amounts) are independent of which instrument (within a class considered) is employed, and no other object is required. By "absolute" we mean the measure "is not dependent on, or without reference to, anything else; not relative" (ibid.). Similar constructs apply to economic and sociological studies.

### 2.7. Operationalization

Two types of operationalization are considered: construct operationalization and action operationalization of the analysis results. Constructs are abstractions that describe a phenomenon of theoretical interest. Measurable data are an operationalization of underlying constructs. The relationship between the underlying construct and its operationalization can vary, and its level relative to the goal is another important aspect of InfoQ. The role of construct operationalization is dependent on the goal, and especially on whether the goal is explanatory, predictive, or descriptive. In explanatory models, based on underlying causal theories, multiple operationalizations might be acceptable for representing the construct of interest. As long as the data are assumed to measure the construct, the variable is considered adequate. In contrast, in a predictive task, where the goal is to create sufficiently accurate predictions of a certain measurable variable, the choice of operationalized variable is critical. Action operationalizing results refers to three questions posed by W. Edwards Deming (1982):

- What do you want to accomplish?
- By what method will you accomplish it?
- How will you know when you have accomplished it?

Questions to ask to figure out the strength of construct operationalization include:

- Are the measured variables themselves of interest to the study goal, or is their underlying construct of interest?
- What are the justifications for the choice of variables?

Questions to ask to figure out the strength of operationalizing results include:

- Who can be affected (positively or negatively) by the research findings?
- What can he or she do about it?
- Who else?

A low rating on operationalization indicates that the study might have academic value but in fact has no practical impact.

### *2.8. Communication*

Effective communication of the analysis and its utility directly impacts InfoQ. There are plenty of examples where the miscommunication of valid results has led to problematic outcomes. For a study of how to make National Assessment of Educational Progress (NAEP) and state test score reporting scales and reports more understandable, see Hambleton (2002). Questions that a reviewer should ask to figure out the strength of this dimension include:

- Is the exposition of the goal, data, and analysis clear?
- Is the exposition level appropriate for the readership of this report?

A low rating on communication can indicate that poor communication might cover the true value of the analysis and, thereby, dump the value of the information provided by the analysis.

## 3. Quality Standards for Official Statistics

A concept of *Quality of Statistical Data* was developed and used in European official statistics and international organizations such as the International Monetary Fund (IMF) and the Organisation for Economic Co-operation and Development (OECD). This concept refers to the usefulness of summary statistics produced by national statistics agencies and other producers of official statistics. In this context, quality is evaluated in terms of the statistics' usefulness for a particular goal. The OECD uses seven dimensions for quality assessment: *relevance, accuracy, timeliness and punctuality, accessibility, interpretability,* *coherence,* and *credibility* (see chap. 5 in Giovanini 2008). Eurostat's quality dimensions are *relevance of statistical concept*, *accuracy of estimates, timeliness and punctuality in disseminating results, accessibility and clarity of the information, comparability, coherence,* and *completeness*. (Eurostat 2003, 2009). See also Statistics Canada (2002), UK Department of Health (2004) and Office for National Statistics (2007).

In the United States, the National Center for Science and Engineering Statistics (NCSES), formerly the Division of Science Resources Statistics, was established within the National Science Foundation with general responsibility for statistical data. Part of its mandate is to provide information that is useful to practitioners, researchers, policy-makers, and the public. NCSES prepares about 30 reports a year based on surveys.

The purpose of survey standards is to set a framework for assuring data and reporting quality. Guidance documents are meant to help

(1) increase the reliability and validity of data,
(2) promote a common understanding of the desired methodology and processes,
(3) avoid duplication and promote the efficient transfer of ideas, and
(4) remove ambiguities and inconsistencies.

The goal is to provide the clearest possible presentations of data and their analysis. Guidelines typically focus on technical issues involved in the work rather than issues of contract management or publication formats (Biemer et al. 2003, 2012).

Specifically, NCSES aims to adhere to the ideals set forth in "Principles and Practices for a Federal Statistical Agency" (Citro and Straf 2006). As NCSES is a US federal statistical agency, NCSES surveys must follow guidelines and policies as set forth in the Paperwork Reduction Act and other legislation related to surveys. For example, NCSES surveys must follow the implementation guidance, survey clearance policies, response-rate requirements, and related orders prepared by the Office of Management and Budget (OMB). The following standards are based on US government standards for statistical surveys (see www.nsf.gov/statistics/). We list them in Table 1 with an annotation mapping them onto InfoQ dimensions, where relevant.

## I. Development of Concepts, Methods, and Design

### Survey Planning

Standard 1.1: Agencies initiating a new survey or major revision of an existing survey must develop a written plan that sets forth a justification, including: goals and objectives; potential users; the decisions the survey is designed to inform; key survey estimates; the precision required of the estimates (e.g., the size of differences that need to be detected); the tabulations and analytic results that will inform decisions and other uses; related and previous surveys; steps taken to prevent unnecessary duplication with other sources of information; when and how frequently users need the data; and the level of detail needed in tabulations, confidential microdata, and public-use data files.

This standard requires the explicit declaration of goals and methods for communicating results. It also raises the issue of data resolution in terms of dissemination and generalization (estimate precision).

*Table 1. Relationship between NCSES standards and InfoQ dimensions.*

| NCSES standard | | Data resolution | Data structure | Data integration | Temporal relevance | Data-Goal chronology | Generalizability | Operationalization | Communication |
|---|---|---|---|---|---|---|---|---|---|
| **I. Development of concepts, methods and design** | Survey planning | ■ | | | | | ■ | | ■ |
| | Survey design | ■ | ■ | ■ | | | ■ | | |
| | Survey response rate | | | | | | ■ | ■ | |
| | Survey pretesting | ■ | | | | | | | |
| **II. Collection of data** | Developing sampling frames | | | | | ■ | | ■ | |
| | Required notifications to potential survey respondents | | | | | | | ■ | |
| | Data-collection methodology | | | | | | | | |
| **III. Processing and editing of data** | Data editing | | | | | | | | |
| | Nonresponse analysis and response-rate calculation | | | | | | ■ | | |
| | Coding | | | | | | | | |
| | Data protection | | | | | | | | |
| | Evaluation | | | | | | | | ■ |
| **IV. Production of estimates and projections** | Developing estimates and projections | | | | | | ■ | | |
| **V. Data analysis** | Analysis and report planning | | | | | | | | |
| | Inference and comparisons | | | | | | ■ | | |
| **VI. Review procedures** | Review of information products | | | | | | | | |
| **VII. Dissemination of information products** | Releasing information | | | | | ■ | | | ■ |
| | Data protection and disclosure avoidance for dissemination | | | | | | | | |
| | Survey documentation | | | | | | | | ■ |
| | Documentation release of public-use microdata | | | | | | | | |

## Survey Design

Standard 1.2: Agencies must develop a survey design, including defining the target population, designing the sampling plan, specifying the data-collection instrument and methods, developing a realistic timetable and cost estimate, and selecting samples using generally accepted statistical methods (e.g., probabilistic methods that can provide estimates of sampling error). Any use of nonprobability sampling methods (e.g., cut-off or model-based samples) must be justified statistically and be able to measure estimation error. The size and design of the sample must reflect the level of detail needed in tabulations and other data products, and the precision required of key estimates. Documentation of each of these activities and resulting decisions must be maintained in the project files for use in documentation (see Standards 7.3 and 7.4).

This standard advises on data resolution, data structure and data integration. The questionnaire design addresses the issue of construct operationalization, and estimation error relates to generalizability.

### Survey Response Rates

Standard 1.3: Agencies must design the survey to achieve the highest practical rates of response, commensurate with the importance of survey uses, respondent burden, and data-collection costs, to ensure that survey results are representative of the target population so that they can be used with confidence to inform decisions. Nonresponse bias analyses must be conducted when unit or item response rates or other factors suggest the potential for bias to occur.

The main focus here lies on statistical generalization. In a sense, this standard also deals with action operationalization. The survey must be designed and conducted in a way that encourages respondents to take action and respond.

### Pretesting Survey Systems

Standard 1.4: Agencies must ensure that all components of a survey function are as intended when implemented in the full-scale survey and that measurement error is controlled by conducting a pretest of the survey components or by having successfully fielded the survey components on a previous occasion.

Pretesting relates to data resolution and the question of whether the collection instrument is sufficiently reliable and precise.

## II. Collection of Data

### Developing Sampling Frames

Standard 2.1: Agencies must ensure that the frames for the planned sample survey or census are appropriate for the study design and are evaluated against the target population for quality.

Sampling-frame development is crucial for statistical generalization. Here we also ensure the chronology of data and goal in terms of the survey deployment.

### Required Notifications to Potential Survey Respondents

Standard 2.2: Agencies must ensure that each collection of information instrument clearly states the reasons the information is planned to be collected; the way such information is planned to be used to further the proper performance of the functions of the agency; whether responses to the collection of information are voluntary or mandatory (citing authority); the nature and extent of confidentiality to be provided, if any, citing authority;

an estimate of the average respondent burden together with a request that the public direct to the agency any comments concerning the accuracy of this burden estimate and any suggestions for reducing this burden; the control number; and a statement that an agency may not conduct and a person is not required to respond to an information collection request unless it displays a currently valid control number.

This is another aspect of action operationalization.

Data-Collection Methodology

Standard 2.3: Agencies must design and administer their data-collection instruments and methods in a manner that achieves the best balance between maximizing data quality and controlling measurement error while minimizing respondent burden and cost.

## III. Processing and Editing of Data

The standards in this section focus upon the data component, and in particular upon assuring data quality and confidentiality.

Data Editing

Standard 3.1: Agencies must edit data appropriately, based on available information, to mitigate or correct detectable errors.

Nonresponse Analysis and Response-Rate Calculation

Standard 3.2: Agencies must appropriately measure, adjust for, report, and analyze unit and item nonresponse to assess their effects on data quality and to inform users. Response rates must be computed using standard formulas to measure the proportion of the eligible sample that is represented by the responding units in each study, as an indicator of potential nonresponse bias.

Coding

Standard 3.3: Agencies must add codes to collected data to identify aspects of data quality from the collection (e.g., missing data) in order to allow users to appropriately analyze the data. Codes added to convert information collected as text into a form that permits immediate analysis must use standardized codes, when available, to enhance comparability.

Data Protection

Standard 3.4: Agencies must implement safeguards throughout the production process to ensure that survey data are handled to avoid disclosure.

Evaluation

Standard 3.5: Agencies must evaluate the quality of the data and make the evaluation public (through technical notes and documentation included in reports of results or through a separate report) to allow users to interpret results of analyses, and to help designers of recurring surveys focus improvement efforts.

This relates to communication.

## IV. Production of Estimates and Projections

Developing Estimates and Projections

Standard 4.1: Agencies must use accepted theory and methods when deriving both direct survey-based estimates as well as model-based estimates and projections that use survey data. Error estimates must be calculated and disseminated to support assessment of the appropriateness of the uses of the estimates or projections. Agencies must plan and implement evaluations to assess the quality of the estimates and projections.

This standard is aimed at statistical generalizability and focuses on the quality of the data analysis (deriving estimates can be considered part of the data-analysis component).

## V. Data Analysis

Analysis and Report Planning

Standard 5.1: Agencies must develop a plan for the analysis of survey data prior to the start of a specific analysis to ensure that statistical tests are used appropriately and that adequate resources are available to complete the analysis.

This standard once again focuses on analysis quality.

Inference and Comparisons

Standard 5.2: Agencies must base statements of comparisons and other statistical conclusions derived from survey data on acceptable statistical practice.

## VI. Review Procedures

Review of Information Products

Standard 6.1: Agencies are responsible for the quality of information that they disseminate and must institute appropriate content/subject-matter, statistical, and methodological review procedures to comply with OMB and agency Information Quality Guidelines.

## VII. Dissemination of Information Products

Releasing Information

Standard 7.1: Agencies must release information intended for the general public according to a dissemination plan that provides for equivalent, timely access to all users and provides information to the public about the agencies' dissemination policies and procedures, including those related to any planned or unanticipated data revisions.

This standard touches on the chronology of data and goal as well as on communication.

Data Protection and Disclosure Avoidance for Dissemination

Standard 7.2: When releasing information products, agencies must ensure strict compliance with any confidentiality pledge to the respondents and all applicable federal legislation and regulations.

Survey Documentation

Standard 7.3: Agencies must produce survey documentation that includes those materials necessary to understand how to properly analyze data from each survey, as well as the information necessary to replicate and evaluate each survey's results (see also Standard 1.2). Survey documentation must be readily accessible to users, unless it is necessary to restrict access to protect confidentiality.

Proper documentation is essential for proper communication.

Documentation and Release of Public-Use Microdata

Standard 7.4: Agencies that release microdata to the public must include documentation clearly describing how the information is constructed and provide the metadata necessary for users to access and manipulate the data (see also Standard 1.2). Public-use microdata documentation and metadata must be readily accessible to users.

This standard aims at adequate communication of the data (not the results).

These standards provide a comprehensive framework for the various activities involved in planning and implementing official statistics surveys.

The next section focuses on customer-satisfaction surveys, such as the surveys on Service of General Interest (SGI) conducted within the European Union (EU).

## 4. Standards for Customer Surveys

Customer satisfaction, according to the ISO10004:2010 standards of the International Organization for Standardization (ISO), is the "customer's perception of the degree to which the customer's requirements have been fulfilled". It is "determined by the gap between the customer's expectations and the customer's perception of the product [or service] as delivered by the organization" (ISO/TS 10004 2010).

The ISO describes the importance of standards on its website:

> "ISO is a non-governmental organization that forms a bridge between the public and private sectors. Standards ensure desirable characteristics of products and services such as quality, environmental friendliness, safety, reliability, efficiency and interchangeability – and at an economical cost."

ISO's work program ranges from standards for traditional activities such as agriculture and construction, through mechanical engineering, manufacturing and distribution, to transport, medical devices, information and communication technologies, standards for good management practice and for services. Its primary aim is to share concepts, definitions and tools to guarantee that products and services meet expectations. When standards are absent, products may turn out to be of poor quality, might be incompatible with available equipment, or could be unreliable or even dangerous.

The goals and objectives of customer-satisfaction surveys are described clearly in ISO 10004:

> "The information obtained from monitoring and measuring customer satisfaction can help identify opportunities for improvement of the organization's strategies, products, processes and characteristics that are valued by customers, and serve the organization's objectives. Such improvements can strengthen customer confidence and result in commercial and other benefits".

In the following, we give a brief description of the ISO 10004 standard.

## ISO 10004 Guidelines for Monitoring and Measuring Customer Satisfaction

The rationale of the ISO 10004 standard – as reported in Clause 1 – is to provide "guidance in defining and implementing processes to monitor and measure customer satisfaction". It is intended for use "by organizations regardless of type, size or product provided" but it is related only "to customers external to the organization".

The ISO approach outlines three phases in the processes of measuring and monitoring customer satisfaction:

(1) Planning (Clause 6);
(2) Operation (Clause 7);
(3) Maintenance and Improvement (Clause 8).

### *Planning*

The planning phase "refers to the definition of the purposes and objectives of measuring customer satisfaction and the determination of the frequency of data gathering (regularly, on an occasional basis, dictated by business needs or specific events)". For example, an organization might be interested in investigating reasons for customer complaints after the release of a new product or causes of a loss of market share. Alternatively, it might want to compare its position relative to other organizations regularly. Moreover, "[i]nformation regarding customer satisfaction might be obtained indirectly from the organization's internal processes (e.g., customer complaints handling) or from external sources (e.g., reported in the media) or directly from customers".

In determining the frequency of data collection, this clause relates to chronology of data and goal as well as to temporal relevance. "[D]efinition of . . . customer satisfaction" concerns construct operationalization. The collection of data from different sources indirectly touches on data structure and resolution. However, the use of "or" for choice of data source indicates no intention to integrate data.

### *Operation*

The operation phase represents the core of the standard and introduces the operational steps an organization should follow in order to meet the requirements of ISO 10004. These steps are:

(a) identify the customers (current or potential) and their expectations,
(b) gather customer-satisfaction data directly from customers through a survey and/or indirectly examining existing sources of information, after having identified

*Table 2. Relationship between ISO 10004 guidelines and InfoQ dimensions.*

| ISO 10004 Phase | Data Resolution | Data Structure | Data Integration | Temporal Relevance | Data-Goal Chronology | Generalizability | Operationalization | Communication |
|---|---|---|---|---|---|---|---|---|
| **Planning** | | ■ | | ■ | ■ | | ■ | |
| **Operation** | ■ | | ■ | | | ■ | | ■ |
| **Maintenance and improvement** | | | ■ | ■ | | ■ | ■ | ■ |

the main characteristics related to customer satisfaction (product, delivery, or organizational characteristic),

(c) analyze customer-satisfaction data after having chosen the appropriate method of analysis,

(d) communicate customer-satisfaction information,

(e) monitor customer satisfaction at defined intervals to control that "the customer-satisfaction information is consistent with, or validated by, other relevant business performance indicators" (Clause 7.6.5).

Statistical issues mentioned in ISO 10004 relate to the number of customers to be surveyed (sample size), the method of sampling (Clause 7.3.3.3 and Annex C.3.1, C3.2), and the choice of the scale of measurement (Clause 7.3.3.4 and Annex C.4).

Identifying the population of interest and sample design relate to generalization. Communication is central to step (d). Step (e) refers to data integration, and the choice of measurement scale relates to data resolution.

*Maintenance and Improvement*

The maintenance and improvement phase includes periodic review, evaluation, and continual improvement of processes for monitoring and measuring customer satisfaction.

This phase aims at maintaining generalizability and temporal relevance, as well as the appropriateness of construct operationalization ("reviewing the indirect indicators of customer satisfaction"). Data integration is used to validate the information against other sources, and communication and actionable operationalization are also mentioned. Table 2 summarizes the relationship between the three phases of the ISO 10004 and the eight InfoQ dimensions.

## 5. Conclusions and Discussion

The present article begins by referring to two examples where official statistics data are combined with organizational data in order to derive information of higher quality through analysis. In the Intesa Sanpaolo Bank example, the competitiveness of an enterprise was assessed using factors such as innovation and R&D, intangibles (such as human capital, brands, quality, and environmental awareness), and foreign direct investment. Some of the challenges encountered when establishing a coherent integrated database included incomplete matching using "tax ID number" as the key, since patent, certification and trademark archives contain only the business name and address of the enterprise. As a result, an algorithm was developed for matching a business name and address to other databases containing both the same information and the tax ID number. With this approach, different business names and addresses may appear for the same enterprise (for instance, abbreviated names, acronyms with or without full stops, presence of the abbreviated legal form, etc.). The tax ID number of an enterprise may also change over the years. Handling these issues properly is key to the quality of information generated by regression analysis. These aspects are related to data resolution, data structure, data integration, temporal relevance, and chronology of data and goal, four of the InfoQ dimensions. Therefore, considering each of these InfoQ dimensions, with their associated questions, can help guide the analyst to detect and formalize the challenges to the final information quality.

In the study of Dalla Valle (2014), data from a survey about the economic climate, firms' activity and production, and the number and types of employees of 167 firms located in the provinces of Milan and Lodi in the North of Italy were combined with official statistics data from the Italian national stock exchange. An analysis combining Vines and Bayesian networks permits the proper calibration of the data, thus strengthening the quality of the information derived from the survey. The InfoQ dimensions involved in this work also include data resolution, data structure, data integration, temporal relevance and chronology of data and goal. These two examples demonstrate how concern for the quality of the information derived from an analysis of a given data set requires that attention be paid to several dimensions beyond the quality of the analysis method used. The eight InfoQ dimensions provide a general template for identifying and evaluating such challenges.

With the increased availability of data sources and ubiquity of analytic technologies, the challenge of transforming data into information and knowledge is growing in importance (Kenett 2008). Official statistics play a critical role in this context and applied research, using official statistics, needs to ensure the generation of high-information quality. In this article, we discuss the various elements that determine the quality of such information and describe several proposed approaches for achieving it. Specifically, we compare the InfoQ concept of Kenett and Shmueli (2014) with NCSES and ISO standards. InfoQ is a general approach that has been applied to a wide range of applied research applications such as education, healthcare, risk management and customer-satisfaction surveys (for more examples see Kenett and Shmueli 2016). Here we discuss examples of how official statistics data and data from internal sources are integrated to generate higher information quality. These various guidelines and initiatives focus on identifying what was learned from the data analysis and was actually done within the analysis framework. These two aspects are related to the issue of reproducibility and replicability of research. The terminology in this context is not unified (Kenett and Shmueli 2015). For example, quoting Drummond (2009):

> "Reproducibility requires changes; replicability avoids them. A critical point of reproducing an experimental result is that irrelevant things are intentionally not replicated. One might say, one should replicate the result not the experiment."

In contrast, Banks (2011) notes:

> "As a former editor of the Journal of the American Statistical Association, my own sense is that very few applied papers are perfectly reproducible. Most do not come with code or data, and even if they did, I expect a careful check would find discrepancies from the published paper. The reasons are innocent: code written by graduate students is continually tweaked and has sketchy documentation. The exact data cleaning procedures are not perfectly remembered when the final version of the paper is written, or may be muddled by miscommunication among multiple authors. And even if a conscientious researcher provided a full description of every cleaning step, every model fitting choice, and all aspects of variable selection, the resulting paper would be so long and tedious that no doubt the foolish editor would demand that it be shortened."

The expanded view of information quality, embedded in the InfoQ dimensions, is an attempt to clearly map both reproducibility and replicability components in a research study. From any report derived from data analysis we would like to understand both what was learned and how it was achieved. Our work is an initial step in such an endeavor and further tools, and methodology needs to be developed in order to support it.

## 6. References

Banks, D. 2011. "Reproducible Research: A Range of Response." *Statistics, Politics, and Policy* 2: 2151–7509. Doi: http://dx.doi.org/10.2202/2151-7509.1023.

Biemer, P. and L. Lyberg. 2003. *Introduction to Survey Quality*. Hoboken: John Wiley and Sons.

Biemer, P.P., D. Trewin, H. Bergdahl, L. Japec, and Å. Pettersson. 2012. A Tool for Managing Product Quality. Eur. Conf. Quality in Official Statistics, Athens.

Citro, C.F. and M.L. Straf. 2006. Principles and Practices for a Federal Statistical Agency: Fifth Edition, The National Washington, DC: Academies Press.

Dalla Valle, L. 2014. "Official Statistics Data Integration Using Copulas." *Quality Technology and Quantitative Management* 11: 111–131. Doi: http://dx.doi.org/10.1080/16843703.2014.11673329.

Dalla Valle, L. and R.S. Kenett. 2015. "Official Statistics Data Integration to Enhanced Information Quality." *Quality and Reliability Engineering International* 31: 1281–1300. Doi: http://dx.doi.org/10.1002/qre.1859.

Deming, W.E. 1982. *Out of the Crisis*. Cambridge: MIT Press.

Drummond, C. 2009. "*Replicability is not Reproducibility: Nor is it Good Science*." In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, June 14–18, Montreal, Canada. Available at: https://www.researchgate.net/publication/228709155_Replicability_is_Not_Reproducibility_Nor_is_It_Good_Science (accessed September 5, 2016).

EPA. 2005. Uniform Federal Policy for Quality Assurance Project Plans: Evaluating, Assessing, and Documenting Environmental Data Collection and Use Programs. Available at: www.epa.gov/fedfac/pdf/ufp_qapp_v1_0305.pdf (accessed 20 October, 2015).

Eurostat. 2003. *Standard Quality Report*. Luxembourg: Eurostat.

Eurostat. 2009. *Handbook for Quality Reports*. Luxembourg: Eurostat.

Figini, S., R.S. Kenett, and S. Salini. 2010. "Optimal Scaling for Risk Assessment: Merging of Operational and Financial Data." *Quality and Reliability Engineering International* 26: 887–897. Available at: http://onlinelibrary.wiley.com/doi/10.1002/qre.1158/full (accessed September 2016).

Forbes, S. and D. Brown. 2012. "Conceptual Thinking in National Statistics Offices." *Statistical Journal of the IAOS* 28: 89–98. Doi: http://dx.doi.org/10.3233/SJI-2012-0751.

Foresti, G., F. Guelpa, and S. Trenti. 2012. "Enterprise in a Globalised Context and Public and Private Statistical Setups." *Proceedings of the 46th Scientific Meeting of the Italian Statistical Society*. (SIS), June 20–22, Rome. Available at: http://meetings.sis-statistica.org/index.php/sm/sm2012/paper/viewFile/2393/267 (accessed 22 September, 2016)

Giovanini, E. 2008. *Understanding Economic Statistics*. Paris: OECD Publishing. Available at: https://www.oecd.org/std/41746710.pdf (accessed 22 October, 2016).

Hambleton, R.K. 2002. "How Can We Make NAEP and State Test Score Reporting Scales and Reports More Understandable?" In *Assessment in educational reform*, edited by R.W. Lissitz and W.D. Schafer. 192–205. Boston, MA: Allyn & Bacon.

ISO/TS 10004. 2010 (E). Quality Management — Customer Satisfaction — Guidelines for Monitoring and Measuring, First Edition, the International Organization for Standardization, Geneva, Switzerland.

Kenett, R.S. 2008. "From Data to Information to Knowledge." *Six Sigma Forum Magazine*: 32–33.

Kenett, R.S. and Y. Raanan. 2010. *Operational Risk Management: A Practical Approach to Intelligent Data Analysis*. Chichester, UK: John Wiley and Sons.

Kenett, R.S. and S. Salini. 2012. *Modern Analysis of Customer Satisfaction Surveys: With Applications Using R*. Chichester, UK: John Wiley and Sons.

Kenett, R.S. and G. Shmueli. 2014. "On Information Quality (with discussion)." *Journal of the Royal Statistical Society, Series A* 177: 3–38.

Kenett, R.S. and G. Shmueli. 2015. "Clarifying the Terminology that Describes Scientific Reproducibility (correspondence)." *Nature Methods* 12: 699. Doi: http://dx.doi.org/10.1038/nmeth.3489.

Kenett, R.S. and G. Shmueli. 2016. *Information Quality: The Potential of Data and Analytics to Generate Knowledge*. Chichester, UK: John Wiley and Sons.

Kenett, R.S. and S. Zacks. 2014. *Modern Industrial Statistics: With Applications in R, MINITAB and JMP*. Chichester, UK: John Wiley and Sons.

McNutt, M. 2014. "Reproducibility." *Science* 343: 229. Doi: http://dx.doi.org/10.1126/science.1250475.

Office for National Statistics 2007. *Guidelines for measuring statistical quality*. London: Office for National Statistics.

Penny, R.N. and M. Reale. 2004. "Using Graphical Modelling in Official Statistics." *Quaderni di Statistica* 6: 31–48. Available at: http://www.labstat.it/home/wp-content/uploads/2015/03/reale.pdf (accessed September 2016).

Rasch, G. 1977. "On Specific Objectivity: An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements." *The Danish Yearbook of Philosophy* 14: 58–93.

Statistics Canada 2002. *Statistics Canada's quality assurance framework*. Ottawa: Statistics Canada.

UK Department of Health 2004. *A strategy for NHS information quality assurance—consultation draft*. London: Department of Health.

Vicard, P. and M. Scanu. 2012. "Applications of Bayesian Networks in Official Statistics." In *Advanced Statistical Methods for the Analysis of Large Data-Sets*, edited by A. di Ciaccio, M. Coli, and J.M. Angulo Ibanez. 113–123. Berlin: Springer Verlag.

Received September 2014
Revised September 2015
Accepted August 2016

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 887–905, http://dx.doi.org/10.1515/JOS-2016-0046

# The Use of Official Statistics in Self-Selection Bias Modeling

*Luciana Dalla Valle*[1]

Official statistics are a fundamental source of publicly available information that periodically provides a great amount of data on all major areas of citizens' lives, such as economics, social development, education, and the environment. However, these extraordinary sources of information are often neglected, especially by business and industrial statisticians. In particular, data collected from small businesses, like small and medium-sized enterprizes (SMEs), are rarely integrated with official statistics data.

In official statistics data integration, the quality of data is essential to guarantee reliable results. Considering the analysis of surveys on SMEs, one of the most common issues related to data quality is the high proportion of nonresponses that leads to self-selection bias.

This work illustrates a flexible methodology to deal with self-selection bias, based on the generalization of Heckman's two-step method with the introduction of copulas. This approach allows us to assume different distributions for the marginals and to express various dependence structures. The methodology is illustrated through a real data application, where the parameters are estimated according to the Bayesian approach and official statistics data are incorporated into the model via informative priors.

*Key words:* Bayes theorem; copulas; Heckman's two-step method; informative priors; small and medium-sized enterprizes.

## 1. Introduction

Official statistics are a fundamental source of information about many aspects of citizens' lives, about health, education, public and private services, as well as about the economic climate, the financial situation, and the environment.

Official statistics represent precious and rich data sources not only for public institutions, but also for firms that need to compare their performance against their competitors, measure the satisfaction of their customers, explore new markets and identify the most profitable locations to establish new subsidiaries.

However, the use of official statistics by firms, and in particular by medium-sized enterprizes (SMEs), is still rather limited.

Due to the recent growth of the number of available data sources and the increase in data quality, the use of innovative methods to aggregate results obtained from

[1] School of Computing, Electronics and Mathematics, Plymouth University, Drake Circus, PL4 8AA Plymouth, Devon, UK. Email: luciana.dallavalle@plymouth.ac.uk

**Acknowledgments:** I thank the anonymous referees for their insightful comments, which significantly improved the presentation of the article.

**Disclaimer:** The contents of this publication, including analysis and statistical elaborations, are the sole responsibility of the author and can in no way to be attributed to ISTAT. The views expressed in this article, as well as the information included in it, do not necessarily reflect the opinion or position of the European Commission and in no way commit the institution.

official statistics and from specific datasets is fundamental in order to obtain reliable analyses.

The issue of data quality may invalidate statistical results, in which case integrating different data sources and methods to improve data quality is needed.

According to the literature, the reliability of the results of a survey is reduced by the existence of nonsampling errors or errors related to data-collection methods.

The major types of nonsampling errors are measurement, coverage, and self-selection errors (see Nicolini and Dalla Valle 2012).

A *coverage error* is observed when the total number of subjects (target population) and their list (frame population), available to the creator of the sampling list used to select surveyed units, do not coincide.

A *measurement error* is given by the difference between the real value of an item related to a surveying unit and the corresponding observed value. This type of error frequently has been attributed to the presence of an interviewer.

Finally, a *self-selection error*, or unit nonresponse error, takes place when the selected unit does not answer or does not fill in the questionnaire form. This nonresponse may be caused by the inability to reach the subject or by his/her refusal to join the survey. The self-selected subjects who have provided answers to the questionnaire form a nonprobabilistic sample of the population.

In this article we focus on self-selection error, which is associated with subjects' independent decision to take part in the survey.

The main issue with self-selection is that the responders differ from nonresponders and therefore estimating an effect from only the responders might confound the effect and the choice to respond. Typically responders have common characteristics (i.e., they may all be young, middle-class women). In this case the sample is biased, since it does not represent the population it is related to, and the sample distribution of the variables differs from the same variables in the population.

The literature proposes some methods to correct the bias caused by self-selection. The Propensity Score Matching method was first introduced by Rubin (1974) and later developed by Rosenbaum and Rubin (1983), and suggests correcting the self-selection bias in probabilistic terms. According to this method, propensity scores are calculated using a multivariate logistic regression, and then each responder (from the so-called treatment group) is matched with a nonresponder (from the so-called control group) with the same score (for more details, please see Nicolini and Dalla Valle 2011). However, Propensity Score Matching requires large samples with substantial overlap between treatment and control groups.

The Heckman two-step Procedure, proposed by Heckman (Heckman 1979), considers two equations tied together by a latent factor that allows the missing data associated with the nonresponding subjects to be estimated. Heckman's method and its variants have been an essential tool for applied economics. Hamilton and Nickerson (2003) apply Heckman's method to strategic management and in particular to endogenous self-selection, according to which managers choose strategies and organizational forms with the expectation that they will yield high performance. The authors show that the use of corrections for endogeneity may yield more accurate estimates of the costs and benefits of alternative strategic choices. Lucchetti and Pigini (2013) use Heckman's self-selection model to propose a test for bivariate normality in imperfectly observed models, based on the information matrix test for censored models with bootstrap critical values via Monte Carlo simulation. However, Heckman's approach requires restrictive assumptions that limit its flexibility and makes it difficult to adapt it to various dependence structures in the data.

We propose a novel approach allowing us on the one hand to correct self-selection bias and on the other hand to integrate specific data with official statistic data. This innovative approach combines the virtues of a flexible generalization of Heckman's two-step method using copulas and the Bayesian approach. The use of copulas to generalize Heckman's method relaxes the assumptions of normality and permits the accommodation of different types of dependencies, while the Bayesian approach allows the integration of official data by means of prior information. Moreover, our method can be applied successfully when dealing with small samples.

The remainder of this article is organized as follows: in Section 2 we introduce copulas and we present the main results of copula theory; Section 3 is devoted to the self-selection model as proposed by Heckman; Section 4 illustrates the characteristics of the proposed approach, using copulas within the self-selection model and integrating information with the Bayesian approach; Section 5 introduces an illustrative example and presents the results of the application of our model; finally, concluding remarks are given in Section 6.

## 2. Introduction to Copulas

### *2.1. Definition of Copula*

The copula allows us to model the joint distribution of two or more random variables in a flexible way, incorporating their dependency effects. The word copula is derived from Latin, meaning to bind, tie, connect, and was first adopted by Sklar (Sklar 1959). In this context, the term refers to the ability of the copula to link the marginal distributions of random variables to a multivariate distribution, generating a stochastic dependence relationship. The main advantage of the copula is that it allows us to explicitly express the dependence structure of a set of random variables, whatever the distribution of these variables.

More formally, the copula is a multivariate distribution function defined over the unit cube $[0, 1]^d$ (where $d$ is the dimension of the copula), $C : [0, 1]^d \rightarrow [0, 1]$, linking two or more marginals distributed as uniforms. In the bivariate case, our focus in the remainder of the article, $d = 2$ and the copula is expressed as

$$C_\theta(u_1, u_2) = \Pr(U_1 < u_1, U_2 < u_2), \tag{1}$$

where $C$ is the bivariate copula, $U_1, U_2$ are uniformly distributed random variables, with support belonging to the set $[0, 1]^2$, and $\theta$ is the copula dependence parameter vector.

The most important result in copula theory is Sklar's theorem (Sklar 1959), stating that if $F$ is a joint bivariate distribution function with marginals $F_1$ and $F_2$, then there exists a bivariate copula $C$ such that for $(x_1, x_2)$

$$F(x_1, x_2) = C_\theta(F_1(x_1), F_2(x_2)). \tag{2}$$

If $F_1$ and $F_2$ are continuous functions, then the copula is unique for any $(x_1, x_2) \in \mathbb{R} \cup \{-\infty, +\infty\}$. Thus, although the marginals are arguments of the copula, it is independent of them, since it separates the distributions of the marginals from their dependence structure, parametrized by $\theta$.

Nelsen's (1999) corollary suggests a method of generating copulas via the inversion method. If $F$ is a continuous bivariate joint distribution function with univariate marginals $F_1$ and $F_2$ and generalized inverses $F_1^{-1}$ and $F_2^{-1}$, then for $(u_1, u_2)$ there exists a unique copula $C$ such that

$$C_\theta(u_1, u_2) = F\big(F_1^{-1}(u_1), F_2^{-1}(u_2)\big). \tag{3}$$

## 2.2. *Types of Copulas*

The two main families of copulas are the Elliptical and Archimedean copulas (see Joe 1993, 1997).

Elliptical copulas are the copulas of elliptical distributions and their form is generally obtained using Nelsen's corollary (3). They are multivariate distributions sharing many of the tractable properties of the multivariate normal distribution.

The most popular elliptical copula is the Normal or Gaussian copula, whose characteristics are summarized in Table 1.

Another example of a copula that is particularly useful for its mathematical simplicity is the Farlie-Gumbel-Morgenstern (FGM) copula (Morgenstern 1956; Gumbel 1960; Farlie 1960).

The Archimedean family includes copulas expressed in a simple form based on the mathematical theory of associativity, and covers a variety of dependence structures. Archimedean copulas are constructed based on a generator function $\varphi : [0, 1] \rightarrow [0, \infty]$, with the properties of being a continuous, convex, and decreasing function (i.e., $\varphi(1) = 0$, $\varphi'(t) < 0$ and $\varphi''(t) > 0$ for $0 < t < 1$). The function $\varphi(t)$ generates the copula, in the bivariate case, as follows

$$\varphi(C_\theta(u_1, u_2)) = \varphi(u_1) + \varphi(u_2). \tag{4}$$

When the generator is strict (i.e., $\varphi(0) = \infty$), then the inverse $\varphi^{-1}(\cdot)$ exists and the copula is expressed as

$$C_\theta(u_1, u_2) = \varphi^{-1}[\varphi(u_1) + \varphi(u_2)],$$

otherwise a pseudoinverse function $\varphi^{[-1]}$ is used.

Some of the most popular Archimedean copulas are the AMH, Clayton, Gumbel and Frank copula (Ali et al. 1978; Clayton 1978; Gumbel 1960; Frank 1979). The main characteristics of these types of copulas are listed in Table 1 and they will be used in Section 5 to fit our model to real data. The range of Kendall's $\tau$ is reported for comparison purposes. This concordance measure is generally preferred to the copula's dependence parameter $\theta$, since $\tau$ is invariant with respect to the marginals and to strictly increasing transformations of the variables. For more details about transforming the copula parameter $\theta$ into Kendall's $\tau$, please see Smith (2003).

Figure 1 shows the bivariate contour plots of the different types of copulas illustrated in this section, all with standard normal margins and $\tau = 0.5$.

*Table 1. Characteristics of different copulas.*

| Copula | Distribution | Generator function $\varphi(t)$ | $\theta$ range | Kendall's $\tau$ range | Type of dependence |
|---|---|---|---|---|---|
| Gaussian | $\Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta)$ | - | $-1 \leq \theta \leq 1$; $\theta = 0$ independence | $-1 \leq \tau \leq 1$ | both negative and positive equal in upper and lower tails |
| FGM | $u_1 u_2[1 + \theta(1 - u_1)(1 - u_2)]$ | - | $-1 \leq \theta \leq 1$; $\theta = 0$ independence | $-2/9 \leq \tau \leq 2/9$ | both negative and positive equal and weak dependence in tails |
| AMH | $u_1 u_2/[1 - \theta(1 - u_1)(1 - u_2)]$ | $\ln\left(\frac{1-\theta(1-t)}{t}\right)$ | $-1 \leq \theta < 1$; $\theta = 0$ independence | $-0.1817 \leq \tau < 1/3$ | both negative and positive unequal and weak dependence in tails |
| Clayton | $\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}$ | $(1/\theta)(t^{-\theta} - 1)$ | $0 < \theta < \infty$; $\theta \rightarrow 0$ independence | $0 < \tau < 1$ | only positive strong left tail and weak right tail |
| Gumbel | $\exp\left(-\left[(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}\right]^{1/\theta}\right)$ | $(-\ln t)^{\theta}$ | $1 \leq \theta < \infty$; $\theta = 1$ independence | $0 \leq \tau < 1$ | only positive strong right tail and weak left tail |
| Frank | $-\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right)$ | $-\ln\left[\left(e^{-\theta t} - 1\right)\right]/\left(e^{-\theta} - 1\right)$ | $\{\theta \in (-\infty; 0) \cup (0; \infty)\}$; $\theta \rightarrow 0$ independence | $-1 \leq \tau \leq 1$ | both positive and negative equal in upper and lower tails |

Note that $\Phi_2(\cdot\,, \cdot\,; \theta)$ denotes the bivariate cumulative distribution function (cdf) of the standard normal with Pearson's correlation parameter $\theta$ and $\Phi^{-1}(\cdot)$ denotes the inverted cdf of the univariate standard normal.

## 3. The Self-Selection Model

The self-selection model we are proposing is also known as the *Tobit-2* model (as introduced by Tobin 1958). This is a censored regression model where the dependent variable is only observed in a selected sample that is not representative of the population. Censoring occurs when the value of the dependent variable is only partially known. It is a defect in the sample, because if there were no censoring, then the data would be a representative sample from the population of interest.

In 1979, Heckman proposed a model for self-selection, which is made by two linear equations related to each other: the substantial equation and the selection equation.

Supposing that data are missing for $N - n$ observations (the number of nonresponders), we define the selection equation (that represents participation) for individual $i$, $i = 1, \ldots, N$, as follows:

$$y_{1i}^* = \mathbf{x}_{1i}\beta_1 + \varepsilon_{1i}, \tag{5}$$

where $y_{1i}^*$ is an unobserved latent random variable such that $y_{1i}^* > 0$ corresponds to responders, while $y_{1i}^* \leq 0$ corresponds to nonresponders; $\mathbf{x}_{1i}$ is the $i$th vector of variables known for all $N$ subjects, $\beta_1$ is a vector of parameters, and $\varepsilon_{1i}$ is the error.

The substantial equation (that is observed for participants) for individual $i$ is:

$$y_{2i}^* = \mathbf{x}_{2i}\beta_2 + \varepsilon_{2i}, \tag{6}$$

where $y_{2i}^*$ denotes the latent continuous variable of interest, $\mathbf{x}_{2i}$ is the $i$th vector of variables known for all $N$ subjects, $\beta_2$ is a vector of parameters, and $\varepsilon_{2i}$ is the error.

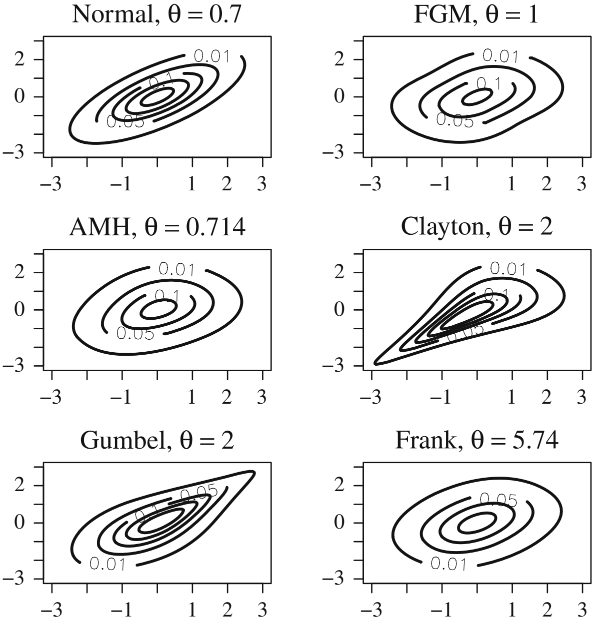


*Fig. 1. Bivariate contour plots of different copulas, with standard normal margins and $\tau = 0.5$. From the top left figure: Normal with $\theta = 0.7$, FGM with $\theta = 1$, AHM with $\theta = 0.714$, Clayton with $\theta = 2$, Gumbel with $\theta = 2$, Frank with $\theta = 5.74$.*

Note that the explanatory variables $\mathbf{x}_1$ and $\mathbf{x}_2$ for the selection and substantial equation may or may not be equal. However, the model is well identified if the exclusion restriction is fulfilled, that is, if $\mathbf{x}_1$ includes a component that has substantial explanatory power but that is not present in $\mathbf{x}_2$ (see Heckman 1979). If the exclusion restriction is not fulfilled, the consequence is perfect multicollinearity and the equations cannot be estimated.

We can now define the observed variables

$$y_{1i} = \begin{cases} 0 & \text{if } y_{1i}^* \leq 0 \\ 1 & \text{if } y_{1i}^* > 0 \end{cases}$$

and

$$y_{2i} = \begin{cases} 0 & \text{if } y_{1i} = 0 \\ y_{2i}^* & \text{if } y_{1i} = 1, \end{cases}$$

where $y_{1i} = 1$ corresponds to a responder and $y_{1i} = 0$ to a nonresponder, and we observe the outcome $y_{2i}$ only if the latent selection variable $y_{1i}^*$ is positive.

Note that the self-selection model can be alternatively written such that the selection equation becomes

$$y_{1i} = \mathbb{I}\left[y_{1i}^* > 0\right]$$

and the substantial equation becomes

$$y_{2i} = \mathbb{I}\left[y_{1i}^* > 0\right] y_{2i}^*,$$

where $\mathbb{I}[\,\cdot\,]$ is the indicator function.

Hence, the likelihood function of the self-selection model is

$$L = \prod_{i=1}^{N} \left\{\Pr\left[y_{1i}^* \leq 0\right]\right\}^{1-y_{1i}} \left\{f_{2|1}\left(y_{2i}|y_{1i}^* > 0\right) \cdot \Pr\left[y_{1i}^* > 0\right]\right\}^{y_{1i}} \tag{7}$$

where the first term is the contribution of nonresponders and the second term is the contribution of responders. In other words, the density of $y_{2i}$ is the same as that of $y_{2i}^*$ for $y_{1i} = 1$ and is equal to the probability of observing $y_{1i}^* \leq 0$ if $y_{1i} = 0$.

The conditional density in Equation (7) can be written as follows

$$f_{2|1}\left(y_{2i}|y_{1i}^* > 0\right) = \frac{1}{1 - F_1(0)}\left[f_2(y_{2i}) - \frac{\partial}{\partial y_2} F(0, y_{2i})\right]$$

where $F_1(0) = \Pr\left\{y_{1i}^* \leq 0\right\} = \Pr\{y_{1i} = 0\}$ and $F(\,\cdot\,,\,\cdot\,)$ is the bivariate joint cdf (cumulative distribution function). Substituting the conditional density form into (7) yields

$$L = \prod_{i=1}^{N} \{F_1(0)\}^{1-y_{1i}} \left\{f_2(y_{2i}) - \frac{\partial}{\partial y_2} F(0, y_{2i})\right\}^{y_{1i}}. \tag{8}$$

## 4. Copulas Applied to Self-Selection

The likelihood function of the self-selection model (8) can be re-expressed in a more flexible way using copulas. In particular, in (8) the derivative of the joint cdf, following Sklar's theorem and its corollary, can be written as

$$\frac{\partial}{\partial y_2} F(0, y_{2i}) = \frac{\partial}{\partial v} C_\theta(F_1(0), v)\bigg|_{v \to F_2} \cdot \frac{\partial F_2}{\partial y_2}.$$

Thus the likelihood function (8) can be written in terms of copulas as follows

$$L = \prod_{i=1}^{N} \{F_1(0)\}^{1-y_{1i}} \left\{ \left[1 - \frac{\partial}{\partial F_2} C_\theta(F_1, F_2)\right] \cdot f_2(y_{2i}) \right\}^{y_{1i}}. \tag{9}$$

### 4.1. Heckman's Model

Heckman's model is also called the Normal model. He supposes that the marginal latent variables $Y_1^*$ and $Y_2^*$ are distributed according to Gaussian models, such that:

$$Y_1^* \sim N(\mathbf{x}_1\beta_1, 1) \qquad Y_2^* \sim N\left(\mathbf{x}_2\beta_2, \sigma_2^2\right),$$

where $\sigma_1^2 = 1$. As a consequence the error terms follow a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \theta \\ \theta & \sigma_2^2 \end{pmatrix}\right).$$

The likelihood function in this case takes the form

$$L = \prod_{i=1}^{N} \{1 - \Phi(\mathbf{x}_{1i}\beta_1)\}^{1-y_{1i}} \left\{ \frac{1}{\sigma_2} \phi\left(\frac{y_{2i} - \mathbf{x}_{2i}\beta_2}{\sigma_2}\right) \Phi\left(\frac{\mathbf{x}_{1i}\beta_1 + \frac{\theta}{\sigma_2}(y_{2i} - \mathbf{x}_{2i}\beta_2)}{\sqrt{1-\theta^2}}\right) \right\}^{y_{1i}}, \tag{10}$$

where the left term corresponds to non self-selection, while the right term corresponds to self-selection.

Heckman's assumption of a joint normal distribution for the error terms is overly restrictive, limiting the applicability of his approach. As pointed out by Lee (1983), the copula approach can be used to relax the traditional assumption that the marginal distributions are normal. Indeed, the marginals are very often not normally distributed, especially financial variables. Smith (2003) provides a general copula-based framework for Heckman's model by demonstrating that copulas can be used to extend the standard analysis to any bivariate distribution with given marginals (see also Smith 2005). The use of normal marginals and normal copula leads us to the traditional Heckman's method, as is shown by comparing Equations (10) and (8) (see Bhat and Eluru 2009). However, with significant departures from normality for the marginals and/or the copula, the traditional Heckman's approach is no longer sufficiently general and the use of the copula approach is essential to provide the flexibility necessary for modelling the data and the dependencies in the correct way. The following sections will demonstrate how the Bayesian approach allows us to incorporate different sources of information into the generalized Heckman's model, and how this technique can be applied to nonresponse modelling.

According to the copula approach, the likelihood has to be calculated using the (8). The expressions of the derivatives $\frac{\partial}{\partial F_2} C_\theta(F_1, F_2)$ for each type of copula are listed in Table 2.

### 4.2. *The Bayesian Approach*

In order to integrate specific data with official data sources, we use the Bayesian approach, specifying informative priors using official information. The Bayesian approach is based upon the idea that the interviewer begins with some prior beliefs about the system and then updates these beliefs on the basis of observed data. This updating procedure is based upon Bayes' Theorem:

$$\pi(\eta|\text{data}) \propto f(\text{data}|\eta)p(\eta) \qquad (\text{posterior} \propto \text{likelihood} \times \text{prior}),$$

where $\eta$ is the parameter vector. Generally, parameter estimates are determined employing the Markov Chain Monte Carlo (MCMC) method, which uses algorithms to sample observations from the posterior distribution based on the construction of a Markov chain that has the posterior as its equilibrium distribution. The state of the chain after a number of steps is then used as a sample of the posterior distribution. If the prior distributions are conjugate, general MCMC algorithms are not needed, but simpler techniques, like the Gibbs Sampler, may be used (see Albert and Chib 1993; Gamerman and Lopes 2006; Armero et al. 2008).

In order to apply the Bayesian approach to the generalized Heckman's model, we specify prior distributions for the vectors of parameters of the selection equation $\beta_1$ and of the substantial equation $\beta_2$, for the variance parameter $\sigma_2^2$, and for the copula dependence parameter $\theta$. Then, we sample from the posterior distribution by implementing a Metropolis-within-Gibbs algorithm.

We assume a multivariate normally distributed vague prior for the selection equation parameter vector $\beta_1 \sim N(\mu_1, \Sigma_1)$ where $\mu_1$ is a $(n_1 + 1)$-dimensional vector of zeros and $\Sigma_1 = 100 I_{n_1+1}$, with $I_{n_1+1}$ the $(n_1 + 1)$-dimensional identity matrix. Like for the parameter vector $\beta_1$, we consider a multivariate normal prior for the substantial equation parameter vector $\beta_2$, but we used information from official statistics to define informative prior distributions. Hence, $\beta_2 \sim N(\mu_2, \Sigma_2)$, where $\mu_2$ is a $(n_2 + 1)$-dimensional vector and

*Table 2. Expressions for the copula derivatives $\frac{\partial}{\partial F_2} C_\theta(F_1, F_2)$.*

| Copula | Expression for $\frac{\partial}{\partial F_2} C_\theta(F_1, F_2)$ |
|---|---|
| Gaussian | $\Phi\left(\frac{\Phi^{-1}(u_1)-\theta\Phi^{-1}(u_2)}{\sqrt{1-\theta^2}}\right)$ |
| FGM | $u_1[1 + \theta(1 - u_1)(1 - 2u_2)]$ |
| AMH | $\frac{(1-\theta)u_1+\theta u_1^2}{(1-\theta(1-u_1)(1-u_2))^2}$ |
| Clayton | $u_2^{-(\theta+1)}\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-\left(\frac{1+\theta}{\theta}\right)}$ |
| Gumbel | $u_2^{-1}(-\ln(u_2))^{\theta-1} \cdot C_\theta(u_1, u_2)\left[(-\ln(u_1))^\theta + (-\ln(u_2))^\theta\right]^{\left(\frac{1}{\theta}-1\right)}$ |
| Frank | $\left[1 - e^{\theta C_\theta(u_1,u_2)}\right](1 - e^{\theta u_2})^{-1}$ |

$\Sigma_2$ is the $(n_2 + 1)$-dimensional prior covariance matrix. For $\sigma_2^2$ we consider the vague prior $\sigma_2^2 \sim \Gamma^{-1}(a, b)$ where $a = 0.001$ and $b = 0.001$. As prior distribution for $\tau$ we consider the vague prior $\tau \sim Beta(\alpha, \beta)$ extended to the range $[-1, 1]$ (Huard et al. 2006), where $\alpha = 1$ and $\beta = 1$.

## 5. Innovation Survey Data

The methodology illustrated in the previous section was tested using two datasets: a national-level survey and an official EU-level survey dataset.

The first dataset is available on the ISTAT (Italian National Institution of Statistics) website and it contains data collected through a survey on innovations introduced and innovative activities undertaken by a sample of Italian firms between 2008 and 2010.

The Italian Innovation Survey, carried out on a two-year basis, collects information about new or significantly improved goods or services (product innovations) and new or significantly improved processes, logistics or distribution methods (process innovations), as well as about organizational and marketing innovation. The original data were perturbed by ISTAT, in order to guarantee the privacy of respondents (see ISTAT 2013).

From the original ISTAT dataset, we only selected SMEs, that, according to the definition provided by the European Union, include enterprises which employ fewer than 250 persons and which have an annual turnover not exceeding 50 million euros, and/or an annual balance sheet total not exceeding 43 million euros.

Moreover, we restricted our attention to the reference period of 2010, hence limiting the number of firms in the dataset to 4,266.

Therefore, from a total number of 3.8 millions of Italian SMEs in 2010, we only considered survey information of a small sample of about 4,200 firms.

The variables we used from the innovation survey dataset are described in Table 3.

We integrated the ISTAT innovation survey data with a second dataset, the 2010 Innovation Union Scoreboard (IUS) provided by the European Union (see European Commission 2010). IUS provides a comparative assessment of the research and innovation performance of the EU Member States and the relative strengths and weaknesses of their research and innovation systems.

In particular, we used data about human resources, firms' activities, and outputs, considering the following variables:

- human resources who completed tertiary education,
- business R&D firm expenditure,
- non-R&D innovation firm expenditure,
- firms introducing product or process innovations,
- firms introducing marketing/organizational innovations,
- knowledge-intensive services exports,
- sales of new-to-market and new-to-firm innovations.

We used the IUS variables listed above to define informative prior distributions for the substantial equation parameters $\beta_2$ of the generalized Heckman's model, described in Section 4. The parameters of these informative priors were defined based on the empirical distributions of the corresponding IUS variables. This approach allows us to integrate the

*Table 3. Description of the innovation survey dataset variables.*

| Innovation survey dataset | |
|---|---|
| Variable names | Variable label |
| turn | turnover |
| rrdinx | expenses for activities of R&D |
| rrdexx | expenses for acquisitions of R&D services |
| rmacx | expenses for acquisition of machinery and equipment |
| roekx | expenses for acquisition of other external technologies |
| rdsgx | expenses for design activities |
| rprex | expenses for other innovative activities |
| rtrx | expenses for education on innovative activities |
| rmarx | expenses for marketing of innovative products |
| empdeg | number of employees with a university degree |
| turnmar | turnover coming from new products or services (or significantly improved products and services) for the reference market |
| turnin | turnover coming from new products or services (or significantly improved products and services) for the firm only |

ISTAT national data source with the more general IUS international data source, provided by the European Commission.

## 5.1. The Model

We suppose the firms that did not respond to the questionnaire are those belonging to the business and other services and nonmarketed services NACE macrosectors. The percentage of respondent firms is 85.07%, while the percentage of nonrespondent firms is 14.93%.

We assume a Normal distribution for the marginal $Y_1^*$ (selection equation)

$$Y_1^* \sim N(\mathbf{x}_1\beta_1, 1)$$

and a log-normal distribution for $Y_2^*$ (substantial equation), after a graphical examination of the variable and the application of the Kolmogorov-Smirnov test, which accepts log-normality:

$$\log Y_2^* \sim N\left(\mu_l, \sigma_l^2\right),$$

where $\mu_l = e^{\mathbf{x}_2\beta_2 + \sigma^2/2}$ and $\sigma_l^2 = \left(e^{\sigma^2} - 1\right)e^{2\mathbf{x}_2\beta_2 + \sigma^2}$. Figure 2 shows the histogram of the variable Turnover.

In the model, the target variable $y_2$ is *turn*; the vector $\mathbf{x}_1$ comprises the above eleven variables listed in Table 3. The model is well identified if the exclusion restriction is fulfilled, that is, if $\mathbf{x}_1$ includes a component (*empdeg*) that has substantial explanatory power but that is not present in $\mathbf{x}_2$.

## 5.2. Results

We run the Metropolis-within-Gibbs algorithm for 10,000 iterations and discarded the first 2,000 iterations as the burn-in period. Because of space considerations, we here analyze the MCMC traceplots of the model using the Clayton copula, since the results obtained

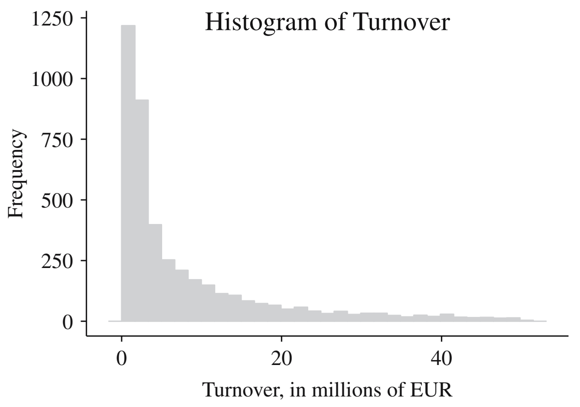


*Fig. 2. Histogram of Turnover.*

with the other choices of copulas are very similar to those presented. The trace plots of the parameters $\beta_1$, $\beta_2$, $\sigma^2$ and $\theta$ are listed in the Appendix. The sample paths show that the chains are well mixing, freely exploring the sample space.

Parameter estimates for the selection and substantial equations are very stable, as shown in Figures 3 and 4, representing credible intervals for $\beta_1$ and $\beta_2$, respectively. A credible interval is computed from the posterior distribution and is the interval within which the probability of the parameter of interest falling in is given by the level of credibility. The credible intervals are all very similar for the different choices of copula. The only exceptions are the credible intervals of the $\beta_2$ parameters modeled with the independence copula. However, this was expected, since the independence copula assumes no association between the selection and substantial equations. The results of the $\beta_1$ parameters indicate which variables are associated with response. From Figure 3, the variables with a significant negative influence on the response are *rrdinx*, *roekx* and *empdeg*, while the variable with a significant positive influence on the response is *rmarx*. This means that firms that invest in R&D and external technologies, do not invest in marketing, and employ several graduates, are nonrespondents. The $\beta_2$ parameters indicate which variables explain the firms' turnover. Figure 4 suggests that the variables with a significant positive

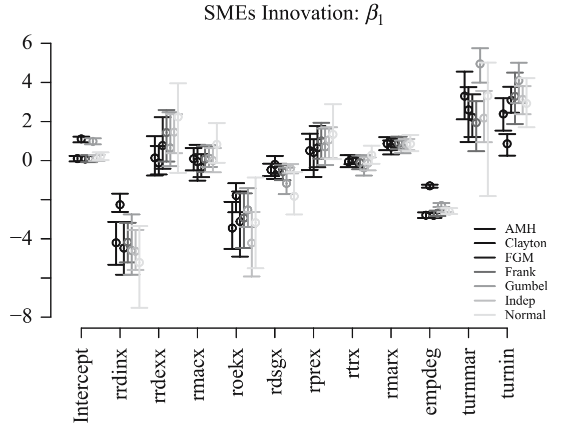


*Fig. 3. Credible intervals of $\beta_1$ for all copulas considered at 95% level.*

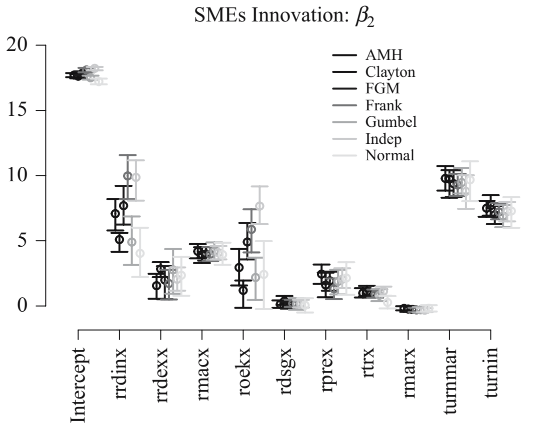


*Fig. 4. Credible intervals of $\beta_2$ for all copulas considered at 95% level.*

influence on the firms' turnover are *rrdinx*, *turnmar*, *turnin*, *rmacx*, *rrdexx*, and *rprex*. This means that firms investing in R&D, machinery equipment, new products and services, and other innovative activities show a high turnover.

Figures 5 and 6 show the boxplots of the posterior distributions of the parameters $\theta$ and $\tau$. As can be seen from the plots, the dependence parameters $\tau$ are positive, meaning that the nonrespondent SMEs (firms that did not fill in the questionnaire) are those with high turnover. The values of Kendall's $\tau$ denote a moderate degree of dependence for almost all the different types of copulas.

### 5.2.1. Model Comparison

We compare the performances of the different copula models using the Deviance Information Criterion (DIC), which has the following expression

$$DIC = \bar{D} + p_D$$

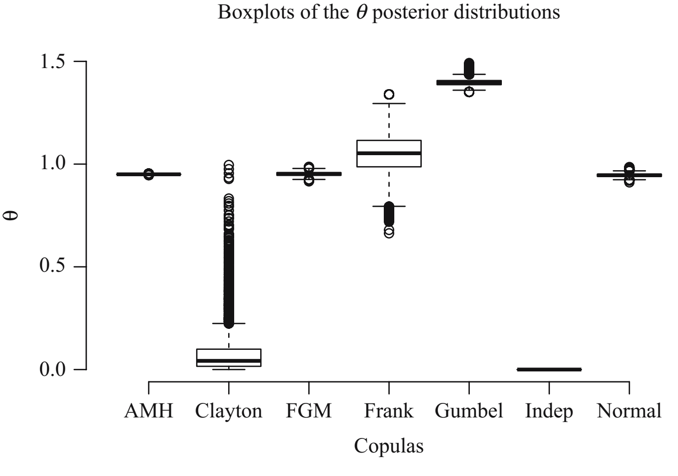


*Fig. 5. Boxplots of the posterior distributions of $\theta$ for the different copulas.*

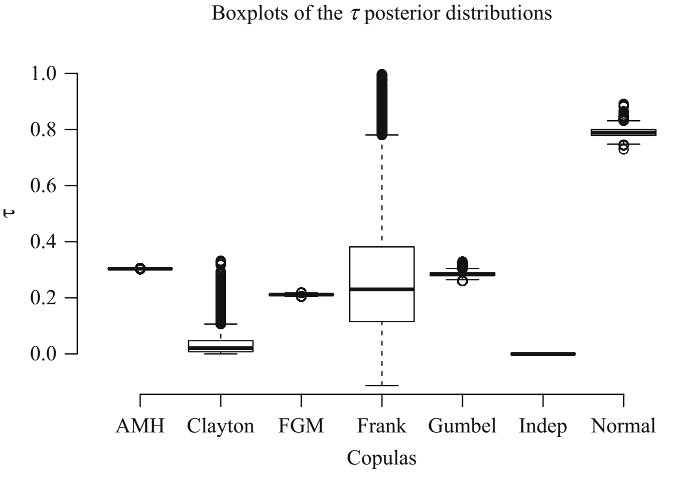


*Fig. 6. Boxplots of the posterior distributions of $\tau$ for the different copulas.*

where $\bar{D} = E(-2\log[L(\text{data}|\eta)])$ is the average of the log-likelihoods calculated at the end of each MCMC iteration, $p_D = \bar{D} - \hat{D}$ and $\hat{D} = -2\log[L(\text{data}|\eta*)]$ is the log-likelihood calculated using the parameter posterior means. Models with smaller DIC are better supported by the data.

Table 4 lists the DIC results for the different copulas. The Clayton copula model outperforms the others, since it has the lowest DIC value. Therefore, the Clayton copula is the one that best models the relationship between Heckman's equations. The main advantage that the Clayton copula offers over the Normal is that the unequal tail dependence, which is stronger in the left tail, is properly accounted for, leading to more accurate results.

Finally, in order to correctly estimate our target variable, that is the turnover of the SMEs, we need to consider the dependence value estimated through the most suitable copula for our data. The mean turnover can be calculated as

$$E\left[Y|Y_1^* > 0\right] = \int_0^\infty y f_{2|1}\left(y|Y_1^* > 0\right)dy = \frac{1}{1 - F_1(0)}\left(E(Y) - \int_0^\infty y\frac{\partial}{\partial F_2}C_\theta(F_1, F_2)f_2 dy\right)$$

where the result was evaluated at $\mathbf{x} = \bar{\mathbf{x}}$, the covariate averages across the total number of firms. Figure 7 shows the histogram of the mean turnover value for the SMEs,

*Table 4. Model comparison.*

| | DIC |
|---|---|
| AMH | −43599.06 |
| Clayton | **−50993.80** |
| FGM | −43678.94 |
| Frank | −43681.95 |
| Gumbel | −45594.26 |
| Indep | −43767.92 |
| Normal | −43593.09 |

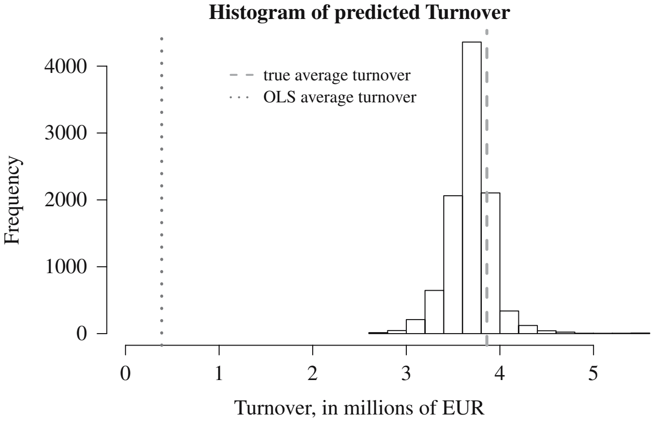


*Fig. 7. Histogram of Turnover predicted via the Bayesian generalized Heckman approach. The plot compares the copula estimate for the average turnover with the biased OLS estimates.*

calculated from the MCMC simulations. The dashed line represents the true average value of turnover for the observed dataset, while the dotted line represents the average value of turnover predicted by the traditional OLS model, which is based on the Normal copula and the log-transformation of $y_2$. Please note that the true value of turnover is available, since self-selection was artificially introduced in the Innovation survey dataset, as explained in Subsection 5.1. This result shows that the use of the OLS model in presence of self-selection is completely unrealistic and underestimates the true value of the target variable. The generalized Heckman's model using the Clayton copula performs well and accurately predicts the true value of turnover, since the predicted turnover is very close to its true value. The Clayton copula in this case is more flexible than the traditional Normal copula in capturing asymmetric tail dependence, and it gives more reliable predictions.

## 6. Concluding Remarks

This article illustrated the application of the Bayesian generalized Heckman approach to correct the self-selection bias integrating different sources of information.

This approach has a number of potential applications, especially where survey data are employed. The use of official statistics in sector and marketing analysis by firms is one of them. However, this approach can be successfully implemented in education, medical, and social studies.

A limitation of the study could be the computational complexity in some cases. However, the main advantage is the accuracy of the results compared to traditional approaches.

Further studies may include the analysis of additional families of copulas and their rotated versions.

## Appendix

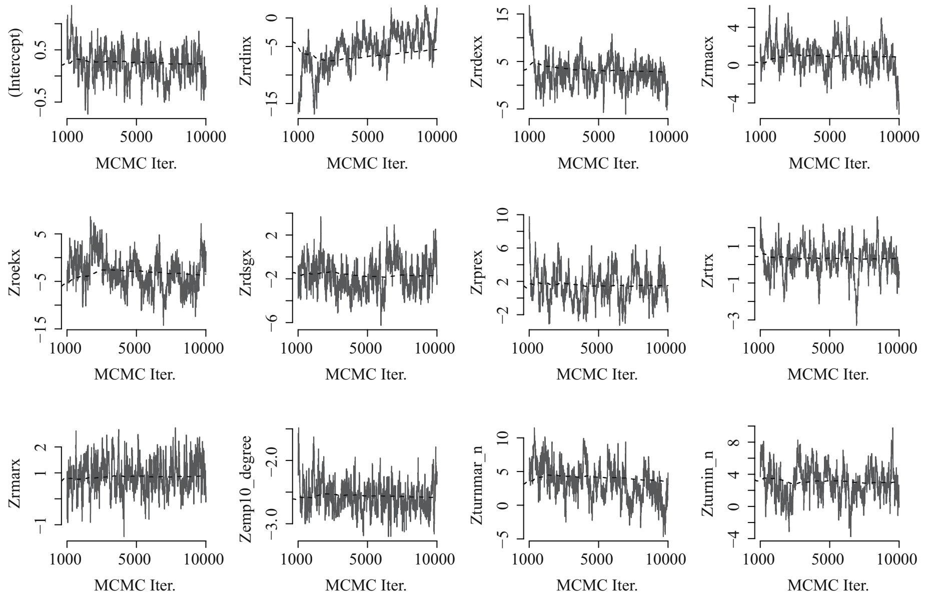


*Fig. 8. Trace plots of the $\beta_1$ parameters for the Clayton copula model. The labels on the vertical axes refer to the names of the variables.*

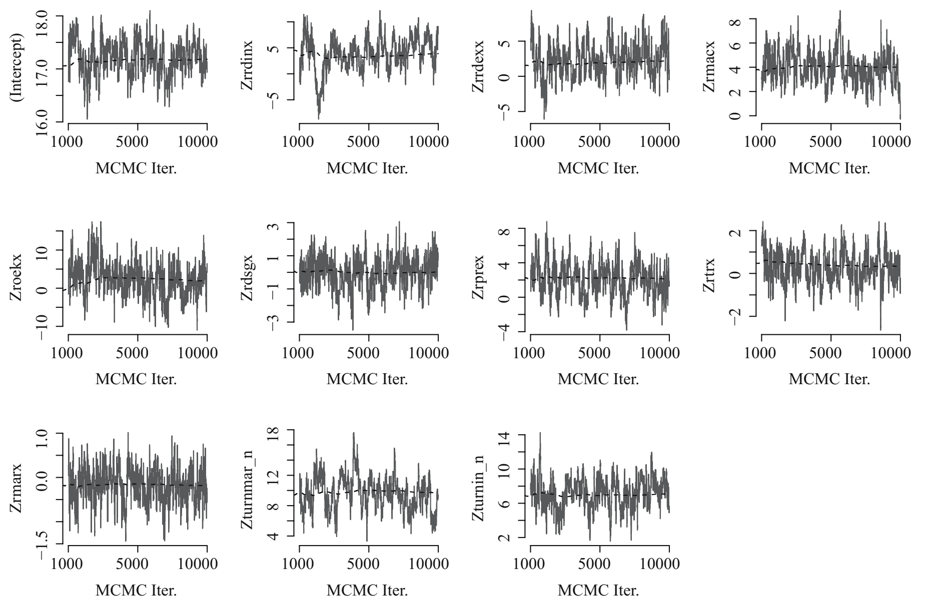


*Fig. 9. Trace plots of the $\beta_2$ parameters for the Clayton copula model. The labels on the vertical axes refer to the names of the variables.*

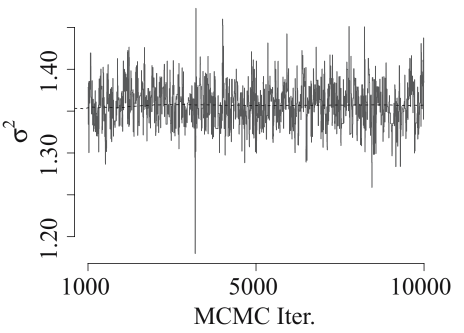


*Fig. 10. Trace plots of the $\sigma^2$ parameter for the Clayton copula model.*

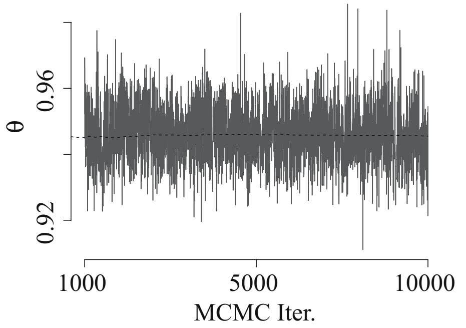


*Fig. 11. Trace plots of the $\theta$ parameter for the Clayton copula model.*

## 7. References


Albert, J.H. and S. Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88: 669–679. Doi: http://dx.doi.org/10.2307/2290350.

Ali, M.M., N.N. Mikhail, and M.S. Haq. 1978. "A Class of Bivariate Distributions Including the Bivariate Logistic." *Journal of Multivariate Analysis* 8: 405–412. Doi: http://dx.doi.org/10.1016/0047-259X(78)90063-5.

Armero, C., A. López-Quílez, and R. López-Sánchez. 2008. "Bayesian Assessment of Times to Diagnosis in Breast Cancer Screening." *Journal of Applied Statistics* 35: 997–1009. Doi: http://dx.doi.org/10.1080/02664760802191397.

Bhat, C.R. and N. Eluru. 2009. "A Copula-Based Approach to Accommodate Residential Self-Selection Effects in Travel Behavior Modeling." *Transportation Research Part B* 43: 749–765. Doi: http://dx.doi.org/10.1016/j.trb.2009.02.001.

Clayton, D.G. 1978. "A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Family Tendency in Chronic Disease Incidence." *Biometrika* 65: 141–151. Doi: http://dx.doi.org/10.1093/biomet/65.1.141.

European Commission. 2010. "Innovation Union Scoreboard 2010." Available at: http://ec.europa.eu/enterprise/policies/innovation/files/ius/ius-2010_en.pdf (accessed February 2014).

Farlie, D.J.G. 1960. "The Performance of Some Correlation Coefficients for a General Bivariate Distribution." *Biometrika* 47: 307–323. Doi: http://dx.doi.org/10.2307/2333302.

Frank, M.J. 1979. "On the Simultaneous Associativity of F(x, y) and x + y − F(x, y)." *Aequationes Mathematicae* 19: 194–226.

Gamerman, D. and H. Lopes. 2006. "Markov Chain Monte Carlo." *Texts in Statistical Science*, 2nd ed. New York: Chapman & Hall.

Gumbel, E.J. 1960. "Bivariate Exponential Distributions." *Journal of the American Statistical Association* 55: 698–707.

Hamilton, B.H. and J.A. Nickerson. 2003. "Correcting for Endogeneity in Strategic Management Research." *Strategic Organization* 1: 51–78. Doi: http://dx.doi.org/10.1177/1476127003001001218.

Heckman, J.J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47: 153–161.

Huard, D., G. Evin, and A.-C. Favre. 2006. "Bayesian Copula Selection." *Computational Statistics and Data Analysis* 51: 809–822. Doi: http://dx.doi.org/10.1016/j.csda.2005.08.010.

ISTAT. 2013. *Italian Innovation Survey*. Available at: http://www.istat.it (accessed February 2014).

Joe, H. 1993. "Parametric Families of Multivariate Distributions with Given Marginals." *Journal of Multivariate Analysis* 46: 262–282. Doi: http://dx.doi.org/10.1006/jmva.1993.1061.

Joe, H. 1997. *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Lee, L.-F. 1983. "Generalized Econometric Models with Selectivity." *Econometrica* 51: 507–512. Doi: http://dx.doi.org/10.2307/1912003.

Lucchetti, R. and C. Pigini. 2013. "A Test for Bivariate Normality with Applications in Microeconometric Models." *Statistical Methods and Applications* 22: 535–572. Doi: http://dx.doi.org/10.1007/s10260-013-0236-5.

Morgenstern, D. 1956. "Einfache Beispiele zweidimensionaler Verteilungen." *Mitteilungsblatt für Mathematische Statistik* 8: 234–235.

Nelsen, R.B. 1999. *An Introduction to Copulas*. New York: Springer.

Nicolini, G. and L. Dalla Valle. 2011. "Errors in Customer Satisfaction Surveys and Methods to Correct Self-Selection Bias." *Quality Technology & Quantitative Management* 8: 167–181. Doi: http://dx.doi.org/10.1080/16843703.2011.11673254.

Nicolini, G. and L. Dalla Valle. 2012. "Census and Sample Surveys." In *Modern Analysis of Customer Surveys, with Applications Using R*, edited by Ron S. Kenett and Silvia Salini, 37–53. Statistics in Practice Series. Hoboken, NJ: John Wiley & Sons.

Rosenbaum, P. and D. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55. Doi: http://dx.doi.org/10.1093/biomet/70.1.41.

Rubin, D.B. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies." *Journal of Educational Psychology* 66: 688–701. Doi: http://dx.doi.org/10.1037/h0037350.

Sklar, A. 1959. "Fonctions de répartition a *n* dimensions et leurs marges." *Publications de l'Institut de Statistique de l'Universite de Paris* 8: 229–231.

Smith, M. 2003. "Modelling Sample Selection Using Archimedean Copulas." *Econometrics Journal* 6: 99–123. Doi: http://dx.doi.org/10.1111/1368-423X.00101.

Smith, M.D. 2005. "Using Copulas to Model Switching Regimes with an Application to Child Labour." *The Economic Record* 81: S47–S57.

Tobin, J. 1958. "Estimation of Relationships for Limited Dependent Variables." *Econometrica* 26: 24–36. Doi: http://dx.doi.org/10.2307/1907382.

Received April 2015
Revised October 2015
Accepted March 2016

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 907–912, http://dx.doi.org/10.1515/JOS-2016-0047

# Invited Commentary Special Section: The Role of Official Statistics in Statistical Capacity Building

*John Pullinger*[1]

## 1. Introduction and Outline

The United Nations Fundamental Principles of Official Statistics place official statistics as an indispensable element in the information system of a democratic society (United Nations Economic and Social Council 2013). In that light, official statistics are themselves a critical element of national capability, providing a trustworthy foundation for decision makers and an essential tool for citizens to make good choices in their own lives. Official statistics also hold decision makers, in the public and private sectors, to account.

The capacity built within systems of official statistics is also part of wider civil society. This capacity supports a wider infrastructure within nations, regions, and globally, incorporating academia, business, and civil society, including national statistical societies and the International Statistical Institute.

Equally, statistical capacity developed through the education system in schools and universities and through continuous professional development in businesses, provides a statistical ecosystem that is mutually reinforcing, fostering a resource that can benefit official statistics as well as other fields of endeavour.

Especially salient is statistical literacy, which needs to be nurtured within all parts of society including amongst politicians, officials, and the media as well as in the worlds of finance, commerce, and academia. High levels of statistical literacy in society create a virtuous circle – the greater the statistical expertise of the using community the greater the demand for high quality data.

The current imperative to build statistical capacity of official statisticians as discussed by Deutsch, as well as Forbes and Keegan in the special section of this issue, is strong as data becomes the defining resource of our age. The ability to mobilise the power of data to help individuals, organisations, and states make better decisions is becoming the differentiator between those who thrive and those who fail to do so.

## 2. Context – History and Current

The root of the word statistics hints at the fundamental value it has in the effective functioning of a state. Notions from the 18th century that the purpose of statistics is to improve the quantum of happiness of the people reference its potential to empower citizens and transform lives (Pullinger 2013). However, information is always associated with power, and that power can be used benevolently or malevolently. A vital aspect of the

[1] National Statistician, UK Statistics Authority, 1 Drummond Gate, London SW1V 2QQ, UK. Email: national.statistician@statistics.gov.uk

public value of statistics is openness – helping to challenge decisions as well as to make them.

The history of official statistics in many nations links to the ups and downs of the countries themselves. In several instances (positive and negative) critical moments in official statistics parallel critical moments for the nation (Seltzer 1994).

This is also so at the international level. The role of the Conference of European Statisticians during the period before 1989 enabled it to be ready in the years afterwards to play a pivotal role in supporting the new democracies of Central and Eastern Europe. These nations needed to develop statistical capability, by, among other things, the adoption of fundamental principles of official statistics. These principles were taken up by other regions and finally adopted by the United Nations General Assembly in 2014, the same year as the UN Secretary General published his report on the impact of the data revolution on global development (United Nations 2014b), emphasising the need for continuous evolution of official statistics, which MacFeely discusses in the special section of this issue.

## 3. Official Statistics as a Part of National and Global Capacity

Over the last two years the role of official statistics as part of national and global capacity has been given sharpened focus by the adoption by the UN General Assembly of Sustainable Development Goals for the period 2015–2030 (United Nations Department of Economic and Social Affairs 2014). The goals include specific proposals to provide the means of implementation for the ambitions set out. These include enhanced capacity-building support to developing countries to increase significantly increase the availability of high-quality, timely, and reliable data disaggregated by income, gender, age, race, ethnicity, migratory status, disability, geographic location, and other characteristics relevant in national contexts (United Nations 2014a). The agreement reached in Addis Ababa on financing for development provides a strong underpinning for this in the call for international support and partnerships for implementing effective and targeted capacity-building in developing countries (United Nations Department of Economic and Social Affairs 2015).

The United Nations Statistical Commission has now agreed on a package of indicators to provide a practical starting point for measuring progress on sustainable development in the years ahead. There is work to be done in all countries but especially in Africa, the least developed countries, landlocked developing countries, and small island developing states. The world data forum that takes place in South Africa in 2017 will provide an opportunity for partners from all sectors to meet and consider how best to work together to build the capability required (United Nations Statistical Commission 2016).

The history of statistical capacity building merits detailed reflection. Despite the valiant work of PARIS21 (Soremho-Ramos 2015) and others arising from the follow up to the Millennium Development Goals, sustained and successful capacity development remain elusive. We must learn from this experience to be more successful in the years ahead.

## 4. Official Statistics Supporting National Statistical Capability

Of course, official statistics themselves already provide a resource that can help grow wider capability. As an employer, National Statistical Institutes can draw on academic training and there are many examples of positive symbiosis. The European Union Masters in Official Statistics programme is an example (Eurostat (n.d.)). The plan in the UK to develop an Economic Statistics Centre of Excellence is another highlighting an increasing trend towards using partnership approaches to develop capacity in the medium term (Office for National Statistics 2016).

At another level, the services provided by statisticians can help generate innovation and entrepreneurship in other sectors. The UK has been in vanguard of supporting Open Data with many positive outcomes across the economy (Open Government Partnership (n.d.)), also highlighting benefits for small and medium-sized enterprises, which Coleman details in the special section of this issue. Other UK initiatives, such as the Justice Data Lab, that enables third sector organisations to benefit from statistical analysis of sensitive data without compromising confidentiality, can provide a boost to the contribution statistics makes to wider society (Ministry of Justice 2014). Another example is the Integrated Data Infrastructure in New Zealand, which has helped solve complex issues such as on crime and vulnerable children to improve outcomes for citizens (Statistics New Zealand 2016).

## 5. National Statistical Capability Supporting Official Statistics

Mutual dependence goes further. Official statistics in a country benefits enormously from a vibrant national statistical capability in the wider community. Of particular importance is the role of national statistical societies (and at the global level of the International Statistical Institute). The Royal Statistical Society (RSS) for example has instituted a strategy designed to play a pivotal role in supporting statistical systems. The strategy is focused on four core themes: the discipline of statistics; the statistical profession; statistics in public policy; and statistical literacy (Royal Statistical Society 2013). All of these areas benefit and add capability to official statistics and reflect the public value offered by strong statistical communities in universities, in workplaces (especially those that recognise the contribution made by chartered statisticians), and in the public realm.

The nurturing of statistical literacy has a special place. Indeed, the adoption of the Royal Statistical Society strategy was accompanied with the launch of the Getstats campaign for statistical literacy (Royal Statistical Society (n.d.)). The campaign was designed to raise levels of statistical understanding across society and thereby raise national statistical capability. The campaign focused initially on two critical groups – politicians and the media. Seminars and skills training opportunities for Members of Parliament and their staff were well received. Journalists had already had strong engagement with the RSS awards for statistical excellence and were ready to take up further opportunities (Royal Statistical Society 2016). These examples show that when statistics presents itself as something that can help people do their own jobs better, they will want to get involved. Statistics also complement wider developments such as the review of the way the BBC used statistics in the most recent UK general election (BBC Trust 2016).

Within schools, initiatives such as CensusAtSchool has demonstrated the value of good curriculum based materials in the classroom (CensusAtSchool UK (n.d.)). This and other initiatives paved the way to make the case for better statistical and data curriculum content across the age and subject range, especially in the 16–18 age group. Also at university level, the Q-Step programme developed by the Economic and Social Research

Council and the Nuffield Foundation has provided quantitative training for university students of a wide variety of social science disciplines from politics to sociology (Nuffield Foundation. (n.d.)).

Another valuable initiative has been that taken by the policy profession in the UK civil service to incorporate statistical training in the core curriculum for their professional development. What links all these initiatives is the fact that if the consumers of official statistics are more statistically literate they will be better able to get value from them. They will also be likely to be more critical of what they get, thereby helping to drive up quality in a virtuous circle.

The role of official statistics in statistical capacity building can thus be characterised as the combination of developing both the supply side and the demand side of the equation.

## 6. The Opportunity of the Data Revolution

The data revolution presents an opportunity to turn this analysis into action in a profound way. As data becomes the defining resource of our age, those who can add value to that resource and create meaning and insight from it will have a special contribution to make to our economies and societies. The thinking required is illustrated by the idea of moving from quality to information quality, see Kenett and Shmueli in the special section of this issue. Building the data capability of statistical organisations is thus a pressing imperative. At the global level there is the potential for a grand partnership between governments, businesses, and civil society. There is an opportunity for those with a special position linking these communities to step up. The International Statistical Institute, International Association of Official Statisticians, International Association of Statistical Education and many others have much to contribute. Similarly within the regions and nations there are partnerships to be encouraged.

In the UK, the mutual benefits accruing to UK official statistics and the Royal Statistical Society from a vibrant relationship could inspire others.

The launch of a Data Science Campus at the ONS headquarters in Newport is a symbol to the organisation and to the nation that we are taking that opportunity (Office for National Statistics (n.d.)). At the heart of our strategy is the enhancement of technological and human capability with major investments in a learning academy for our staff and apprenticeships to train the next generation of colleagues (Office for National Statistics 2014). We are also ready reach out and make common cause with those interested in the development of official statistics that serve the public good. Central to this is the desire to work with colleagues in other countries, especially those who struggle to access the necessary resources internally. Here we can work with UN and other agencies to form new kinds of partnerships and together mobilise the power of data to help the world make better decisions.

## 7. References

BBC Trust. 2016. "BBC Trust Impartiality Review: Making Sense of Statistics." Available at: http://downloads.bbc.co.uk/bbctrust/assets/files/pdf/our_work/stats_impartiality/report.pdf (accessed September 2016).

CensusAtSchool UK. (n.d.). *Welcome to CensusAtSchool UK*. Available at: http://www.censusatschool.com/ (accessed September 2016).

Eurostat. (n.d.). *European Master in Official Statistics*. Available at: http://ec.europa.eu/eurostat/web/european-statistical-system/emos (accessed September 2016).

Ministry of Justice. 2014. *Accessing the Justice Data Lab service*. Available at: https://www.gov.uk/government/publications/justice-data-lab (accessed September 2016).

Nuffield Foundation. (n.d.). *Q-Step*. Available at: http://www.nuffieldfoundation.org/q-step (accessed September 2016).

Office for National Statistics. 2014. "Better Statistics, Better Decisions: Strategy for UK statistics, 2015 to 2020." Strategy launched on 1 November 2014. Available at: https://www.statisticsauthority.gov.uk/archive/news/uk-statistics-authority-statement-of-strategy-2015-2020-better-statistics-better-decisions.pdf (accessed September 2016).

Office for National Statistics. 2016. *Economic Statistics Centre of Excellence*. Available at: https://www.ons.gov.uk/news/statementsandletters/economicstatisticscentreofexcellenceescoe (accessed September 2016).

Office for National Statistics. (n.d.). *Data Science Campus*. Available at: https://www.ons.gov.uk/aboutus/whatwedo/datasciencecampus (accessed September 2016).

Open Government Partnership. (n.d.). *United Kingdom*. Available at: http://www.opengovpartnership.org/country/united-kingdom (accessed September 2016).

Pullinger, J. 2013. "Statistics Making an Impact." *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 176: 819–839. Doi: http://dx.doi.org/10.1111/rssa.12023.

Royal Statistical Society. 2013. *Strategic Plan 2014-18: A World with Data at the Heart of Understanding and Decision Making*. Available at: https://www.rss.org.uk/Images/PDF/about/strategy-summary_flr.pdf (accessed September 2016).

Royal Statistical Society. 2016. *Statistical Excellence in Journalism*. Available at: http://www.rss.org.uk/RSS/About/Recognising_Statistical_Excellence/Journalism_awards/RSS/About_the_RSS/Recognising_statistical_excellence_sub/Statistical_Excellence_in_Journalism_Awards.aspx?hkey=f10b8e1e-466a-4ef1-af2a-f8f53e04f338 (accessed September 2016).

Royal Statistical Society. (n.d.). *Statistical Literacy*. Available at: http://www.rss.org.uk/RSS/Influencing_Change/Statistical_literacy/RSS/Influencing_Change/Statistical_literacy.aspx?hkey=821bf2f4-8a09-413c-8d22-290e2209a92a (accessed September 2016).

Seltzer, B. 1994. "Politics and Statistics: Independence, Dependence or Interaction?". Department for Economic and Social Interaction and Policy Analysis Working Paper Series No. 6. Available at: http://unstats.un.org/unsd/statcom/FP-Seltzer.pdf (accessed September 2016).

Soremho-Ramos, S. 2015. "Paris21: Light Evaluation of Paris21, Final Report." Report published February 2015 commissioned at the Paris21 Board Meeting in 2014. Available at: http://www.paris21.org/sites/default/files/PARIS212015-Evaluation_0.pdf (accessed September 2016).

Statistics New Zealand. 2016. *The Integrated Data Infrastructure*. Available at: http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure.aspx (accessed September 2016).

United Nations. 2014a. *Sustainable Development Goal 17: Strengthen the Means of Implementation and Revitalize the Global Partnership for Sustainable Development*. Available at: https://sustainabledevelopment.un.org/sdg17 (accessed September 2016).

United Nations. 2014b. "A World That Counts: Mobilising the Data Revolution for Sustainable Development." Report prepared at the request of the United Nations Secretary-General, by the Independent Expert Advisory Group on a Data Revolution for Sustainable Development. Available at: http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf (accessed September 2016).

United Nations Department of Economic and Social Affairs. 2014. *Sustainable Development Goal indicators*. Available at: http://unstats.un.org/sdgs/ (accessed September 2016).

United Nations Department of Economic and Social Affairs. 2015. *Addis Ababa Action Agenda: of the Third International Conference on Financing for Development*. Available at: http://www.un.org/esa/ffd/wp-content/uploads/2015/08/AAAA_Outcome.pdf (accessed September 2016).

United Nations Economic and Social Council. 2013. "Fundamental Principles of Official Statistics." Resolution adopted by the Economic and Social Council on 24 July 2013. Available at: http://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf (accessed September 2016).

United Nations Statistical Commission. 2016. *High Level Forum of Official Statistics: Dialogue Towards the UN World Data Forum*. Meeting 7 March 2016 organised by the United Nations Statistics Division. Available at: http://unstats.un.org/unsd/statcom/47th-session/side-events/20160307-3A-towards-un-world-data-forum/ (accessed September 2016).

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 913–916, http://dx.doi.org/10.1515/JOS-2016-0048

# Invited Commentary
# Special Section: Addressing the Needs of Official Statistics Users: The Case of Eurostat

*Dr. Marleen De Smedt*[1]

As the producer of official statistics for the European Union, Eurostat aims to produce relevant and accurate European statistics (official statistics at the European level) that are directly accessible to and understood and used by a wide range of users.

On a monthly basis, the Eurostat website records more than three million visits, over four million page views, some 700,000 pdf downloads and more than one million extractions of data. These metrics rank the site among the top five websites of the European Commission (Bautier et al. 2014). But who are these users? And what are they looking for?

Eurostat obtains information about the users of its data in various ways: via bilateral contacts with Commission policy services (via "annual hearings" and through written agreements – "Memorandum of Understanding"), via requests that users address to the Eurostat Statistical Support services (which include the Eurostat User Support network and separate central services for media and institutional requests) and via the annual Eurostat User Satisfaction Surveys.

The Eurostat User Support network consists of Eurostat and Support Centres in all 28 EU countries and in a number of countries in the wider European area. This network has been in operation since 2004. Users can put forward specific questions to this network, which in turn gives a first insight of the type of users and the main type of support wanted.

In the period April 2015 to September 2015, the Eurostat User Support network processed 3,757 requests and, in the same period, 1,271 requests were sent directly to Eurostat. Of these requests sent directly to Eurostat a further analysis was made on the type of user (see Table 1).

Through an annual Eurostat User Satisfaction Survey, which includes general questions about clarity and user-friendliness of the data, Eurostat is reaching out to its users – inside and outside the Commission services – in order to learn about the type of users and uses, users' assessment of the quality of, and trust in, European statistics and on its dissemination.

The report on the 2015 User Satisfaction Survey indicated that students, academic and private users account for 43.5% of users and that "research" and "general background information" were the most common purposes for all users combined. Results also showed a high user satisfaction rate with the renewed Eurostat website, a significant improvement

[1] Advisor to the Director-General, Eurostat, European Commission, Luxembourg. Email: marleen.desmedt@ec.europa.eu

*Table 1. Type of request according to type of user of European statistics in the period April 2015 to September 2015.*

| Type of user | Type of request | | | | | Total requests |
|---|---|---|---|---|---|---|
| | Availability of data/ publication | Verification of data | Methodological request | Technical problems | Other | |
| Private user | 59 | 6 | 16 | 2 | 8 | 91 |
| Student or academic | 104 | 12 | 22 | 0 | 13 | 151 |
| Commercial company/enterprise | 137 | 26 | 43 | 13 | 5 | 224 |
| Press and other media | 10 | 2 | 2 | 1 | 1 | 16 |
| EU institution/agency | 25 | 2 | 4 | 3 | 5 | 39 |
| International organisation | 22 | 4 | 7 | 0 | 1 | 34 |
| National Statistical Institute | 334 | 109 | 136 | 53 | 29 | 661 |
| Public administration/government | 25 | 0 | 7 | 3 | 2 | 37 |
| Political parties and political organisations | 1 | 0 | 1 | 0 | 0 | 2 |
| Other | 7 | 1 | 3 | 2 | 3 | 16 |
| **Grand total** | 724 | 162 | 241 | 77 | 67 | 1271 |

Source: Eurostat Statistical Support services report (2016).

in ease of understanding European statistics, and a high level of trust among users of European statistics.

Previous User Satisfaction Surveys (before 2014) indicated that many users wanted easier access to data, "ready-to-use" statistical tables and more context information. This has resulted in continuous improvements of data presentation on the Eurostat website, which now offers – in addition to the detailed data base and extraction tools – a series of most popular tables, key figures and domains in focus (i.e., "Sustainable Development" and "GDP and beyond") as well as *Statistics Explained*.

*Statistics Explained* is a special website that uses Wiki techniques and presents a number of articles that make up an encyclopedia of European statistics for everyone, supplemented by a statistical glossary clarifying all terms used and by numerous links to further information. It is a portal for occasional as well as for regular users. From January to July in 2015, *Statistics Explained* had over three million page views with a high current number of visits and rising. For specific domains, such as "quality of life", "you in the EU", "young Europeans" users can go directly from Eurostat's main page to "Infographics", which presents statistics in a user-friendly way. In addition, mobile apps can be downloaded (for instance concerning the EU economy and country profiles).

In addition, Eurostat has invested in organising training sessions for Commission policy departments and provides practical training on how to access, interpret and use statistics to groups of visitors and journalists. The European Statistical Training Programme (ESTP), coordinated by Eurostat, helps officials and employees of the ESS improve statistical literacy, taking into account the different levels of statistical knowledge and working experiences. It includes courses in Official Statistics, IT applications, Research and Development and Statistical Management. As a result of another programme for Training and Education in Official Statistics within existing Master programmes at European universities, the European Statistical System Committee (ESSC) awarded – in May 2015

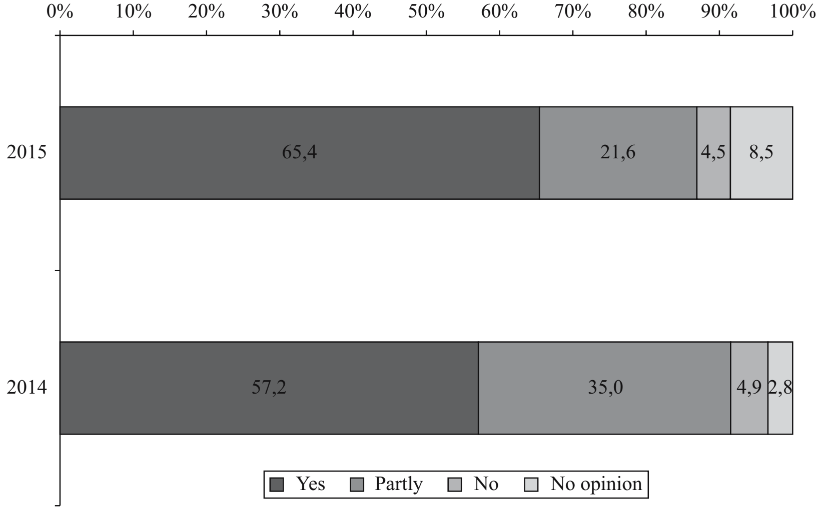


*Fig. 1. Assessment of the presentation of the statistics on the website, in %. (Are European statistics presented in an easy-to-understand way?)*

for the first time – twelve university master programmes with the *European Master in Official Statistics* (*EMOS*) label.

The 2015 User Satisfaction Survey showed an important rise in the percentage of users satisfied with the presentation of Eurostat data (Figure 1).

These are satisfactory results, but there is still room for continued efforts to improve clarity and accessibility of European statistics. In addition, the data environment is constantly evolving with new openings such as open data, big data, and use of geographic information.

To respond to these challenges, in 2014 Eurostat and its partners in the ESS agreed on a Vision2020 priority area for action on "Identifying user needs and cooperation with stakeholders" (European Statistical System Committee 2014). In 2015, a statistical project, DIGICOM, was launched among the ESS partners to support their actions on user analysis and on creating a new dissemination and communication strategy using innovative technologies.

Users of European statistics can also reach through to the European official statistics providers within the framework of the biannual meetings of the European Statistical Advisory Committee (ESAC). This Committee has set up a "classification" of users, thereby allowing producers of European statistics to focus on different sets of users and different contexts (Vichi et al. 2015).

The above illustrates Eurostat's efforts aimed at satisfying the needs of its users. However, before any further steps in this direction are taken it might be wise to first thoroughly assess statistical literacy of the current groups of users to better target their needs – now and in the future. This assessment should also include user groups represented by official statisticians and by small and medium enterprises (SMEs) – as discussed by Deutsch, Forbes and Keegan, as well as Coleman in the special section of this issue.

## References

Bautier, P., C. Laevaert, and B. Le Goff. 2014. *Tracking Users for a Targeted Dissemination*, Paper presented at the Conference of European Statistics Stakeholders in Rome, 2015.

European Statistical System Committee. 2014. *Vision2020*. Available at: http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020 (accessed September 2016).

Vichi, M., M. João Valente Rosa, and F. Ruane. 2015. *The Users of Statistics and Their Role in the European Society*, ESAC document 2015/1175; http://ec.europa.eu/eurostat/web/european-statistical-advisory-committee-esac/activities/esac-documents (accessed September 2016).

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 917–945, http://dx.doi.org/10.1515/JOS-2016-0049

# Measuring and Detecting Errors in Occupational Coding: an Analysis of SHARE Data

*Michele Belloni*[1], *Agar Brugiavini*[1], *Elena Meschi*[1], *and Kea Tijdens*[2]

This article studies coding errors in occupational data, as the quality of this data is important but often neglected. In particular, we recoded open-ended questions on occupation for last and current job in the Dutch sample of the "Survey of Health, Ageing and Retirement in Europe" (SHARE) using a high-quality software program for ex-post coding (CASCOT software). Taking CASCOT coding as our benchmark, our results suggest that the incidence of coding errors in SHARE is high, even when the comparison is made at the level of one-digit occupational codes (28% for last job and 30% for current job). This finding highlights the complexity of occupational coding and suggests that processing errors due to miscoding should be taken into account when undertaking statistical analyses or writing econometric models. Our analysis suggests strategies to alleviate such coding errors, and we propose a set of equations that can predict error. These equations may complement coding software and improve the quality of occupational coding.

*Key words:* ISCO; coding software; coding error; cognitive functioning; education.

## 1. Introduction

Knowledge concerning the occupations of individuals is important in many fields of the social sciences. For example, in economics, sociology, and other disciplines, occupation is

[1] Department of Economics, Ca' Foscari University of Venice. S. Giobbe, Cannaregio, 873 30121 Venice, Italy. Emails: michele.belloni@unive.it, brugiavi@unive.it, and elena.meschi@unive.it
[2] University of Amsterdam/Amsterdam Institute for Advanced Labour Studies (AIAS), Nieuwe Prinsengracht 130, 1018 VZ Amsterdam, Netherlands. Email: k.g.tijdens@uva.nl
**Acknowledgments:** This article builds on research carried out for the DASISH project (Data Service Infrastructure for Social Sciences and Humanities, funded by EU-FP7, Contract no. 283646) and the SHARE project (Survey of Health, Ageing and Retirement in Europe). The authors are grateful to researchers at CentERdata in the Netherlands for their efforts in coding Dutch job titles. The authors would like to thank Eric Balster, Peter Elias, Eric Harrison, Maurice Martens, Sue Westerman, and all participants of the "CASCOT: Occupational Coding in Multi-national Surveys" Workshop in Venice (10-11 April 2014) for helpful suggestions on earlier draft of this work. The research leading to these results has received support under the European Commission's 7th Framework Programme (FP7/2013-2017) under grant agreement n°312691, InGRID – Inclusive Growth Research Infrastructure Diffusion. The last author would like to acknowledge the contribution of WEBDATANET [COST Action IS1004]. This paper uses data from SHARE wave 1 release 2.5.0, as of 23 August 2011. The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life), through the 6th Framework Programme (projects SHARE-I3, RII-CT-2006-062193, COMPARE, CIT5-CT-2005-028857, and SHARELIFE, CIT4-CT2006-028812) and through the 7th Framework Programme (SHARE-PREP, N° 211909, SHARE-LEAP, N° 227822 and SHARE M4, N° 261982). Michele Belloni is also affiliated with CeRP – Collegio Carlo Alberto. Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21 AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions).

often considered – either in itself or as part of an index – as a proxy for socioeconomic status. In sociology and labour economics, occupation is a key variable in a wide range of studies, such as the 'task approach' to labour markets and job polarisation (e.g., Autor 2013; Autor et al. 2006; Goos and Manning 2007); the definition of skill mismatch and overeducation (for an extensive overview of this literature, cf. e.g., Hartog 2000; Leuven and Oosterbeek 2011); the analysis of the effect of occupation on health status (e.g., Fletcher et al. 2011; Ravesteijn et al. 2013); the dynamics of occupational mobility (e.g., Moscarini and Thomsson 2007; Perales 2014); and the analysis of socioeconomic status (e.g., Rose and Harrison 2007).

The quality of occupational data is rarely discussed in this literature, despite the fact that the measurement of occupation in social surveys is a rather complex issue. Handbooks by international institutions such as the International Labour Organization (ILO) detail how to ask about occupation in labour force surveys and censuses (e.g., ILO 2010). However, empirical research on best practices and miscoding is scarce. The difficulty of providing researchers with an accurate measure of occupation concerns, first, the choice of questions to include in the questionnaire, second, the training of interviewers and, third, the conversion of job titles and descriptions that are often recorded in open text fields into occupational codes.

The statistical agencies of 150 countries associated with the ILO have adopted the International Standard Classification of Occupations (ISCO) to normalise the measurement of occupations. The first classification dates back to 1958, with updates in 1968, 1988, and recently in 2008. The Commission of the European Communities (2009) has adopted ISCO-08 as its occupational classification standard, and the European statistical agency Eurostat has made efforts to support European countries to develop coding indexes for occupation data collected through their own labour force and similar surveys. In 2012, almost half of the 150 countries associated with the ILO used the ISCO standard, while the other half either did not classify occupations or maintained their own classification standard (UN 2014).

The ILO provides a classification standard as well as task descriptions for all four-digit occupational units in ISCO (ILO 2014). The task descriptions also provide a coding index, but only in English. Therefore, the coding of occupations becomes particularly challenging in international surveys – such as the "Survey of Health, Ageing and Retirement in Europe" (SHARE) and the "European Social Survey" (ESS), where the occupational codes should be fully comparable across countries – because it is sometimes problematic for countries to map their specific occupations and job titles onto the international ISCO categories.

Researchers are often not aware of the complex preparatory work behind occupational coding, and often consider the published variable of 'occupation' as free of error. This is not the case if processing errors arise during the coding of the variable. Processing errors are one source of nonsampling errors that contribute to total survey errors (see Biemer and Lyberg 2003). Processing errors arise during the data-processing stage and comprise editing errors, coding errors, data-entry errors, and programming errors. For example, in coding answers to open-ended questions related to economic characteristics – such as occupation – coders may deviate from the procedures laid out in coding manuals and therefore assign wrong codes to these characteristics.

Elias (1997) highlighted possible sources of error in occupational data, surveying the few existing studies that evaluated the quality of occupational data through recoding. He found that agreement rates (i.e., the percentage of verbatim responses coded equally after recoding) increased with higher levels of aggregation, thus at one or two digits. At three digits, agreement rates in excess of 75% were hard to obtain. Ellison (2014) pointed out that agreement rates tend to be higher for mother's, father's and last jobs than for an individual's current job. The intuitive explanation for these results is that individuals tend to give too many details about their current job because they think that their job is complex and thus do not provide a simple description, while this occurs to a lesser extent for parents' and last job.

In this article, we will first demonstrate that occupational coding is in fact susceptible to processing errors. In addition, we will test whether such processing errors are random or correlated to some specific individual or job-related characteristics. Finally, we will present our recommendations for reducing this type of error and will propose a novel predictive equation for coding error, given some individual and job-related characteristics, which may be particularly useful if used during interviews.

To attain our aims, we conducted the following empirical analysis. First, we recoded the verbatim response to the open-ended questions on current and last occupation for the Dutch sample of SHARE data using a well-known and high-quality software program for ex-post coding called CASCOT. Second, we compared SHARE data as originally published with recoded occupational variables. Finally, we analysed which individual and job-related characteristics (such as age, gender, education, or industry) were associated with the probability of coding error. The article proceeds as follows: Section 2 discusses the alternative methods used to collect and code information on individuals' occupations and describes the main features of CASCOT. In Section 3, we describe our empirical study and present the data and the methodology adopted. The results of our analysis are presented and discussed in Section 4. Finally, Section 5 presents our conclusions and suggests some directions for further research.

## 2. Coding Occupations in Survey Data: Alternative Methods

Most occupational information in survey data is obtained from direct questions addressed to respondents. The question about occupation is usually asked in an open text field (e.g.,: 'What occupation did you perform in your principal job during the week of . . . to . . . ?'; for an overview of survey questions see Tijdens 2014b; for question design see Jackle 2008 and DESA 2010). Open-ended questions allow the classification of occupations at a detailed level of disaggregation, but the text fields require coding afterwards ('office coding'). Promising attempts to code job titles during CAPI interviews are currently being made using a look-up table or coding index. One notable example of these new coding methods is the semantic text-string matching algorithm (the 'Jobcoder') developed by CentERdata (http://www.centerdata.nl/) and used for the first time in SHARE Wave 6. The occupational coding process in this wave of SHARE followed a two-step approach. In the first step, verbatim responses to the open-ended question on occupation were stored for future possible checks. In the second step, the verbatim responses were forwarded to the 'Jobcoder', which searched its job titles database and checked whether there was an entry that corresponded precisely. If such an entry was found, the software coded the text immediately; otherwise, the interviewer was given the opportunity to ask the interviewee for a more precise job description.

In the more standard case of 'office coding', the classification of occupational information is achieved after the interview through a coding process that can be done manually or semi-automatically using a computerised coding system ('computer-assisted coding') or by a combination of both. Manual coding requires a lot of training for coders and coder supervisors (see Hoffmann et al. 1995; Ganzeboom 2008). Semiautomatic coding tools are becoming increasingly reliable instruments that use semantic matching with previously coded occupations. Machine-learning algorithms also appear to be a promising recent development, requiring a substantial number of manually coded occupations to be used as training data for the automatic classification (Bethmann et al. 2014; Cheeseman Day 2014).

CASCOT is a software tool for coding text automatically or manually (http://www2.warwick.ac.uk/fac/soc/ier/software/cascot/). It was developed at the Institute for Employment Research (IER) in 1993 and since then has been continuously updated and used by over 100 organisations in the UK and abroad. The software developed at IER is able to code job titles in the UK into various editions of the Standard Occupational Classification (SOC) and International Standard Classification of Occupations (ISCO). CASCOT software is coupled with an editor which allows users to modify internal coding rules and permits the software to use alternative occupational classification structures. High-quality coding requires high-quality job descriptions. The recorded text should ideally contain sufficient information to distinguish it from alternative text descriptions which may be coded to other categories within the classification, but it should not contain superfluous words. The recorded text should also be free of typing errors if possible. This ideal will not always be achieved, but CASCOT has been designed to perform a complicated analysis of the words in the text, understand common typing errors and compare these words to those in the classification, ultimately providing a list of recommended codes. If the input text is not sufficiently distinctive, the topmost recommendation may not necessarily be the correct code. When CASCOT assigns a code to a piece of text, it also calculates a score from 1 to 100, which represents the degree of certainty that the given code is the correct one. When CASCOT encounters a word or phrase that is descriptive of an occupation but lacks sufficient information to distinguish it from other categories (i.e., without any further qualifying terms), CASCOT will attempt to suggest a code but the score will be limited to below 40 to indicate the uncertainty associated with the suggestion (e.g., cases such as 'Teacher' or 'Engineer').

The user may run CASCOT in three different modes: fully automatic, semiautomatic, or manual/one-by-one. The fully automatic mode does not require any human intervention: once a list of job descriptions is provided in the software, a series of corresponding codes plus the associated scores is produced. If the software considers the quality of a given job description too low to be able to attribute any reasonable code, it reports 'no conclusion' for that specific text. The semiautomatic mode works by setting a minimum score: in all cases in which CASCOT attributes a score greater than the minimum value, it codes the text automatically; otherwise it asks for human intervention. In these cases, the operator is asked to choose manually from a list of recommendations. The operator's decision may be supported by ancillary variables if they are available in the data: a pop-up window opens in CASCOT and shows, for example, the industry in which the individual is/was working. In manual mode, CASCOT provides a list of recommended codes with corresponding scores for each job description, and leaves the final choice of the best code to the operator.

CASCOT output was compared to a selection of high-quality manually coded data, with the overall results showing that 80% of the records receive a score greater than 40 and, of these, 80% are matched to manually coded data. When using CASCOT, one can expect this level of performance with similar data, but the performance depends on the quality of the data input (for more information about the software, see Elias et al. 1993; Jones and Elias 2004).

Statistics Netherlands (CBS) has developed a Dutch version of CASCOT, building on the English version. Since 2012, this software (henceforth CASCOT-NL) has been used in the Netherlands to code job titles in the most relevant social surveys, including the Dutch Labour Force Survey. CASCOT-NL is suitable for implementation in CAPI, CATI, and CAWI modes.

In this study, we use a version of CASCOT-NL that was used by CBS to classify job descriptions given in the Dutch Labour Force Survey into four-digit ISCO-08 codes. A noticeable difference between CASCOT-UK and CASCOT-NL (the 'classification file ISCO v1.1') is that the latter includes a special category for vague responses called '99'. Very often, a certainty score equal to 99 is given to these cases originally coded '99'. This is because – once tagged in this way – these especially problematic answers go through subsequent coding steps. These steps exploit information from additional variables such as sector of work, the individual's educational attainment and tasks and duties involved in the job. Finally, the most difficult cases are manually coded by a team of experts (see CBS 2012 and Westerman 2014 for further details on CBS coding procedures).

## 3. Data and Empirical Strategy

Our analysis is based on SHARE data. SHARE is a cross-national longitudinal survey on health, socioeconomic status and social and family networks representative of the population aged 50 and over. Four waves of SHARE are currently available. We focus on the first wave of data (collected in 2004–2005) because this is the only one in which information on occupation was gathered using an open-ended question (in the subsequent waves 2 to 5, the occupation question uses a tick list of ten occupational titles). In particular, in SHARE Wave 1, respondents were asked the following question: "What [is/was] your [main/last] job called? Please give the exact name or title." This question was directed at both employed/self-employed and retired/unemployed individuals (the latter conditional on having worked earlier in life). Note that SHARE also collects information on respondents' second job, parents' job and former partner's job. Parents' jobs are intrinsically more difficult to code than respondents' jobs because the former may have been excluded from recent job classifications. There are very few observations for respondents' second job and former partner's job. Thus we excluded these additional variables from our analysis.

SHARE country teams manually coded the text strings on respondents' job titles using ISCO-88 (COM) codes – the International Standard Classification of Occupations used at that time. Each country team hired and trained coders independently. Coders were asked to follow a protocol providing them with guidelines on how to code 'critical' jobs (e.g., managers in agriculture or teachers). These guidelines were partly common to all countries and partly language specific. SHARE coders also made use of ancillary information on training and qualifications needed for the job and on the industry the respondent was working in based on the question: "What kind of business, industry or services do you work in (that is, what do they make or do at the place where you work)?" SHARE coders were asked to code job descriptions at the maximum possible level of detail, that is, at four-digit (or Unit group) ISCO-88 level. It was also suggested that they code vague responses by means of trailing zeros: this means that if they were unsure whether a given job description could be attributable to a given Unit group, they should attribute it to either a Minor (i.e., three digits), Sub-major (two digits) or Major (one digit) group. The ISCO-88 codes generated for two variables – one for current main job (*ep016_*) and one for last job (*ep052_*) – were then published (for further details, see MEA 2013, 29). The first wave of SHARE covered eleven European countries and Israel. Our recoding exercise only uses the Dutch sample of this wave because CASCOT is currently available in two languages – English and Dutch – and the English language is not present in SHARE data.

We recoded job descriptions using CASCOT-NL in its semiautomatic mode by setting a minimum score of 70 and with the assistance of an expert coder who was a Dutch native speaker and who has been involved in occupational coding and occupational databases for many years. As mentioned above, in all cases in which CASCOT-NL attributed a score greater than 70, it coded the text automatically. The expert coder manually coded all the residual cases. Consistent with what is done in SHARE, the operator coded vague responses by means of trailing zeros. The manual recoding was done twice: with and without ancillary information. The use of ancillary variables increased the comparability between the SHARE and CASCOT-NL coding. Moreover, the operator made use of the same ancillary variables (on training and qualifications needed for the job and on the industry the respondent was working in) used by SHARE coders. In order to avoid the 'anchoring effect' – that is, the tendency of human coders to select the code already in front of them (see Cheeseman Day 2014) – the expert coder used a recent CBS coding index (see http://www.cbs.nl/nl-NL/menu/methoden/classificaties/overzicht/sbc/default.htm) including 4,705 job titles rather than the list of codes recommended by the CASCOT-NL classification file ISCO v1.1. We believe that the combination of a high-quality software program (which automatically coded a high proportion of cases at the four-digit level, see below), an expert coder, the use of ancillary information, and the use of an extensive external job titles list ensured a high level of coding and provided better coding than manual SHARE coding. In the following, we will therefore consider the CASCOT-NL coding (the version exploiting ancillary variables) as our benchmark.

Tables 1a and 1b show the number of recoded cases available for our statistical analysis: 2,790 observations, of which 1,773 concern last job (Table 1a) and 1,017 current job (Table 1b). The higher frequency for last job in comparison with current job primarily reflects the distribution of respondents by work status in the first wave of SHARE. Two points are worth mentioning with respect to Tables 1a and 1b: first, the number of cases automatically coded (scoring above 70) at four-digit level is high (40%, i.e., 708 out of the 1,773 total observations for last job; 55%, i.e., 557 out of the 1,017 observations for

Table 1a. *Output of CASCOT-NL recoding at different number of digits by score level and use of ancillary variables – Last job: frequencies, and row percentages (in italics).*

| | | 4 digit | 3 digit | 2 digit | 1 digit | Total |
|---|---|---|---|---|---|---|
| Score above 70* | | 708 | 108 | 142 | 71 | 1029 |
| | | *69* | *10* | *14* | *7* | *100* |
| Score below 70** | *No ancillary* | 336 | 146 | 115 | 73 | 670 |
| | | *50* | *22* | *17* | *11* | *100* |
| | *With ancillary* | 596 | 98 | 23 | 27 | 744 |
| | | *80* | *13* | *3* | *4* | *100* |
| Total | | | | | | **1,773** |

*automatically coded; **manually coded.

current job); second, making use of ancillary information dramatically increases the number of digits at which the observations are coded. For example, for last job, the percentage of cases coded at four-digit level among those which scored below 70 increased from 50 to 80% when using ancillary variables.

The main issue arising when comparing codes from SHARE and CASCOT-NL is the lack of homogeneity in the classification structure. SHARE Netherlands coded job descriptions at three-digit ISCO-88 level (note that all other countries coded jobs at ISCO-88 four-digit level, see above), while CASCOT-NL, as described above, coded to ISCO-08 four-digit level. We therefore homogenised the two sets of codes as follows. First, we converted CASCOT-NL codes from ISCO-08 into ISCO-88 using an official correspondence table (ILO 2014). Unfortunately, according to this table, there is a 'many-to-one' correspondence between ISCO-88 and ISCO-08, that is, multiple ISCO-88 codes are associated with the same four-digit ISCO-08 code. In our data, this occurs for about 20% of the sample. In these cases, we associated multiple ISCO-88 codes with the same job description. Considering the issue of nonunivocal correspondence between different versions of ISCO, we decided that a job description would have a different code if the ISCO-88 code attributed by SHARE coders is not equal to *any* of the ISCO-88 codes resulting from the conversion of the CASCOT-NL output into ISCO-88. Otherwise, the

Table 1b. *Output of CASCOT-NL recoding at different number of digits by score level and use of ancillary variables – Current job: frequencies, and row percentages (in italics).*

| | | 4 digit | 3 digit | 2 digit | 1 digit | Total |
|---|---|---|---|---|---|---|
| Score above 70* | | 557 | 87 | 0 | 7 | 651 |
| | | *86* | *13* | *0* | *1* | *100* |
| Score below 70** | *No ancillary* | 188 | 104 | 37 | 37 | 366 |
| | | *51* | *28* | *10* | *10* | *100* |
| | *With ancillary* | 241 | 53 | 42 | 30 | 366 |
| | | *66* | *14* | *11* | *8* | *100* |
| Total | | | | | | **1,017** |

*automatically coded; **manually coded.

job description has the same code. Second, we only considered three digits. In summary, we compared codes from SHARE and CASCOT-NL in terms of three-digit ISCO-88 codes.

## 4. Results

### 4.1. Descriptive Statistics

Figures 1a and 1b show the distribution of occupations by ISCO-88 Major groups according to both SHARE and CASCOT-NL coding, and for last and current job respectively. Given the fact that multiple codes are sometimes associated with the same individual in our recoding exercise due to the lack of one-to-one correspondence between ISCO-08 and ISCO-88, we used weighting to construct these figures. In particular, when *n* codes are associated with the same individual, we attributed a weight equal to *1/n* to each of them.

The figures reveal sizable differences between ISCO distributions of current and last job. The share of professionals and associate professionals (ISCO Major groups 2 and 3) is much higher for current job than for last job, whereas the opposite occurs for lower-skilled occupations. This fact may reflect changes in occupational structure over time, possibly due to technological change or international trade, as last job may often refer to occupations started early in an individual's working career. There is in fact extensive literature showing that technological progress and increased competition from low-wage countries have changed labour demand in favour of more skilled occupations (e.g., Autor et al. 2003; Feenstra and Hanson 1996). In addition, these differences in the distribution of occupations may also be due to selective retirement: manual workers may retire earlier from the labour force than nonmanual workers and therefore may be overrepresented in

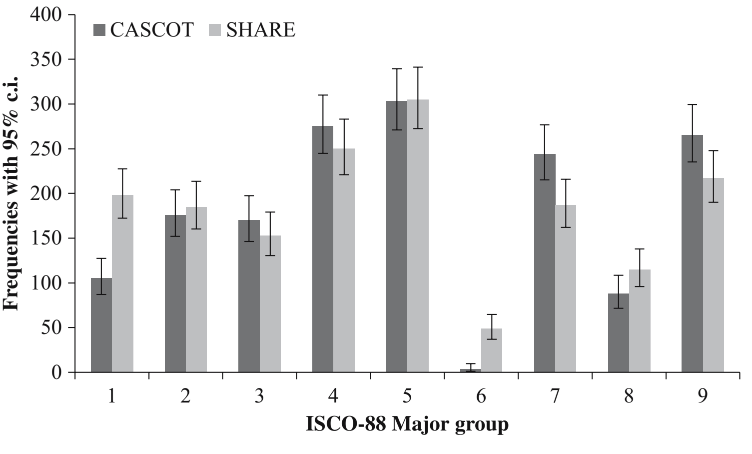


*Fig. 1a. Distribution of occupation by ISCO-88 Major group, CASCOT-NL and SHARE coding – Last job (frequencies with 95% confidence intervals). Legend: 1 = Legislators, senior officials and managers, 2 = Professionals, 3 = Technicians and associate professionals, 4 = Clerks, 5 = Service workers and shop and market sales workers, 6 = Skilled agricultural and fishery workers, 7 = Craft and related trades workers, 8 = Plant and machine operators and assemblers, 9 = Elementary occupations.*

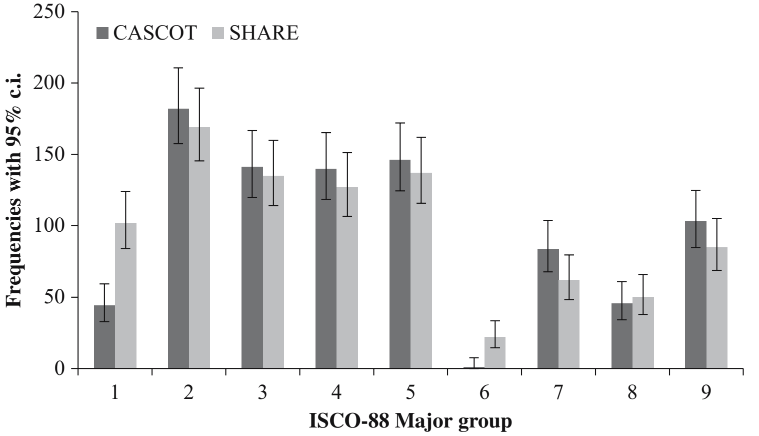


*Fig. 1b. Distribution of occupation by ISCO-88 Major group, CASCOT-NL and SHARE coding – Current job (frequencies with 95% confidence intervals). Legend: 1 = Legislators, senior officials and managers, 2 = Professionals, 3 = Technicians and associate professionals, 4 = Clerks, 5 = Service workers and shop and market sales workers, 6 = Skilled agricultural and fishery workers, 7 = Craft and related trades workers, 8 = Plant and machine operators and assemblers, 9 = Elementary occupations.*

the last job variable. The contrary may occur for professionals, who may remain in the labour market even beyond the standard retirement age. The issue of selective retirement is non-negligible in countries favouring part-time work such as the Netherlands. Finally, note that the number of observations for each Major group is limited; consequently, statistical analyses disaggregated by ISCO groups at 2/3 digits are not presented in this section.

Tables 2a and 2b report frequency and percentage of same and different codes for last and current job respectively. The percentage coded differently (which we call 'disagreement rate' hereafter) appears high even when the comparison is made at the one-digit level (28% for last job and 29% for current job). As expected, such percentages rise with the number of digits at which the comparison is performed. This result is in line with the meta-analysis of the results from occupational recoding studies carried out by Elias (1997) and cited in the introduction. The disagreement rate is slightly higher for current job than for last job: for example, at three-digit level, 47% of texts for current job are coded differently, compared with 43% for last job. A possible explanation for this last finding is related to sample composition: we have seen that the ISCO-88 Major group distribution for current and last job are different for good reasons (Figure 1), and some

*Table 2a. Observations coded equally and differently by CASCOT-NL and SHARE at different number of digits – Last job (frequencies and percentages).*

| | 1 digit | | 2 digit | | 3 digit | |
|---|---|---|---|---|---|---|
| ISCO-88 Code: | *Freq.* | *Percent* | *Freq.* | *Percent* | *Freq.* | *Percent* |
| Same | 1,195 | 72 | 1,086 | 65 | 937 | 56 |
| Different | 464 | 28 | 573 | 35 | 722 | 44 |
| Total | 1,659 | 100 | 1,659 | 100 | 1,659 | 100 |

*Table 2b. Observations coded equally and differently by CASCOT-NL and SHARE at different number of digits – Current job (frequencies and percentages).*

| | 1 digit | | 2 digit | | 3 digit | |
|---|---|---|---|---|---|---|
| ISCO-88 Code: | *Freq.* | *Percent* | *Freq.* | *Percent* | *Freq.* | *Percent* |
| Same | 631 | 71 | 555 | 62 | 465 | 52 |
| Different | 258 | 29 | 334 | 38 | 424 | 48 |
| Total | 889 | 100 | 889 | 100 | 889 | 100 |

ISCO groups may be more subject to coding errors than others (see Table 3). Finally, this finding is consistent with the explanation proposed by Ellison (2014) and mentioned earlier in this article: individuals tend to give too many details about their current job because they think it is complex, while this occurs to a lesser extent for last job.

Table 3 reports disagreement rates for ISCO-88 Major groups, for both last and current job. There is wide variety in the disagreement rate across groups, with groups 1 ("Legislators, senior officials and managers") and 3 ("Technicians and associate professionals") being those with the highest values. The percentage of observations coded differently is also high for the current job variable in group 6 ("Skilled agricultural and fishery workers"). Agricultural workers are known to be difficult to code and some occupations in this category were subject to changes in classification from ISCO-88 to ISCO-08. The high disagreement rate for this category may be due to the fact that the ISCO-88 Unit groups of 1221, "Production and operations department managers in agriculture forestry and fishing", and 1311, "General managers in agriculture, forestry and fishing", were removed from Major group 1 in the ISCO-08 classification. The occupations included within this category were moved to Sub-Major Group 61 and merged with the relevant supervisory groups (UN 2007). Therefore, "General managers in agriculture, hunting, forestry and fishing" were classified as ISCO-88 Unit group 1311, and should not be included within Major group 6.

In addition to disagreement rates, in the following we attempt to quantify the degree of disagreement between the two sets of codes in terms of skill levels (where a hierarchical order and a measure of difference – or 'distance' – among groups in terms of skills can be established). The ILO in fact maps ISCO Major groups into skill levels (Elias 1997; ILO 2012) which can then be mapped onto education levels defined by ISCED-97 (see Table A1 in the Appendix). For example, the difference in skill level between a job in ISCO-88 Major group 9 (Elementary Occupations, skill level 1) and 2 (Professionals, skill level 4) is equal to 3. We first performed the Wilcoxon signed-rank test for paired data (Wilcoxon 1945). The results of this test were very different for last and current job. While for last job the null hypothesis that SHARE and CASCOT-NL coding distributions will be the same was not rejected ($p$-value $= 0.12$), for current job this hypothesis was rejected even at one percent significance level ($p$-value $= 0.0004$). Tables 4a and 4b present the bivariate distributions – SHARE vs CASCOT-NL skill-level groups – for last and current job respectively. The tables show that most of the coding disagreement occurs within similar groups of occupations. Looking at last job, 85% of occupations coded into skill group 1 in SHARE are coded into the same skill group in CASCOT-NL. The percentages of correct

*Table 3. Disagreement rate at different number of digits for ISCO Major groups – last job and current job.*

| | Last job | | | Current job | | |
|---|---|---|---|---|---|---|
| | Disagreement rate (%) | | | Disagreement rate (%) | | |
| ISCO Major group* | 3 digit | 2 digit | 1 digit | 3 digit | 2 digit | 1 digit |
| Legislators, Senior Officials, and Managers | 83 | 71 | 62 | 82 | 70 | 63 |
| Professionals | 37 | 24 | 23 | 31 | 21 | 18 |
| Technicians and Associate Professionals | 62 | 52 | 48 | 65 | 53 | 47 |
| Clerks | 38 | 22 | 20 | 35 | 24 | 20 |
| Service Workers and Shop and Market Sales | 27 | 26 | 20 | 29 | 28 | 20 |
| Skilled Agricultural and Fishery Workers | 45 | 43 | 18 | 91 | 86 | 27 |
| Craft and Related Trades Workers | 37 | 29 | 14 | 56 | 32 | 16 |
| Plant and Machine Operators and Assemblers | 57 | 47 | 43 | 46 | 40 | 26 |
| Elementary Occupations | 29 | 20 | 14 | 45 | 34 | 20 |

Note: The disagreement rate is the percentage of observations coded differently by CASCOT-NL and SHARE; *ISCO-88 Major groups, as coded in SHARE.

*Table 4a. Skill levels bivariate distributions – SHARE vs CASCOT-NL – Last job (%).*

| CASCOT → SHARE ↓ | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| 1 | **85** | 14 | 1 | 0 | 100 |
| 2 | 6 | **83** | 10 | 1 | 100 |
| 3 | 3 | 28 | **52** | 17 | 100 |
| 4 | 0 | 9 | 19 | **72** | 100 |
| Total | 16 | 56 | 17 | 11 | 100 |

coding are around 83% for skill group 2, 52% for skill group 3 and 72% for skill group 4. As suggested by the Wilcoxon signed-rank test, these percentages are lower when considering current job, with the exception of skill level 4. We currently have no explanation for the latter.

In the remainder of the article, we investigate which individual characteristics are more likely to be associated with coding disagreement. To this end, we performed both univariate and multivariate analyses. The tables reporting univariate statistics can be found in the Appendix. In particular, Table A2 presents the disagreement rate according to education level, Table A3 according to gender and Tables A4a and A4b according to industry for last and current job respectively. The figures clearly show that the rates of coding disagreement differ substantially across education and gender, with higher rates for more educated individuals (only for last job) and for males. No clear patterns emerge from the tables on disagreement rates for industry, probably because of the very low number of observations in some groups. In the following subsection, we explore these results in more detail based on a multivariate analysis.

### *4.2. Multivariate Analysis: Predicting Coding Errors*

In this section, we estimate a set of Linear Probability Models (LPM) that can be used to predict coding errors. They can also provide information about which ISCO groups are more difficult to code. An LPM is a multiple linear regression model with a binary dependent variable (Wooldridge 2010). As a robustness check, we also estimated the same equations using nonlinear methods and the results were almost the same. The dependent variable in these models allows for the possibility of multiple correspondences in the ISCO-08 to ISCO-88 conversion tables. In other words, in our models the dependent variable is a dummy variable equal to 1 if the three-digit ISCO-88 code provided by SHARE is not equal to any of the three-digit ISCO-88 codes resulting from the conversion of the ISCO-08 CASCOT-NL code into ISCO-88; otherwise, the dependent variable is equal to 0. We estimated weighted regressions to account for the multiple correspondences in the ISCO-08 to ISCO-88 conversion tables (where each observation is given a weight that is inversely related to the number of correspondences). The results for the unweighted regressions were virtually unchanged. Moreover, we considered two alternative dependent variables, namely a dummy for being coded differently at one- or two-digit ISCO level. Again, the results of these regressions were similar to those reported in the paper. All these additional results are available from the authors upon request.

*Table 4b. Skill levels bivariate distributions – SHARE vs CASCOT-NL – Current job (%).*

| CASCOT → <br> SHARE ↓ | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| 1 | **80** | 16 | 2 | 2 | 100 |
| 2 | 4 | **82** | 13 | 1 | 100 |
| 3 | 4 | 18 | **50** | 28 | 100 |
| 4 | 1 | 8 | 14 | **77** | 100 |
| Total | 12 | 47 | 21 | 21 | 100 |

The set of LPM we estimated differ in terms of the set of explanatory variables. We estimated separate models for last and current job. By pooling these two variables, we would have considerably increased the number of observations and perhaps improved the precision of our estimates. However, the descriptive findings outlined earlier suggest that coding disagreements for current and last job have different patterns: our econometric results (see below) clearly confirm that pooling current and last job – assuming that explanatory variables have the same effect on the probability of miscoding for current and last job – would have led to misspecification.

Table 5a reports LPM estimates for the probability of the last job being miscoded at three-digit level. We present six specifications in this table. The first three columns include only individual and job-related characteristics, that is, they do not include any variable that results from the coding process. Models 1–3 can be used during (or before) 'office' coding: if the survey containing the questions on occupation also provides information on the explanatory variables included in the estimated equation, their values can be 'plugged in' and used to predict the likelihood that any attributed code is correct or incorrect.

Specification 1 includes basic individual characteristics present in almost all surveys as explanatory variables, namely gender, educational attainment (four aggregated ISCED-97 groups), whether the individual is self-employed (controlling for self-employment is also important to identify the gender effects, as females are overrepresented within this group of workers), and whether the individual is foreign born. Our results indicate that females are 29% less likely to be miscoded than males. Remarkably, there is a strong positive gradient between education and coding disagreement: relative to individuals with no or primary education, those with an upper and postsecondary degree (ISCED 3-4) have a 14% higher probability of being miscoded; this percentage rises to about 18% for individuals holding a tertiary education degree (ISCED 5-6). Being self-employed translates into about ten percent higher chance of being miscoded. The same holds for being born abroad. With the exception of the dummy variable of ISCED 2, all of the explanatory variables included in this model were significant at least at five percent level. This very basic model with few right-hand side variables is able to explain about twelve percent of the variability in the dependent variable (see the R-squared statistic at the bottom of the table).

Specification 2 includes two additional regressors: age and a cognitive skills index. These individual characteristics (especially the latter) might be particularly important for predicting miscoding when looking at mature (50+) individuals. It might be expected that older individuals and individuals with less cognitive functioning provide poorer job

*Table 5a. Linear Probability Model for the probability of miscoding at ISCO three-digit level – Estimation results for last job.*

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Female | −0.290*** | −0.291*** | −0.234*** | −0.156*** | −0.155*** | −0.075** |
| | (0.024) | (0.024) | (0.030) | (0.031) | (0.031) | (0.031) |
| Lower secondary education (ISCED 2) | 0.047 | 0.049 | 0.017 | 0.001 | −0.012 | −0.040 |
| | (0.030) | (0.032) | (0.035) | (0.034) | (0.033) | (0.030) |
| Higher and postsec. ed. (ISCED 3–4) | 0.141*** | 0.137*** | 0.078* | 0.053 | 0.048 | 0.002 |
| | (0.035) | (0.038) | (0.041) | (0.041) | (0.040) | (0.037) |
| Tertiary education (ISCED 5–6) | 0.185*** | 0.181*** | 0.149*** | 0.107** | 0.105** | 0.060 |
| | (0.041) | (0.044) | (0.050) | (0.053) | (0.051) | (0.048) |
| Self-employed | 0.106** | 0.100** | 0.124** | 0.006 | 0.031 | −0.059 |
| | (0.043) | (0.043) | (0.048) | (0.049) | (0.048) | (0.047) |
| Foreign born | 0.101** | 0.079 | 0.043 | 0.042 | 0.029 | 0.015 |
| | (0.049) | (0.051) | (0.056) | (0.054) | (0.052) | (0.048) |
| Age | | 0.001 | 0.000 | −0.000 | 0.000 | 0.000 |
| | | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Cognitive skill index | | 0.006 | −0.003 | −0.023 | −0.009 | −0.003 |
| | | (0.019) | (0.020) | (0.020) | (0.019) | (0.017) |
| Not elsewhere classified | | | | 0.041 | −0.029 | −0.061 |
| | | | | (0.100) | (0.097) | (0.092) |
| **Additional controls**: | | | | | | |
| *Industry dummy (31 groups)* | *No* | *No* | *Yes* | *Yes* | *Yes* | *Yes* |
| *ISCO one-digit dummy (10 groups)* | *No* | *No* | *No* | *Yes* | *No* | *No* |
| *ISCO two-digit dummy (28 groups)* | *No* | *No* | *No* | *No* | *Yes* | *No* |
| *ISCO three-digit dummy (90 groups)* | *No* | *No* | *No* | *No* | *No* | *Yes* |
| **Ancillary statistics**: | | | | | | |
| *Wald test H0: no joint significance Industry dummy variables (p-value)* | | | 0.0003*** | 0.0047*** | 0.0585* | 0.007*** |
| *Wald test H0: no joint significance ISCO dummy variables (p-value)* | | | | 0.000*** | 0.000*** | 0.000*** |
| Observations | 1,629 | 1,607 | 1,421 | 1,421 | 1,421 | 1,421 |
| R-squared | 0.119 | 0.119 | 0.148 | 0.218 | 0.286 | 0.454 |

Note: Dependent variable: ‘miscoding’ = a dummy variable equal to 1 if the three-digit ISCO-88 code provided by SHARE is not equal to any of the three-digit ISCO-88 codes resulting from the conversion of the ISCO-08 CASCOT-NL code into ISCO-88. Standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$; Reference categories: male, no or primary education (ISCED 0-1), employee, Italian born.

descriptions, which are thus more difficult to code. A 'Cognitive Functioning' module included in SHARE reports the results of simple tests of verbal fluency, such as counting the number of items that can be named in one minute; recalling as many words as possible from a ten-word list; and testing daily life numerical calculations (see e.g., Christelis et al. 2010). Based on these tests, we built an index of cognitive abilities (Leist et al. 2013). This index is given by the average value of the standardised results of these tests. The higher the value of the index, the higher the cognitive abilities. We did not find any significant effect of these two variables on the probability of miscoding last job.

In Specification 3, we additionally controlled for industry by including a set of 31 industry dummy variables in the model. Industry was classified using NACE Codes, Version 4 Rev. 1 1993 (see http://www.top500.de/nace4-e.htm for a description of NACE Version 4 Rev. 1 and MEA 2013, 32–33 for the shorter classification used in SHARE). They jointly affect the probability of coding error, as indicated by the result of the Wald test reported at the bottom of the table ($p$-value $= 0.0003$). Remarkably, even after controlling for industry, the effects of gender, educational attainment, and being self-employed on coding disagreement remained significant, although they were somewhat attenuated. This richer specification is able to explain about 15% of the observed miscoding.

Specifications 4 to 6 add a set of variables to individual and job-related characteristics that result from coding the verbatim response to the open-ended questions on occupation. Specification 4 includes ten ISCO one-digit (Major) groups fixed effects, Specification 5 includes 28 ISCO 2-digits (Sub-major) groups fixed effects, and Specification 6 includes 90 ISCO three-digit (Minor) groups fixed effects. Moreover, all models include a dummy variable for being coded as "Not elsewhere classified" (NEC). This was constructed by looking at the ISCO-88 four-digit codes, as coded by CASCOT-NL software. This NEC dummy was equal to 1 if the ISCO-88 fourth digit was equal to 9, which, according to ILO's guidelines, refers to occupational categories that are not classified into other specific categories within the classification. This variable includes ISCO categories which usually contain many types of clerical jobs. We thus expect NEC jobs to be more likely to be miscoded.

These extended specifications can be used to predict coding errors during CAPI interviews. In addition to proposing a given ISCO code, the coding software (such as the 'Jobcoder', see Section 2) would be able to evaluate the quality of the proposal by determining the probability that it is correct (similarly to the score produced by CASCOT). If this probability is low, the interviewee can be asked for additional information. Another possible use of the predictive equations 4 to 6 is to 'double check' office coding. After an ISCO code has been attributed to the occupation, all of the explanatory variables are in fact available for error prediction.

These specifications – especially Specification 6 – are very demanding in terms of data requirements, and we expect to have limited variability in individual and job-related characteristics once we condition on being coded in a given ISCO group. Nonetheless, the negative coefficient for "female" remained significant at five percent even after controlling for ISCO Minor groups. The same occurred for the industry dummy variables (the $p$-value of the Wald test for no joint significance of the industry dummy variables is almost equal to 0 in Specification 6). This independent source of variation increases the overall explanatory power of our error-predicting equations.

Adding ISCO dummy variables to the model dramatically improves the model fit: the R-squared in fact increases from 15% (Specification 3; no information on ISCO codes) to 22% (Specification 4) and progressively increases further with the number of ISCO digits, up to about 45% (Specification 6). The $p$-value of the Wald test for no joint significance of the ISCO group dummy variables is always equal to 0.

As outlined at the beginning of this section, the estimated equations can also provide information about which ISCO groups are more likely to be miscoded. Figures 2 and 3 present the predictions of models 4 and 5 respectively for last job. Figure 2 shows that ISCO Major groups 1, 3, and 8 are the most miscoded groups. Prediction uncertainty is limited at ISCO one-digit level, with the exception of group 6. These predictions can be compared with the disagreement rates reported in Table 3. In some cases, remarkable differences emerge. For example, the disagreement rate of ISCO Major group 9 ("Elementary occupations") is equal to 29% in Table 3, whereas it is much higher (the point estimate being around 40%) in Figure 2. This difference is due to composition effects – mainly related to industry – which are accounted for in Equation (4). Figure 3 highlights that ISCO groups 11 ("Legislators and senior officials"), 12 ("Corporate managers"), 33 ("Teaching associate professionals"), 82 ("Machine operators and assemblers") and – with higher uncertainly – groups 62 ("Subsistence agricultural and fishery workers") and 81 ("Stationary-plant and related operators") are the ISCO Sub-major groups most subject to coding error. We do not present predictions for ISCO Minor groups from Specification 6 since they were too imprecise to be reliable out of sample.

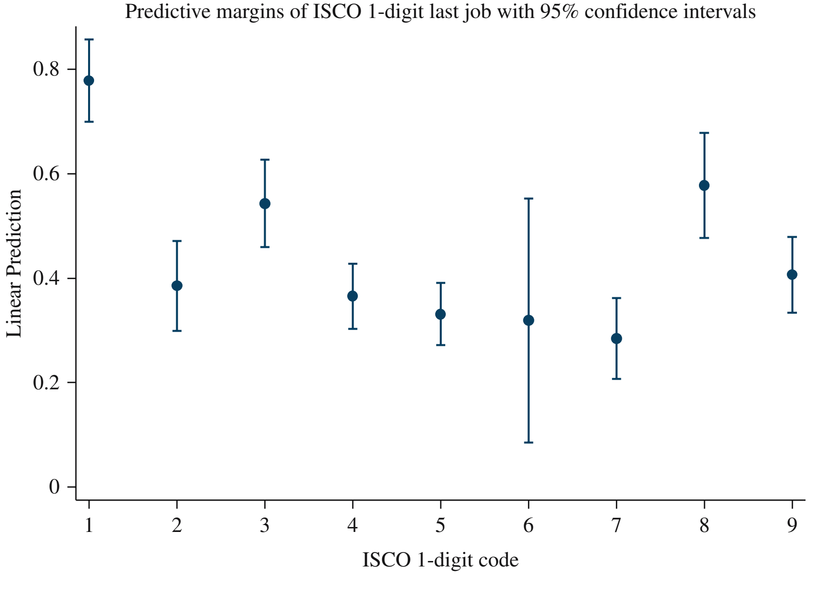


*Fig. 2. Predicted probability of coding error with 95% confidence intervals for ISCO one-digit level – Last job. Legend: 1 = Legislators, senior officials and managers, 2 = Professionals, 3 = Technicians and associate professional, 4 = Clerks, 5 = Service workers and shop and market sales workers, 6 = Skilled agricultural and fishery workers, 7 = Craft and related trades workers, 8 = Plant and machine operators and assemblers, 9 = Elementary occupations. Note: Predictions from Specification 4, Table 5a.*

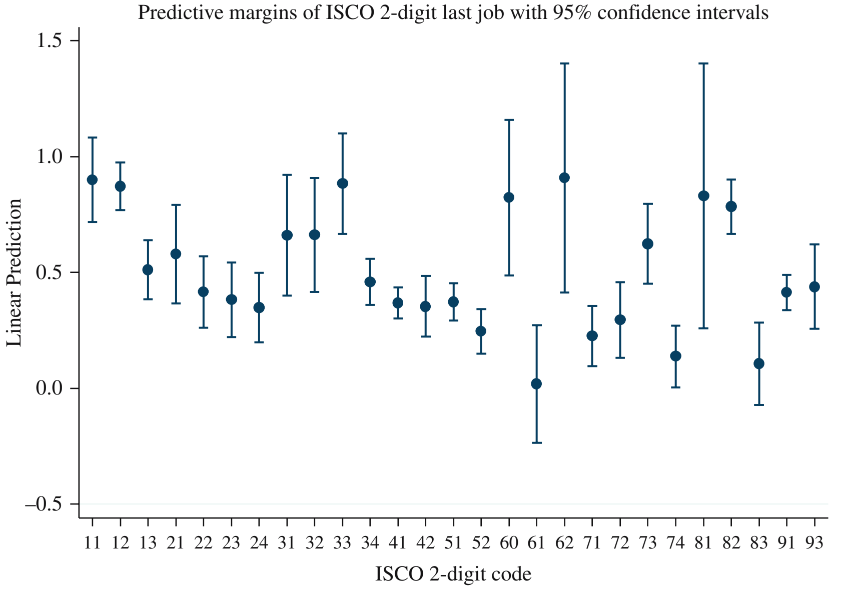


*Fig. 3. Predicted probability of coding error with 95% confidence intervals for ISCO two-digit level – Last job. Note: Predictions from Specification 5, Table 5a.*

Table 5b reports LPM estimates for the probability of the current job being miscoded at ISCO three-digit level. To facilitate comparability, we report the same six specifications presented in Table 5a. The results for current job are very different from those obtained for last job. First, there is no education miscoding gradient for the current job variable. Second, the cognitive skills index has a sizable and significant effect on miscoding, even when ISCO Minor group is controlled for. According to Specification 6, one standard deviation (0.67) increase in this variable (corresponding roughly to a change from its sample median to its 90$^{th}$ percentile) determines a reduction in the probability of miscoding equivalent to 5.7% ($=-0.086*0.67$). Counterintuitively, age has a negative sign, but its effect is quantitatively very small and disappears once the ISCO two-digit level is included in the model (Specification 5). Gender, industry and ISCO groups maintain their strong explanatory power (see the results of corresponding Wald tests at the bottom of the table for the last two groups of variables).

Figures 4 and 5 present coding error predictions for current job using model specifications 4 and 5 respectively. Figure 4 highlights that ISCO Major groups 1, 3, and 6 are the most miscoded groups. The most relevant difference with respect to last job concerns Major group 6: although not precisely estimated, the point estimate of the predicted error is about 90% for current job (it is about 30% for last job; this difference is statistically significant at five percent level). We provided an explanation for the high value of group 6 miscoding for current job in the previous section. Error prediction for group 8 is much higher for last job (point estimate, 0.58) than for current job (0.35). Figure 5 shows that the predicted probabilities of coding error for current job are much higher than for last job for the following ISCO Sub-major groups: 34 (“Other associate professionals”), 61 (“Market-oriented skilled agricultural

*Table 5b. Linear Probability Model for the probability of miscoding at ISCO three-digit level – Estimation results for current job.*

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Female | −0.267*** | −0.273*** | −0.220*** | −0.179*** | −0.161*** | −0.073* |
| | (0.033) | (0.034) | (0.042) | (0.041) | (0.041) | (0.040) |
| Lower secondary education (ISCED 2) | −0.005 | 0.019 | 0.021 | 0.012 | 0.029 | −0.013 |
| | (0.068) | (0.070) | (0.075) | (0.071) | (0.069) | (0.066) |
| Upper and postsec. ed. (ISCED 3–4) | −0.000 | 0.042 | 0.053 | −0.019 | −0.027 | −0.079 |
| | (0.069) | (0.073) | (0.079) | (0.076) | (0.075) | (0.071) |
| Tertiary education (ISCED 5–6) | 0.025 | 0.078 | 0.105 | −0.017 | −0.015 | −0.072 |
| | (0.070) | (0.075) | (0.083) | (0.084) | (0.082) | (0.078) |
| Self-employed | 0.048 | 0.063 | 0.088 | 0.024 | −0.007 | −0.065 |
| | (0.048) | (0.049) | (0.055) | (0.054) | (0.053) | (0.055) |
| Foreign born | 0.043 | −0.012 | 0.044 | 0.085 | 0.058 | 0.044 |
| | (0.068) | (0.070) | (0.076) | (0.072) | (0.070) | (0.066) |
| Age | | −0.007** | −0.006* | −0.005* | −0.003 | −0.003 |
| | | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| Cognitive skill index | | −0.068** | −0.089*** | −0.097*** | −0.080*** | −0.086*** |
| | | (0.027) | (0.029) | (0.027) | (0.027) | (0.025) |
| Not elsewhere classified | | | | 0.146 | 0.134 | 0.102 |
| | | | | (0.102) | (0.099) | (0.098) |
| **Additional controls**: | | | | | | |
| *Industry dummy (31 groups)* | *No* | *No* | *Yes* | *Yes* | *Yes* | *Yes* |
| *ISCO one-digit dummy (10 groups)* | *No* | *No* | *No* | *Yes* | *No* | *No* |
| *ISCO two-digit dummy (28 groups)* | *No* | *No* | *No* | *No* | *Yes* | *No* |
| *ISCO three-digit dummy (90 groups)* | *No* | *No* | *No* | *No* | *No* | *Yes* |
| **Ancillary statistics**: | | | | | | |
| *Wald test H0: no joint significance industry dummy variables (p-value)* | | | 0.0036*** | 0.1452 | 0.0196** | 0.0027*** |
| *Wald test H0: no joint significance ISCO dummy variables (p-value)* | | | | 0.000*** | 0.000*** | 0.000*** |
| Observations | 882 | 850 | 747 | 747 | 747 | 747 |
| R-squared | 0.074 | 0.085 | 0.155 | 0.265 | 0.335 | 0.513 |

Note: Dependent variable: 'miscoding' = a dummy variable equal to 1 if the three-digit ISCO-88 code provided by SHARE is not equal to any of the three-digit ISCO-88 codes resulting from the conversion of the ISCO-08 CASCOT-NL code into ISCO-88. Standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Reference categories: male, no or primary education (ISCED 0-1), employee, Italian born.

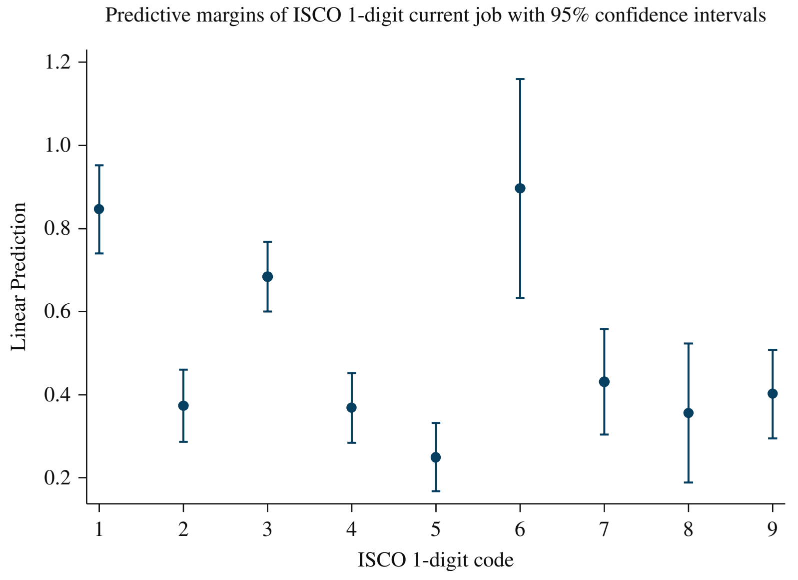


*Fig. 4. Predicted probability of coding error with 95% confidence intervals for ISCO one-digit level – Current job. Legend: 1 = Legislators, senior officials and managers, 2 = Professionals, 3 = Technicians and associate professionals, 4 = Clerks, 5 = Service workers and shop and market sales workers, 6 = Skilled agricultural and fishery workers, 7 = craft and related trades workers, 8 = plant and machine operators and assemblers, 9 = elementary occupations. Note: Predictions from Specification 4, Table 5b.*

and fishery workers"), 71 ("Extraction and building trades workers"), and 74 ("Other craft and related trades workers"). In contrast, and for good reasons, they are lower for ISCO groups 32 ("Life science and health associate professionals"), 42 ("Customer services clerks"), 52 ("Models, salespersons and demonstrators"), and 83 ("Drivers and

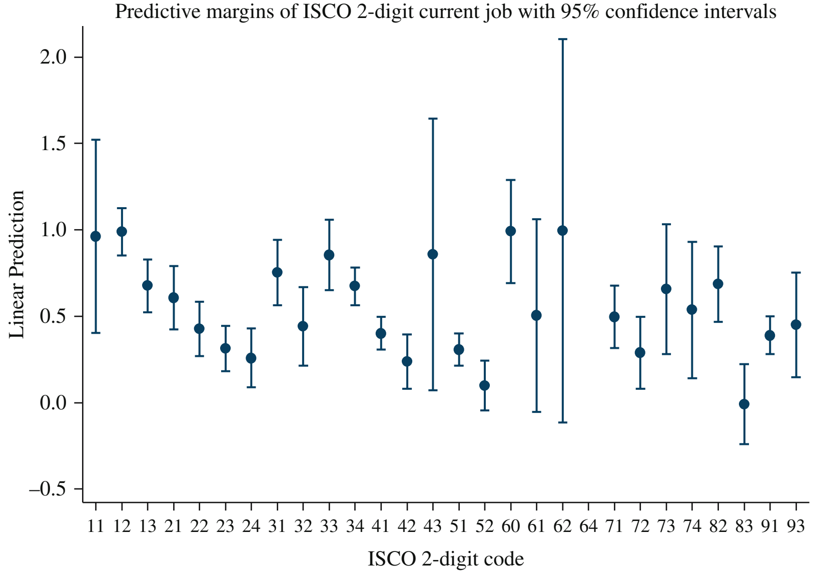


*Fig. 5. Predicted probability of coding error with 95% confidence intervals for ISCO two-digit level – Current job. Note: Predictions from Specification 5, Table 5b.*

mobile-plant operators”). Most of these differences are, however, not statistically significant.

## 5. Discussion and Conclusions

There is growing use of information on occupation in research in the fields of labour economics and sociology, but the quality of occupational data, which is of key importance, is often neglected. Most occupational information in survey data is obtained from direct questions addressed to respondents who provide answers in an open text field. This allows the classification of occupations at a detailed level of disaggregation, but requires coding afterwards (‘office coding’). Promising attempts to code occupational data during CAPI interviews are currently being made using a look-up table or coding index, such as the semantic text-string matching algorithm used in SHARE Wave 6.

In this study, we recoded open-ended questions on occupation for last and current job in the Dutch sample of SHARE data using CASCOT, a well-known and high-quality software program for automatic ex-post coding. We used a Dutch version of CASCOT (CASCOT-NL) in its semiautomatic mode. The combination of a high-quality software program, an expert coder, the use of ancillary information, and the use of an extensive external job titles list ensured a high level of accuracy in coding. A key novelty of our article is the provision of equations that can predict coding errors. We estimated two sets of equations. The first included only individual and job-related characteristics. These equations can be used during (or before) ‘office’ coding: if the survey containing the questions on occupation also provides information on the explanatory variables included in the estimated equation, their values can be used to predict the likelihood that any code attributed is (in)correct. The second set of equations also includes ISCO codes and can be used to predict coding errors during the CAPI interview. In addition to proposing a given ISCO code, the coding software can use these equations to determine the probability that the code is correct. If this probability is low, the interviewee can be asked for additional information. Another possible use of the second set of equations is to recheck office coding: after an ISCO code has been attributed to the occupation, all of the explanatory variables are in fact available for error prediction.

The main findings of this study were: first, the incidence of miscoding in SHARE is high even when comparison is performed at one-digit level – at 28% for last job and 30% for current job. Second, the use of ancillary information drastically increases the number of digits at which the observations are coded. Third, coding errors in occupation are more pronounced for males than for females. Fourth, for the last job variable, they are more likely for more educated individuals and for the self-employed. Fifth, cognitive abilities seem to play an important role in explaining coding errors for current job. Sixth, predictive error equations have a high explanatory power and, finally, ISCO groups 1, 3, and 6 for current job, and 8 for last job, are more susceptible to miscoding.

To reduce coding errors after the interview (‘office coding’) we suggest a semiautomatic software program be used, which also exploits the information provided by ancillary variables as much as possible, such as training and qualifications needed for the job and the industry in which the respondent is working. Many multidisciplinary surveys targeted at older individuals collect information on their last and current jobs (e.g., SHARE, the

English Longitudinal Study of Ageing, and the US Health and Retirement Study). When coding occupations in these surveys, one should ideally make use of measures of individuals' cognitive ability to assist in determining the likelihood of the attributed code being correct. Additional specific questions targeted at the abovementioned groups of occupations should be included in the questionnaire. The main advantage of coding during the interview is that if the response is vague or imprecise, the interviewer can ask the respondent for a more precise job description. Predictive error equations such as those presented in this study may complement the coding software in this novel context.

## Appendix

*Table A1. Mapping of ISCO-08 Major groups to skill levels (Cols. 1 and 2) and mapping of the four ISCO-08 skill levels to ISCED-97 levels of education (Cols. 2 and 3).*

| ISCO-08 Major groups | Skill level | ISCED-97 level |
|---|---|---|
| 1. Managers | 3 + 4 | 5b + 6, 5a |
| 2. Professionals | 4 | 6, 5a |
| 3. Technicians and associate professionals | 3 | 5b |
| 4. Clerical support workers | 2 | 4, 3, 2 |
| 5. Services and sales workers | 2 | 4, 3, 2 |
| 6. Skilled agricultural, forestry and fishery workers | 2 | 4, 3, 2 |
| 7. Craft and related trades workers | 2 | 4, 3, 2 |
| 8. Plant and machinery operators and assemblers | 2 | 4, 3, 2 |
| 9. Elementary occupations | 1 | 1 |

Note: ISCED-97 levels of education: Level 1 = Primary education or first stage of basic education; Level 2 = Lower secondary or second stage of basic education; Level 3 = (Upper) secondary education; Level 4 = Postsecondary nontertiary education; Level 5a = First stage of tertiary education, first degree, medium duration; Level 5b = First stage of tertiary education, short or medium duration, practical orientation; Level 6 = Second stage of tertiary education.
Source: ILO (2012), 14.

*Table A2. Disagreement rate at different number of digits by education levels – Last job and Current job.*

| | Last job | | | | Current job | | | |
|---|---|---|---|---|---|---|---|---|
| | | Disagreement rate (%) | | | | Disagreement rate (%) | | |
| | *Frequencies* | 3 digit | 2 digit | 1 digit | *Frequencies* | 3 digit | 2 digit | 1 digit |
| ISCED 0–1 | 337 | 37 | 30 | 22 | 55 | 47 | 42 | 33 |
| ISCED 2 | 711 | 39 | 32 | 24 | 320 | 47 | 37 | 24 |
| ISCED 3–4 | 355 | 51 | 41 | 35 | 244 | 49 | 40 | 32 |
| ISCED 5–6 | 226 | 55 | 40 | 37 | 263 | 47 | 35 | 31 |
| Total | 1629 | 44 | 35 | 28 | 882 | 48 | 38 | 29 |

Note: Disagreement rate is the percentage of observations coded differently by CASCOT-NL and SHARE.

*Table A3. Disagreement rate at different number of digits by gender – Last job and Current job.*

| | Last job | | | | Current job | | | |
|---|---|---|---|---|---|---|---|---|
| | | Disagreement rate (%) | | | | Disagreement rate (%) | | |
| | *Frequencies* | 3 digit | 2 digit | 1 digit | *Frequencies* | 3 digit | 2 digit | 1 digit |
| Male | 752 | 61 | 48 | 39 | 454 | 61 | 46 | 35 |
| Females | 907 | 29 | 23 | 19 | 435 | 34 | 29 | 23 |
| Total | 1659 | 44 | 35 | 28 | 889 | 48 | 38 | 29 |

Note: Disagreement rate is the percentage of observations coded differently by CASCOT-NL and SHARE.

*Table A4a. Disagreement rate at different number of digits by industry – Last job (sorted by disagreement rate at three digits).*

| Industry | *Frequencies* | Disagreement rate (%) 3 digit | 2 digit | 1 digit |
|---|---|---|---|---|
| Recycling | 1 | 100 | 100 | 100 |
| Research and development | 5 | 80 | 60 | 40 |
| Manufacture of coke, refined petroleum products and nuclear fuel | 14 | 79 | 64 | 64 |
| Manufacture of motor vehicles, trailers and semi-trailers | 14 | 79 | 50 | 43 |
| Electricity, gas, steam and hot water supply | 21 | 76 | 52 | 38 |
| Manufacture of other nonmetallic mineral products | 8 | 75 | 63 | 63 |
| Financial services and insurance | 28 | 64 | 21 | 21 |
| Public administration and defence; compulsory social security | 127 | 61 | 55 | 53 |
| Sewage and refuse disposal, sanitation and similar activities | 5 | 60 | 40 | 40 |
| Manufacture of basic metals, metal products except machinery & equipment | 22 | 59 | 50 | 32 |
| Mining | 74 | 58 | 54 | 24 |
| Computer and related activities | 7 | 57 | 57 | 57 |
| Publishing, printing and reproduction of recorded media | 28 | 57 | 54 | 43 |
| Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials | 9 | 56 | 44 | 44 |
| Hotels and restaurants | 20 | 55 | 55 | 20 |
| Recreational, cultural and sporting activities | 37 | 54 | 38 | 24 |
| Transport, post, telecommunications | 66 | 53 | 45 | 38 |
| Real-estate activities; renting of machinery and equipment without operator and of personal and household goods | 10 | 50 | 20 | 20 |
| Construction | 120 | 48 | 38 | 29 |
| Manufacture of food, tobacco, textiles, clothes, bags, leather goods | 101 | 48 | 41 | 35 |
| Manufacture of furniture; manufacturing NEC | 7 | 43 | 43 | 29 |
| Education | 111 | 41 | 17 | 14 |
| Manufacture of electronic or electric machinery and devices | 17 | 41 | 29 | 18 |
| Manufacture of machinery and equipment NEC | 8 | 38 | 38 | 25 |
| Wholesale trade and commission trade, except of motor vehicles and motorcycles | 34 | 38 | 35 | 29 |
| Activities of membership organisation NEC | 17 | 35 | 24 | 18 |
| Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel | 17 | 35 | 35 | 24 |
| Other business activities | 100 | 33 | 27 | 23 |
| Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods | 192 | 29 | 27 | 22 |
| Other service activities | 39 | 26 | 23 | 21 |
| Health and social work | 211 | 25 | 21 | 19 |
| Total | 1470 | 44 | 35 | 28 |

Note: Disagreement rate is the percentage of observations coded differently by CASCOT-NL and SHARE. Industry is classified using NACE Codes, Version 4 Rev. 1 1993 (see http://www.top500.de/nace4-e.htm for a description of NACE Version 4 Rev. 1 and MEA 2013, pp. 32–33 for the shorter classification used in SHARE).

Table A4b. Disagreement rate at different number of digits by industry – Current job (sorted by disagreement rate at three digits).

| | | Disagreement rate (%) | | |
|---|---|---|---|---|
| Industry | *Frequencies* | 3 digit | 2 digit | 1 digit |
| Electricity, gas, steam and hot water supply | 6 | 100 | 83 | 67 |
| Manufacture of motor vehicles, trailers and semi-trailers | 2 | 100 | 50 | 50 |
| Manufacture of other nonmetallic mineral products | 1 | 100 | 100 | 100 |
| Mining | 43 | 84 | 77 | 33 |
| Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials | 7 | 71 | 43 | 29 |
| Construction | 50 | 68 | 44 | 38 |
| Hotels and restaurants | 9 | 67 | 67 | 22 |
| Manufacture of basic metals, metal products except machinery & equipment | 3 | 67 | 67 | 33 |
| Research and development | 3 | 67 | 67 | 67 |
| Financial services and insurance | 15 | 60 | 27 | 20 |
| Manufacture of food, tobacco, textiles, clothes, bags, leather goods | 22 | 59 | 59 | 41 |
| Real-estate activities, renting of machinery and equipment without operator and of personal and household goods | 12 | 58 | 42 | 42 |
| Transport, post, telecommunications | 35 | 57 | 51 | 40 |
| Computer and related activities | 9 | 56 | 56 | 44 |
| Other business activities | 61 | 54 | 41 | 36 |
| Manufacture of coke, refined petroleum products and nuclear fuel | 4 | 50 | 25 | 25 |
| Manufacture of electronic or electric machinery and devices | 6 | 50 | 33 | 17 |
| Public administration and defence; compulsory social security | 68 | 50 | 43 | 35 |
| Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel | 13 | 46 | 38 | 23 |
| Education | 107 | 41 | 19 | 16 |
| Recreational, cultural and sporting activities | 25 | 40 | 32 | 20 |
| Manufacture of machinery and equipment NEC | 8 | 38 | 38 | 25 |
| Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods | 58 | 34 | 31 | 24 |
| Health and social work | 173 | 33 | 30 | 28 |
| Publishing, printing and reproduction of recorded media | 12 | 25 | 25 | 25 |
| Other service activities | 17 | 24 | 18 | 18 |
| Activities of membership organisation NEC | 7 | 14 | 14 | 14 |
| Manufacture of furniture; manufacturing NEC | 3 | 0 | 0 | 0 |
| Wholesale trade and commission trade, except of motor vehicles and motorcycles | 4 | 0 | 0 | 0 |
| Total | 783 | 47 | 37 | 29 |

Note: Disagreement rate is the percentage of observations coded differently by CASCOT-NL and SHARE. Industry is classified using NACE Codes, Version 4 Rev. 1 1993 (see http://www.top500.de/nace4-e.htm for a description of NACE Version 4 Rev. 1 and MEA 2013, pp. 32–33 for the shorter classification used in SHARE).

*Table A5. Educational attainment and gender composition across ISCO-88 one-digit groups.*

| ISCO one digit | % Primary | % Lower secondary | % Upper secondary | % Tertiary | Mean years of education | % Female |
|---|---|---|---|---|---|---|
| 1 | 5.6 | 30.4 | 29.9 | 34.1 | 14.0 | 20.3 |
| 2 | 0.8 | 14.2 | 21.2 | 63.7 | 16.1 | 54.6 |
| 3 | 3.2 | 22.8 | 35.1 | 38.9 | 14.0 | 41.5 |
| 4 | 7.8 | 50.4 | 32.6 | 9.2 | 12.6 | 72.4 |
| 5 | 18.9 | 54.7 | 21.6 | 4.8 | 11.6 | 81.9 |
| 6 | 20.0 | 61.4 | 12.9 | 5.7 | 11.2 | 42.3 |
| 7 | 31.5 | 48.2 | 17.5 | 2.8 | 9.8 | 20.6 |
| 8 | 29.8 | 49.7 | 17.1 | 3.3 | 10.9 | 20.0 |
| 9 | 35.3 | 50.5 | 10.7 | 3.6 | 9.9 | 70.6 |
| Total | 15.1 | 40.2 | 23.7 | 21.0 | 12.5 | 51.2 |

Note: The table refers to current and last job pooled data and SHARE coding.

## 6. References

Autor, D. 2013. "The 'Task Approach' to Labour Markets: an Overview." *Journal of Labour Market Research* 46: 185–199. Doi: http://dx.doi.org/10.1007/s12651-013-0128-z.

Autor, D., L.F. Katz, and M.S. Kearney. 2006. "The Polarization of the US Labor Market." *American Economic Review* 96: 189–194. Doi: http://dx.doi.org/10.1257/000282806777212620.

Autor, D., F. Levy, and R.J. Murnane. 2003. "The Skill Content Of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118: 1279–1333. Doi: http://dx.doi.org/10.1162/003355303322552801.

Bethmann, A., M. Schierholz, K. Wenzig, and M. Zielonka. 2014. *Automatic Coding of Occupations Using Machine Learning Algorithms for Occupation Coding in Several German Panel Surveys*. In: Statistics Canada (Ed.), Beyond traditional survey taking. Adapting to a changing world. Proceedings of Statistics Canada Symposium 2014, Quebec. Available at: http://fdz.iab.de/342/section.aspx/Publikation/k151124301 (accessed October 2016).

Biemer, P.B. and L.E. Lyberg. 2003. *Introduction to Survey Quality*. New York: John Wiley & Sons, Inc.

CBS (Statistics Netherlands). 2012. *Coding Tool Implemented in 2012 for Coding Occupations in Social Surveys*. Internal Document. The Hague: Statistics Netherlands.

Cheeseman Day, J. 2014. *Using an Autocoder to Code Industry and Occupation in the American Community Survey*. Presentation held at the Federal Economic Statistics Advisory Committee Meeting, 13 June 2014. Available at: http://www2.census.gov/adrm/fesac/2014-06-13_day.pdf (accessed October 2016).

Christelis, D., T. Jappelli, and M. Padula. 2010. "Cognitive Abilities and Portfolio Choice." *European Economic Review* 54: 18–38. Doi: http://dx.doi.org/10.1016/j.euroecorev.2009.04.001.

Commission of the European Communities. 2009. "Commission Regulation (EC) No 1022/2009 of 29 October 2009 amending Regulations (EC) No 1738/2005, (EC) No 698/2006 and (EC) No 377/2008 as regards the International Standard Classification of Occupations (ISCO)." *Official Journal of the European Union*, L 283/3, 30 October 2009. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32009R1022&from=EN (accessed October 2016).

DESA. 2010. "Handbook on Population and Housing Census Editing." Revision 1, Series F, No 82 (Studies in Methods (Ser. F)), United Nations Statistics Division. New York: United Nations Publication. Available at: http://unstats.un.org/unsd/publication/SeriesF/seriesf_82rev1e.pdf (accessed October 2016).

Elias, P., K. Halstead, and K. Prandy. 1993. *Computer-Assisted Standard Occupational Coding*. London: HMSO.

Elias, P. 1997. "Occupational Classification (ISCO-88). Concepts, Methods, Reliability, Validity and Cross-National Comparability." *OECD Labour Market and Social Policy Occasional Papers* No. 20, OECD Publishing. Doi: http://dx.doi.org/10.1787/304441717388.

Ellison, R. 2014. *Demonstration of Performance of CASCOT 5.0*. Presentation held at the *CASCOT*: Occupational Coding in Multi-national Surveys Workshop, 10–11 April 2014, Venice. Available at: http://dasish.eu/dasishevents/cascotworkshop/programmepres/ (accessed 15 April, 2014).

Feenstra, R.C. and G.H. Hanson. 1996. "Globalization, Outsourcing, and Wage Inequality." *American Economic Review* 86: 240–245. Available at: http://www.jstor.org/stable/2118130.

Fletcher, J.M., J.L. Sindelar, and S. Yamaguchi. 2011. "Cumulative Effects of Job Characteristics on Health." *Health Economics* 20: 553–570. Doi: http://dx.doi.org/10.1002/hec.1616.

Ganzeboom, H. 2008. *Occupation Coding: Do's And Dont's*. Version 2, 2 August 2008. Available at: http://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/issp/members/codinginfo/ISCO-coding_dos_donts_HG2008.pdf (accessed October 2016).

Goos, M. and A. Manning. 2007. "Lousy and Lovely Jobs: The Rising Polarization of Work in Britain." *Review of Economics and Statistics* 89: 118–133. Doi: http://dx.doi.org/10.1162/rest.89.1.118.

Hartog, J. 2000. "Over-Education and Earnings: Where Are We, Where Should We Go?" *Economics of Education Review* 19: 131–147. Doi: http://dx.doi.org/10.1016/S0272-7757(99)00050-3.

Hoffmann, E., P. Elias, B. Embury, and R. Thomas. 1995. *What Kind Of Work Do You Do? Data Collection and Processing Strategies When Measuring "Occupation" for Statistical Surveys and Administrative Records*. ILO working paper, N.95-1. Geneva: ILO. Available at: http://www.ilo.org/global/statistics-and-databases/WCMS_087880/lang–en/index.htm (accessed October 2016).

Jackle, A. 2008. "Dependent Interviewing: Effects on Respondent Burden and Efficiency of Data Collection." *Journal of Official Statistics* 24: 411–430.

Jones, R. and P. Elias. 2004. "CASCOT: Computer-Assisted Structured Coding Tool." Coventry: Warwick Institute for Employment Research, University of Warwick.

ILO. 2014. ISCO: International Standard Classification of Occupations. Available at: http://www.ilo.org/public/english/bureau/stat/isco/ (accessed 3 October, 2016).

ILO. 2012. *International Standard Classification of Occupations: Structure, Group Definitions and Correspondence tables*. Vol. 1. Geneva: ILO. Available at: http://www.ilo.org/wcmsp5/groups/public/—dgreports/—dcomm/—publ/documents/publication/wcms_172572.pdf (accessed October 2016).

ILO. 2010. *Measuring the Economically Active in Population Censuses: A Handbook*. Studies in Methods Series F, No. 102. New York: ILO and UN.

Leist, A.K., M.M. Glymour, J.P. Mackenbach, F.J. van Lenthe, and M. Avendano. 2013. "Time Away from Work Predicts Later Cognitive Function: Differences by Activity During Leave." *Annals of Epidemiology* 23: 455–462. Doi: http://dx.doi.org/10.1016/j.annepidem.2013.05.014.

Leuven, E. and H. Oosterbeek. 2011. "Overeducation and Mismatch in the Labor Market." In *Handbook of the Economics of Education*, edited by E. Hanushek, S. Machin and L. Woessmann, 283–326. Amsterdam: Elsevier.

MEA. 2013. SHARE Release Guide 2.6.0. Waves 1 & 2. Munich: publisher. Available at: Munich Center for the Economics of Ageing (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC) publishing. (accessed October 2016).

Moscarini, G. and K. Thomsson. 2007. "Occupational and Job Mobility in the US." *The Scandinavian Journal of Economics* 109: 807–836. Doi: http://dx.doi.org/10.1111/j.1467-9442.2007.00510.x.

Perales, F. 2014. "How Wrong Were We? Dependent Interviewing, Self-Reports and Measurement Error in Occupational Mobility in Panel Surveys." *Longitudinal and Life Course Studies* 4: 299–316. Doi: http://dx.doi.org/10.14301/llcs.v5i3.295.

Ravesteijn, B., H. van Kippersluis, and E. van Doorslaer. 2013. *The Wear and Tear on Health: What is the Role of Occupation?* Tinbergen Institute Discussion Paper 13-143. Amsterdam: Tinbergen Institute. Available at: http://papers.tinbergen.nl/13143.pdf (accessed October 2016).

Rose, D. and E. Harrison. 2007. "The European Socio-Economic Classification: A New Social Class Schema For Comparative European Research." *European Societies* 9: 459–490. Doi: http://dx.doi.org/10.1080/14616690701336518.

Tijdens, K.G. 2014a. "Drop-Out Rates During Completion of an Occupation Search Tree in Web-Surveys." *Journal of Official Statistics* 30: 23–43. Doi: http://dx.doi.org/10.2478/jos-2014-0002.

Tijdens, K.G. 2014b. *Reviewing the Measurement and Comparison of Occupations Across Europe*. AIAS Working Paper 149. Amsterdam: University of Amsterdam. Available at: http://dare.uva.nl/record/1/432281 (accessed October 2016).

United Nations. 2007. "Updating the International Standard Classification of Occupations (ISCO): Summary of major changes between ISCO-88 and ISCO-08 (Feb 2007 draft)." Paper for discussion by the Expert Group on International Economic and Social Classifications, New York, 16–18 April 2007. Available at: http://unstats.un.org/unsd/class/intercop/expertgroup/2007/AC124-11.PDF (accessed October 2016).

United Nations (UN). 2014. *National Classifications*. Available at: http://unstats.un.org/unsd/cr/ctryreg/ctrylist2.asp (accessed March 2015).

Westerman, S. 2014. "CBS and CASCOT: tuning CASCOT for improved performance." Presentation held at the *CASCOT*: Occupational Coding in Multi-national Surveys Workshop, 10–11 April 2014, Venice. Available at: http://dasish.eu/dasishevents/cascotworkshop/programmepres/ (accessed 15 April, 2014).

Wilcoxon, F. 1945. "Individual Comparisons by Ranking Methods." *Biometrics* 1: 80–83. Available at: http://www.jstor.org/stable/3001968.

Wooldridge, J.M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Received December 2014
Revised April 2016
Accepted July 2016

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 947–962, http://dx.doi.org/10.1515/JOS-2016-0050

# Demographic Projections: User and Producer Experiences of Adopting a Stochastic Approach

*Kim Dunstan*[1] *and Christopher Ball*[2]

Statistics New Zealand is one of the few national statistical agencies to have applied a stochastic (probabilistic) approach to official demographic projections. This article discusses the experience and benefits of adopting this new approach, including the perspective of a key user of projections, the New Zealand Treasury. Our experience is that the change is less difficult to make than might be expected. Uncertainty in the different projection inputs (components) can be modelled simply or with more complexity, and progressively applied to different projection types. This means that not all the different demographic projections an agency produces need to adopt a stochastic approach simultaneously. At the same time, users of the projections are keen to better understand the relative certainty and uncertainty of projected outcomes, given the important uses of projections.

*Key words:* Population; projection; stochastic; uncertainty.

## 1. Introduction

Demographic projections are a high-priority output for most national statistical organisations. They complement other demographic statistics, such as census statistics and population estimates, to give information about possible future changes in the size and composition of populations (and families, households, and the labour force). In doing so, the projections assist planning and decision making in areas such as health, education, housing, retirement planning, and transport.

This article describes the user and producer experiences of adopting a stochastic (probabilistic) approach to official projections of New Zealand's population, ethnic populations, and labour force. It complements the Letter to the Editor of this journal by Bijak et al. (2015) which touches on some related themes. The stochastic approach implemented by Statistics NZ in 2012 was a shift from the conventional deterministic approach used extensively in demographic projections worldwide. Uncertainty and stochastic processes are not exclusive to the population domain, but the future is inherently uncertain, so demographic projections seem an obvious area of application.

This article does not detail the stochastic methods that have been used, or how specific stochastic models are chosen. These are discussed more fully in Dunstan (2011), Woods and Dunstan (2014), and Statistics NZ (2012a, 2012b, 2014, 2015a, 2015b, 2016).

[1] Statistics New Zealand, Private Bag 4741, Christchurch 8140, New Zealand. Email: kim.dunstan@stats.govt.nz
[2] New Zealand Treasury. BP Box 3724, Wellington 6008, New Zealand. Email: christopher.ball@treasury.govt.nz

This process of change offers lessons for others: why change to stochastic projections and how difficult is it to change? Section 2 contrasts the deterministic and stochastic approaches. Section 3 explores the advantages of a stochastic approach to understand why Statistics NZ is one of the few national statistical agencies to have applied a stochastic approach to demographic projections. Section 4 discusses the practical challenges of making the change, while Section 5 presents thoughts on the likely future direction of demographic projections. Section 6 concludes with some recommendations.

## 2. Deterministic and Stochastic Approaches Contrasted

At first glance, Figure 1 suggests little difference between deterministic and stochastic projections of the population. Like most national statistical organisations, Statistics NZ's projections have conventionally been derived deterministically – by combining specific assumptions (e.g., about fertility, mortality, and migration) to produce a single projection. Different projections or scenarios can be produced by systematically combining different assumptions (Figure 1a). Collectively, those different projections can convey something about the relative certainty or uncertainty of different outcomes, but not in any quantified way. Hence, the probability that an outcome will be above or below a given scenario is unknown and is not estimated in a deterministic projection.

Stochastic or probabilistic projections are produced in much the same way as deterministic projections. The most important difference is that the projection assumptions – the critical inputs to the projection model – also include a measure of variability. The stochastic approach typically involves creating multiple simulations for each of the projection assumptions. The simulations vary randomly according to the probability distributions of each assumption derived from empirical models using historical data, or from judgements (Lutz 2009; Booth and Tickle 2008). These simulations of the assumptions are combined in a conventional way, namely using the cohort-component method (Statistics NZ 2012b, 2014, 2015a, 2016), to produce a population simulation or projection.

The input assumptions and resulting population simulations have realistic trajectories, with all the year-to-year fluctuations inherent in the real world. However, the real value of

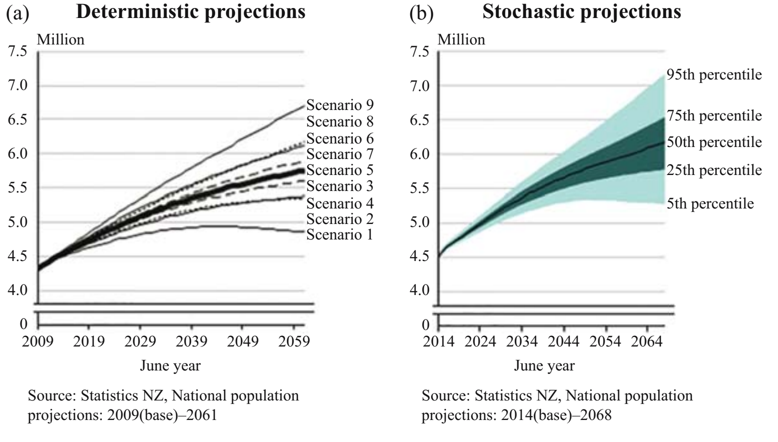


*Fig. 1. Alternative projections of the New Zealand population.*

the stochastic approach is not in the individual simulations per se, but in repeating the process and creating many (e.g., 2,000) simulations. From this collective, we can build up a probability distribution of population size and other characteristics. The distribution can be summarised by percentiles, and illustrated using fan charts (Figure 1b), which indicate the probability of different outcomes. In this case, the fifth percentile indicates a five percent chance that the given outcome will be lower than this percentile.

There are many different ways to model uncertainty and to produce stochastic projections. For the purposes of this article, it is sufficient to note that the different approaches share a common aim of conveying uncertainty in a quantified way by allowing the inputs into the models to vary randomly according to probability distributions. Both expectation (e.g., Billari et al. 2014; Lutz et al. 2014; Lutz 2009; Lutz and Scherbov 1998) and empirical approaches are used internationally to model uncertainty, and both have their advantages and disadvantages. The empirical approaches include the propagation of deviations from past projections (e.g., Stoto 1983; Keilman 1997) as well as statistical models (e.g., Wisniowski et al. 2015; United Nations 2014; Raftery et al. 2012), although even these latter approaches include a mix of expert-based inputs. Empirical models have been most intensively applied where demographic trends have been largely monotonic and sustained, as in the case of death rates and life expectancy (e.g., Lee and Carter 1992).

In Statistics NZ's case, statistical models have been developed where possible to generate stochastic measures of the critical projection assumptions, notably for mortality (Woods and Dunstan 2014). However, other assumptions retain the conventional approach of using the expectations (or so-called 'expert judgement') of demographers within Statistics NZ, reflecting that assumptions need to be plausible in both the short term (over the next ten years) and long term (beyond 50 years). Current methods for projections of New Zealand's total population are summarised in Table 1 and highlight that Statistics NZ uses a mix of expectation and empirical approaches. This is a pragmatic approach if not a purely statistical one, but would be the case whether Statistics NZ's projections were deterministic or stochastic.

For projections of ethnic populations – of four broad and overlapping ethnic groups (European, Māori, Asian, and Pacific) – and labour force, a mix of expectation and empirical approaches also applies. These projections include extra assumptions on paternity and interethnic mobility (Statistics NZ 2015a), and labour-force participation and average hours worked (Statistics NZ 2015b). In the case of ethnic projections, greater emphasis is put on expectation due to the shorter historical time series that are available compared with the total population. From a user's perspective, however, this different emphasis makes no visible difference to the projection results.

## 3. Benefits of a Stochastic Approach

The advantages of a stochastic approach are discussed in Alho (1997, 2005), Booth (2006), Bryant (2003, 2005), and Keilman (1991) among others. In this section we raise the discussion above a purely academic or theoretical level by discussing the real benefits experienced in the production and publication of stochastic projections since 2012.

By providing quantification of uncertainty, stochastic projections have assisted the interpretation of projections. There are several aspects to the improved interpretation.

*Table 1. Method of assumption formulation for New Zealand population projections (published 2016).*

| Assumption | Median (50th percentile) | Variance/distribution of values |
|---|---|---|
| Base population | Empirical model: official population estimates based on census and post-enumeration survey | Expectation (judgement): variance varies by age-sex |
| Fertility | Expectation (judgement): long-term total fertility rate of 1.85 births per woman | Empirical model: ARIMA (0,1,0) model fitted to total fertility rate for 1977–2016 June years |
| Mortality | Empirical model: coherent functional demographic model fitted to age-specific death rates for 1977–2015 June years | |
| Migration | Expectation (judgement): long-term annual net migration of 15,000 | Empirical model: ARIMA (1,0,1) model fitted to net migration for 1988–2016 June years |
| Sex ratio at birth | Empirical model: median and variance from sex ratio at birth for 1900–2015 December years | |

Firstly, the probability distribution of a given variable or characteristic in the stochastic population projections is often skewed, which is typically not conveyed by conventional scenarios. Secondly, the resulting fan charts are intuitive. The 50th percentile or median is analogous to the mid-range deterministic projection that used to be derived. For those people who want one number, they can take that median or indeed another percentile. Finally, the stochastic projections offer more in terms of interpretation. The fan chart of the stochastic projections in Figure 1b, for example, indicates that there is a 50 percent chance that the actual population will lie within the dark band and a 90 percent chance that the actual population will lie within the wider combined lighter and dark band.

By contrast, deterministic scenarios give a poor indication of uncertainty for some key demographic characteristics (e.g., dependency ratios, death numbers). Even for other characteristics, the uncertainty indicated by the scenarios is neither consistent between characteristics, nor consistent across the projection period. For example, Scenarios 1 and 9 give the lowest and highest New Zealand populations, respectively (Figure 1a), but it is Scenarios 3 and 7 which give the lowest and highest ratio of 65+ population to 15–64 population (Figure 2a). Scenarios 1 and 9 actually give a misleadingly narrow range for the ratio of 65+ population to 15–64 population. This partly reflects that the 'low' and 'high' variant assumptions combined to give the different scenarios are not equivalent to a consistent probability interval between the fertility, mortality, and migration assumptions (Bongaarts and Bulatao 2000; Bryant 2005; Lee 1998). Deterministic projections can therefore risk giving the impression that a variable or characteristic is more certain than what is presented in stochastic projections.

The use of time-series models is often aimed at improving the accuracy of projections (e.g., United Nations 2014; Woods and Dunstan 2014). However, it is important to clarify

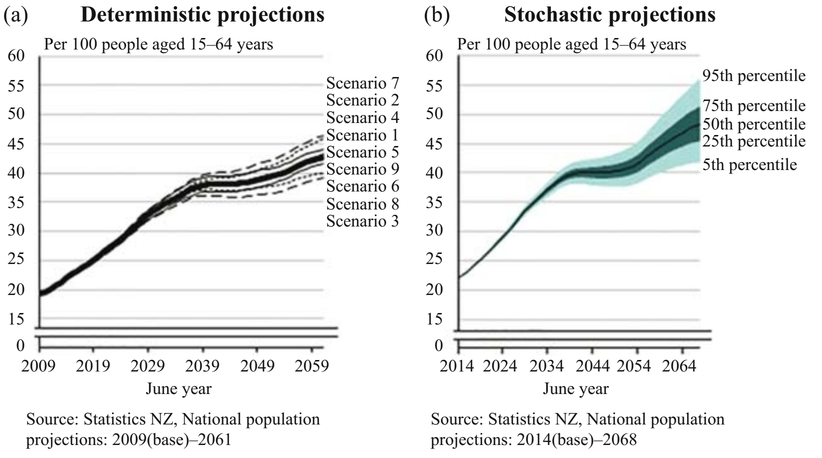


*Fig. 2. Alternative projections of the ratio of 65+ population to 15–64 population.*

that Statistics NZ's shift to stochastic projections was not driven by an expectation of improving the predictive power or accuracy of the projections. The main driver was to improve the interpretation of projections. Partly, this reflects that Statistics NZ's projections are not produced solely using empirical models (Table 1). Given the inherent uncertainty of the future, encouraging users of projections to think about uncertainty is important. It is only feasible for users to think about uncertainty if that uncertainty is conveyed to them appropriately.

Some other benefits of stochastic projections have yet to be realised, but are worth identifying here as we expect these benefits will be forthcoming. Essentially these benefits are extensions of enhanced interpretation.

First, stochastic projections, with their quantified measures of uncertainty, can actually help statistical agencies make and validate decisions to define the projection period. In principle, there is no limit to how far a projection can be extended (the projection horizon). In practice, statistical agencies are wary of publishing projections with very long horizons, on the basis that the large uncertainty makes the projections more misinformative than informative. For projections of different subpopulations, the horizon might be justifiably different. For example, the stochastic projections for ethnic populations (Statistics NZ 2015a) do have wider uncertainty intervals for all characteristics than the stochastic projections of the total population (Statistics NZ 2014). Moreover, for users of projections, the probability distribution can help them make their own informed decisions about the usefulness of different projections across any projection period.

Second, stochastic projections are generally a better input for users who are involved in modelling and projecting other parameters. For example, stochastic projections help them understand the sensitivity of their projections to the demographic inputs. Or, they help them understand the importance of demographic uncertainty relative to uncertainty coming from, say, economic parameters. For users wanting to use the projections in their own models, they will often need to access the full dataset of simulations. Published percentiles may exclude the specific percentiles needed by specific users, and even a full range of percentiles will sometimes be inadequate, especially where users want to extend the stochastic approach to their own models (e.g., with variability estimated for additional parameters).

Population projections are a core input into the New Zealand Treasury's Long Term Fiscal Model (LTFM), which quantifies the sustainability of public finances over a 50-year horizon (New Zealand Treasury 2013). Large government expenditure items – such as public pensions, education and health – are sensitive to changes in the age structure of the population, whereas revenue is sensitive to changes in the size of the labour force. The long-term gap between revenue and expenditure, which is typically used to assess fiscal sustainability, is largely driven by changes in the population and labour-force projections.

Before the availability of stochastic population projections, Treasury typically showed sensitivity of the LTFM to the demographic inputs by using deterministic scenarios which altered one of the fertility, mortality, or migration assumptions. These alternative projections gave a range of possible outcomes, but did not give an indication of how likely a given alternative was relative to another projection (including the mid-range projection). Hence, it is difficult for users to test sensitivity to the population inputs by altering single demographic assumptions. Stochastic projections make that assessment explicit and Treasury is now developing the LTFM to incorporate a stochastic approach (Ball et al. 2015).

Stochastic population projections, conveyed using percentiles and fan charts, give users an idea of the mid-range projection, but also place greater emphasis on the uncertainty in each direction. The uncertainty is conditional on past variability (if using empirical approaches) or expert input (if using expectation). More generally, if modelled variability is not indicative of future variability, then the projections may misrepresent the true probability intervals. See Raftery et al. (2012) for an evaluation of calibration of uncertainty, based on probability intervals for 1990–2010 using variability modelled from 1950–1990 data.

The challenges of conveying uncertainty and the nature of the projections generally are new neither from a user perspective nor from a producer perspective and underscore several aspects. First, they emphasise the importance of communicating to users what the projections are (e.g., an indication of future trends based on current policy settings) and how the assumptions have been derived (e.g., how the variability is estimated for each input assumption). Second, it highlights the importance of regularly updating the projections to incorporate changes to levels and variability in the assumptions, as well as to changes in policy settings. Statistics NZ currently updates its projections every 2–3 years. Third, it reinforces the observation that the value of population projections is only partly defined by whether or not they match reality, especially given the long-term horizon of many projections, but also about whether they are plausible and useful to users at the time the projections were published. Projections aim to form a basis for developing reasonable expectations about the future; to help focus attention on potential events, risks, and opportunities; and to assist people and policy makers to plan and make decisions accordingly. Fourth, this motivates the production of alternative 'what if?' scenarios to complement stochastic projections. These allow particular scenarios of interest to be examined and compared with the benchmark stochastic projections.

## 4. Challenges of Implementing a Stochastic Approach

National statistical agencies are generally wary of adopting new statistical methodologies. This wariness is warranted, as adopting every new statistical development is neither pragmatic nor cost-effective. In this section we outline the main challenges we faced in adopting a stochastic approach and how these were overcome, as a guide to how other producers of projections might negotiate such challenges.

### *4.1. Computing Capacity*

From a production viewpoint, producing large numbers of simulations means that datasets are bigger and programs take longer to run. This can be an issue for users also if they are using those simulations in their own projection models. Computing capacity was a legitimate practical constraint in the past, but it is difficult to justify this as a constraint on implementation in the 21st century.

### *4.2. User Need*

It is sometimes suggested that users of the projections neither want nor need uncertainty to be conveyed in any quantified way. Raftery (2014) identifies five types of general users of projections. Of these, the 'low-stakes user' may have little use for anything but a mid-range projection. However, for other user types, an accurate assessment of projection uncertainty is vital as it affects if and how they use the projections.

Importantly, users of projections get nothing less with Statistics NZ's stochastic projections than they got before. They can still get and use one number or projection according to their needs (e.g., median or other percentile). They can still get scenario-type projections with the 'what if?' scenarios which assume fertility, life expectancy, or net migration at specific levels (Statistics NZ 2014, 2015b, 2016). Users need not worry that the projections look radically different to what they did historically. What users also get, however, is better information about uncertainty.

### *4.3. User Expectations*

Stochastic projections are not necessarily more accurate (or less accurate) than conventional projections when compared with actuality (i.e., observed population change). There is evidence that empirically based stochastic projections are better than expert-based ones (e.g., Alkema et al. 2011; Raftery et al. 2013). Whether or not they are more accurate, stochastic projections do assist interpretation. Communicating these aspects to users is important to ensure their expectations align realistically with what is produced.

Statistics NZ has not conducted any specific research on the use and understanding of the stochastic projections compared with deterministic projections. However, the development of the stochastic projections from 2005 was gradual rather than a sudden substantial shift in approach. This allowed time for discussions with key users, for discussions at population conferences (e.g., Population Association of New Zealand), for national population prototypes to be developed in 2005 and 2010, and for the publication of a working paper (Dunstan 2011). There was therefore the opportunity for both the producer and users to consider the implications of a shift in approach before the new methodology was adopted. Importantly, there has been no negative reaction from users to the adoption of a stochastic approach in the release of official projections since 2012.

### 4.4. User Understanding

Two important aspects of stochastic projections to convey to users is information about the stochastic methods (metadata) and the projections (results) themselves. These aspects are not new for statistical agencies, which have always had the challenge of conveying technical aspects and detailed data to a variety of users.

The experience of Statistics NZ was that existing products and services were suitable for disseminating the stochastic projections. Conventional information releases presented summary results via commentary, graphs and tables (Statistics NZ 2012a, 2012b, 2014, 2015a, 2015b, 2016). More detailed data for each of the projections were disseminated using the existing web-based tool NZ.Stat, a table-builder product powered by software provided by the OECD (Statistics NZ 2012c, OECD 2013).

Conventionally, projections of different characteristics from different scenarios would be published. With the stochastic projections, different percentiles (5th, 10th, 25th, 50th, 75th, 90th, and 95th) of those same characteristics are published. These published percentiles appear to have satisfied user demand, although other percentiles can be readily supplied on request. Given the trend of increasing data dissemination, one can envisage users in future being able to select any percentiles of interest to them, or even the individual simulations for use in their own modelling.

In the two years following the publication of the first stochastic population projections in 2012, the number of unique visitors viewing the main NZ.Stat table (population by age-sex) online averaged more than 100 per month. We are not aware of any of those users requiring assistance to interpret the projections through the website feedback forms or Statistics NZ's free helpline. While this could simply indicate that users are not seeking assistance when they need it, the continued use and downloading of the stochastic projections seems to suggest that users generally understand them.

### 4.5. Spurious Precision

The estimates of uncertainty are themselves uncertain and Statistics NZ has always been upfront about this (Statistics NZ 2012a, 2012b, 2014, 2015a, 2015b, 2016). The projections do not try to include uncertainty arising from catastrophic events (e.g., earthquakes, wars). As they are projections based on current policy settings, they do not try to anticipate major policy changes, so there is the additional uncertainty of 'nondemographic' factors (assuming these factors are not captured in the modelling of 'demographic' uncertainty). The estimates of uncertainty depend on what historical data is used and how uncertainty is modelled. While it is possible to estimate uncertainty based on the historical variability of the demographic parameters, it is more difficult to estimate the uncertainty that arises from the choice of models (for one approach, see Abel et al. 2013), or from the choice of time period(s) that affect the model parameters.

Dowd et al. (2010) refer to these three different types of uncertainty as:

1. model uncertainty (e.g., we do not know the true fertility model),
2. parameter uncertainty (e.g., whatever mortality model we use, we do not know the true values of its parameters),
3. projection uncertainty (e.g., the uncertainty of future migration rates given any particular model and its calibration).

There may be additional uncertainty arising from factors such as errors in historical data and errors in expert judgement (see Keilman 1991; Alho 1997; De Beer 2000). Notably, Bayesian methods are capable of combining various sources of uncertainty in a coherent manner, including, for example, expert uncertainty, model uncertainty, parametric uncertainty, and covariate uncertainty (see Bijak and Bryant 2016). In effect, many stochastic projections (including those of Statistics NZ) model only the projection uncertainty, and are therefore inclined to underestimate the true uncertainty. Nonetheless, this is an improvement on conventional deterministic projections which do not model any of the types of uncertainty. So while the estimated probability intervals in the stochastic projections may appear spuriously precise, this is preferable to a spuriously precise deterministic projection.

One of the challenges for Treasury in moving to a stochastic projection framework is conveying a more accurate assessment of the fiscal pressures without giving the impression of spurious accuracy. In fact, stochastic projections better allow users to focus on broad trends, without overemphasising small changes in the mid-range projection that occur between updates every few years. But there is a challenge in conveying the full uncertainty in fiscal projections in addition to the uncertainty coming from the demographic projections. There are judgements around the effects of government policy and economic variables which are also subject to uncertainty. As a result, the current fiscal projections underestimate the true range of uncertainty.

### 4.6. *Impact on Other Projections*

Like many statistical agencies, Statistics NZ produces a suite of demographic projections, including projections of subpopulations (Figure 3). These projections are not produced

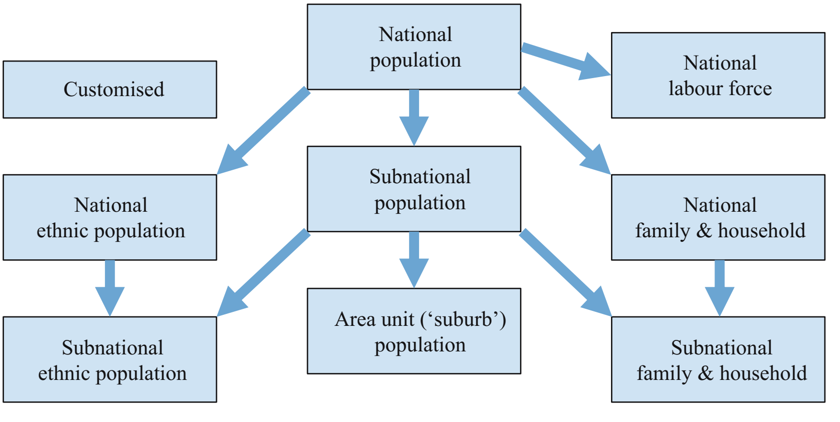


*Fig. 3. Demographic projections produced by Statistics NZ.*

using a single model (e.g., micro-simulation), but are produced using independent projection models, albeit designed explicitly or implicitly to produce an internally consistent set of projections. For example, the mid-range projection for subnational areas will align with the mid-range national projection. This additivity of geographical areas is highly sought by users and contributes to the general coherence of the projections produced.

One of the attractions of Statistics NZ's stochastic approach is that it can be applied to one or more of its projection types without compromising the other projections. For example, a stochastic approach was applied to national population and labour-force projections in 2012, and was extended to national ethnic population projections in 2015. Other projections, including all subnational projections, have continued in their more conventional deterministic form. This is an important consideration for statistical agencies. who might see the challenge of producing stochastic projections for subpopulations a barrier to producing any stochastic projections at all. Long and Hollman (2004) discuss this in relation to ethnic and Hispanic population projections in the United States.

### *4.7. Cost*

A detailed assessment of the costs to develop and produce stochastic projections is beyond the scope of this article, but a few observations based on Statistics NZ's experience are pertinent. First, there are nontrivial investment costs to formulate measures of uncertainty using empirical and/or expectation approaches, to embed the stochastic methods within the organisational production process, and to engage with users in advance of adoption. However, these are initial one-off costs and producers of projections can potentially benefit from the development of stochastic projections by the United Nations, Statistics NZ and elsewhere, thereby reducing the cost and time of their own development. The availability of free open-source software (e.g., Sevcikova et al. 2011; Hyndman 2015) also means cost need not be an obvious deterrent to implementation.

Beyond the development phase, the ongoing production costs have been similar to those of producing conventional deterministic projections. Additional resources are required to formulate measures of uncertainty and to produce multiple simulations when each set of projections is produced, although this is offset by not needing to produce alternative deterministic assumptions (e.g., 'low' and 'high' variants). It was possible to use existing products and services to disseminate metadata and the stochastic projections.

## 5. Future Developments

In common with other statistical agencies, Statistics NZ has an expressed set of strategic objectives to which it aspires. These include being a trusted provider of official statistics, empowering customers in data understanding and use, and driving value for customers through the use of innovative tools and techniques (Statistics NZ 2015c). The development of stochastic projections can be viewed as concomitant with the organisation's core values and strategic direction. As a producer of official (population) statistics, high priority is placed upon the statistical rigour of data, concepts, and methods, including incorporating international best practices. However, there is also an onus on Statistics NZ to provide statistical leadership (e.g., in statistical methodologies), which can run counter to more conservative statistical practices.

Lutz and Scherbov (1998) discuss how "the change of a long-established tradition" generally requires the following:

1. The new practice must have clear advantages when compared with the current one.
2. It should be consistent with other work done by the producing institution, and present an evolution along established lines rather than a discontinuity.
3. The proposed approach should be internally consistent and based on accepted scientific work.
4. It should be practical for both the users and producers, and not cost too much.

What is the experience of Statistics NZ in terms of these criteria? Stochastic projections do offer clear advantages, as discussed in this article. The adoption of a stochastic approach has been a progressive evolution in methodology and, we would argue, not a paradigm shift. For example, the traditional cohort-component method remains the basis of the projections, but with the addition of a stochastic dimension. The stochastic approach builds on a large body of published work from respected demographers and statisticians worldwide. The publication of stochastic projections (Statistics NZ 2012a, 2012b, 2014, 2015a, 2015b, 2016; United Nations 2014) shows that they are practical to produce, and once developed the production costs are similar to conventional methods.

These are important messages for other producers of projections, or for users of projections looking to influence producers of projections. Others can leverage off stochastic developments in New Zealand and elsewhere. Our experience shows that a stochastic approach can be applied to selected projection types while maintaining consistency with other projections produced deterministically. Projection assumptions can be formulated using time series or other empirical models and/or using expert-based approaches. In addition, Statistics NZ continues to publish hypothetical 'what if?' scenarios which are useful in illustrating the effect of specific fertility, mortality, and migration assumptions on population size and structure. A blend of deterministic and stochastic methods is therefore a pragmatic approach for producers of projections, and useful from a user's perspective.

Statistics NZ is aiming to apply stochastic approaches to its other demographic projections – of families and households, and of subnational populations – with the common rationale of increasing the interpretability and usefulness of the projections. Extending the stochastic approach is not without further challenges, and may require different modelling approaches to that used already, such as the use of Bayesian modelling for subnational projections (Bryant and Graham 2013; Bijak and Bryant 2016).

We should expect to see more countries and agencies producing stochastic projections in future. Partly this will be driven by users of projections wanting more informative indications of future demographic change, given the importance of projections for planning and decision making. Partly this will be driven by practical solutions, assisted by technology, to overcome the challenges of adopting a stochastic approach.

More generally, we can expect a growing integration of statistical methods into demographic projections (and other applications) in New Zealand and internationally. This collaboration should be encouraged by producers and users of projections, as the combination of statistical knowhow with demographic knowledge should strengthen the value of projections. In particular, users can have increased confidence in projections which have a strong statistical and data-driven basis, yet remain plausible and interpretable. The development of global stochastic population projections by the United Nations Population Division is likely to add momentum to their uptake.

## 6. Conclusion

From a user's perspective, stochastic projections have been a welcome development. Conventional deterministic projections were poor at conveying the inherent uncertainty of future changes. By quantifying uncertainty, stochastic projections assist interpretation. They clarify which demographic trends are probable and which are improbable. By giving users some quantification of the probability of an event occurring, they can make the case for intervention much sharper. Importantly, there has been no negative reaction from users to the adoption of a stochastic approach in the release of official projections since 2012.

In addition to these benefits, the stochastic projections are inspiring users to quantify additional components of uncertainty. For example, Ball et al. (2015) explore uncertainty in economic parameters such as interest rates, productivity growth, and government expenditure. Such work might be expected to flow through to products dependent on demographic projections, such as long-term fiscal projections, to fully convey the distribution of uncertainty.

The following recommendations provide guidance for producers of projections considering adopting a stochastic approach:

1. Engage with users of projections. Do they understand and use alternative deterministic projections, such as 'low' and 'high' variants? Would they benefit from more informed measures of uncertainty in projections?
2. Identify institutional barriers to adopting a stochastic approach.
3. Look for opportunities to collaborate with other organisations, researchers and academics on using and developing stochastic methods.
4. Utilise existing open-source software to produce stochastic projections in testing and productions.
5. Utilise published examples of how stochastic projections are disseminated in terms of metadata (explanatory information) and data (projection results).
6. Consider producing 'what if?' deterministic scenarios to complement the principal stochastic projections to illustrate specific demographic scenarios.
7. Consider applying a stochastic approach to selected projection types (e.g., national population) before extending further. A progressive development allows projection types to be tackled in order of complexity, and can also help manage user expectations.

From a producer's perspective, stochastic projections are meeting a user need. Increasingly, producers are more focussed on measuring and conveying uncertainty, and less on conveying overly precise point estimates. There are few practical obstacles to producing stochastic population projections, other than the additional resources required to formulate measures of uncertainty and produce multiple simulations. Moreover, given the nature of projections, a stochastic approach is consistent with how statistical agencies would like projections to be conveyed and interpreted.

## 7. References


Abel, G., J. Bijak, J. Forster, J. Raymer, P. Smith, and J. Wong. 2013. "Integrating Uncertainty in Time Series Population Forecasts: An Illustration Using a Simple Projection Model." *Demographic Research* 29: 1187–1226. Doi: http://dx.doi.org/10.4054/DemRes.2013.29.43.

Alho, J. 1997. "Scenarios, Uncertainty and Conditional Forecasts of the World Population." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160: 71–85. Doi: http://dx.doi.org/10.1111/1467-985X.00046.

Alho, J. 2005. "Remarks on the Use of Probabilities in Demography and Forecasting." In *Perspectives on Mortality Forecasting: Probabilistic Models*, edited by N. Keilman. Social Insurance Studies No. 2, Swedish Social Insurance Agency. Available at: www.forsakringskassan.se/wps/wcm/connect/95496cc1-ada3-41a6-af4c-02b4b0b3c233/social_insurance_studies_2.pdf?MOD=AJPERES (accessed 19 May 2016).

Alkema, L., A. Raftery, P. Gerland, S. Clark, F. Pelletier, T. Buettner, and G. Heilig. 2011. "Probabilistic Projections of the Total Fertility Rate for All Countries." *Demography* 48: 815–839. Doi: http://dx.doi.org/10.1007/s13524-011-0040-5.

Ball, C., J. Creedy, and G. Scobie. 2015. *Long-Run Fiscal Projections Under Uncertainty: The Case of New Zealand*. New Zealand Treasury Working Paper 15/10. Available at: www.treasury.govt.nz (accessed 19 May 2016).

Bijak, J., I. Alberts, J. Alho, J. Bryant, T. Buettner, J. Falkingham, J. Forster, P. Gerland, T. King, L. Onorante, N. Keilman, A. O'Hagan, D. Owens, A. Raftery, H. Sevcikova, and P. Smith. 2015. "Letter to the Editor: Probabilistic Population Forecasts for Informed Decision Making." *Journal of Official Statistics* 31: 537–544. Doi: http://dx.doi.org/10.1515/jos-2015-0033.

Bijak, J. and J. Bryant. 2016. "Bayesian Demography 250 Years after Bayes." *Population Studies* 70: 1–19. Doi: http://dx.doi.org/10.1080/00324728.2015.1122826.

Billari, F., R. Graziani, and E. Melilli. 2014. "Stochastic Population Forecasting Based on Combinations of Expert Evaluations Within the Bayesian Paradigm." *Demography* 51: 1933–1954. Doi: http://dx.doi.org/10.1007/s13524-014-0318-5.

Bongaarts, J. and R. Bulatao. 2000. *Beyond Six Billion: Forecasting the World's Population*. Washington, DC: National Academy Press. Doi: http://dx.doi.org/10.17226/9828.

Booth, H. 2006. "Demographic Forecasting: 1980 to 2005 in Review." *International Journal of Forecasting* 22: 547–581. Doi: http://dx.doi.org/10.1016/j.ijforecast.2006.04.001.

Booth, H. and L. Tickle. 2008. "Mortality Modelling and Forecasting: a Review of Methods." *Annals of Actuarial Science* 3: 3–43. Doi: http://dx.doi.org/10.1017/S1748499500000440.

Bryant, J. 2003. "Can Population Projections Be Used for Sensitivity Tests on Policy Models?" New Zealand Treasury Working Paper 03/07. Available at: www.treasury.govt.nz/publications/research-policy/wp/2003/03-07 (accessed 19 May 2016).

Bryant, J. 2005. "What Can Stochastic Population Projections Contribute to Policy Analysis?" *New Zealand Population Review* 31: 1–11. Available at: www.population.org.nz/wp-content/uploads/2010/01/nzpr-vol-31-1_bryant.pdf (accessed 19 May 2016).

Bryant, J. and P. Graham. 2013. "Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources." *Bayesian Analysis* 8: 591–622. Doi: http://dx.doi.org/10.1214/13-BA820.

De Beer, J. 2000. *"Dealing with Uncertainty in Population Forecasting"*. Statistics Netherlands. Available at: www.cbs.nl/nr/rdonlyres/7dc466f9-fe4c-48dc-be90-30216b697548/0/dealingwithuncertainty.pdf (accessed 19 May 2016).

Dowd, K., D. Blake, and A. Cairns. 2010. "Facing Up to Uncertain Life Expectancy: The Longevity Fan." *Demography* 47: 67–78. Doi: http://dx.doi.org/10.1353/dem.0.0083.

Dunstan, K. 2011. "Experimental Stochastic Population Projections for New Zealand: 2009(base)–2111." Statistics New Zealand Working Paper 11–01. Available at: http://www.stats.govt.nz/methods/research-papers/working-papers-original/experimental-stochastic-pop-projections-nz-11-01.aspx (accessed 19 May 2016).

Hyndman, R. 2015. "Demography." Available at: http://robjhyndman.com/software/demography/ (accessed 19 May 2016).

Keilman, N. 1991. "Uncertainty in National Population Forecasting: Issues, Backgrounds, Analyses, Recommendations." *International Journal of Forecasting* 7: 392–393. Doi: http://dx.doi.org/10.1016/0169-2070(91)90018-Q.

Keilman, N. 1997. "Ex-Post Errors in Official Population Forecasts in Industrialized Countries." *Journal of Official Statistics* 13: 245–277. Available at: www.jos.nu/Articles/abstract.asp?article=133245 (accessed 19 May 2016).

Lee, R. 1998. "Probabilistic Approaches to Population Forecasting." In *Frontiers of Population Forecasting*, edited by W. Lutz, J. Vaupel, and D. Ahlburg. 156–190. New York: The Population Council.

Lee, R. and L. Carter. 1992. "Modeling and Forecasting US Mortality." *Journal of the American Statistical Association* 87: 659–671. Doi: http://dx.doi.org/10.2307/2808055.

Long, J. and F. Hollman. 2004. "Developing Official Stochastic Population Forecasts at the US Census Bureau." *International Statistical Review* 72: 201–208. Available at: www.jstor.org/stable/1403854 (accessed 19 May 2016).

Lutz, W. 2009. *Toward a Systematic, Argument-Based Approach to Defining Assumptions for Population Projections*. Interim Report IR-09-037, International Institute for Applied Systems Analysis, Austria. Available at: http://pure.iiasa.ac.at/9115/1/IR-09-037.pdf (accessed 19 May 2016).

Lutz, W., W. Butz, and S. KC. 2014. *World Population and Human Capital in the 21st Century*. Oxford University Press. Doi: http://dx.doi.org/10.1111/j.1728-4457.2015.00032.x.

Lutz, W. and S. Scherbov. 1998. "An Expert-Based Framework for Probabilistic National Population Projections: The Example of Austria." *European Journal of Population* 14: 1–17. Doi: http://dx.doi.org/10.1023/A:1006040321755.

New Zealand Treasury. 2013. "Affording Our Future: Statement on New Zealand's Long-Term Fiscal Position." Available at: http://www.treasury.govt.nz/government/longterm/fiscalposition/2013/affordingourfuture (accessed 19 May 2016).

OECD. 2013. "OECD.Stat Web Browser User Guide". Available at: https://stats.oecd.org/Content/themes/OECD/static/help/WBOS%20User%20Guide%20(EN) (accessed 25 October 2016).

Raftery, A. 2014. "Use and Communication of Probabilistic Forecasts." Available at: http://arxiv.org/abs/1408.4812 (accessed 19 May 2016).

Raftery, A., J. Chunn, P. Gerland, and H. Sevcikova. 2013. "Bayesian Probabilistic Projections of Life Expectancy for All Countries." *Demography* 50: 777–801. Doi: http://dx.doi.org/10.1007/s13524-012-0193-x.

Raftery, A., N. Li, H. Sevcikova, P. Gerland, and G. Heilig. 2012. "Bayesian Probabilistic Population Projections for All Countries." In Proceedings of the National Academy of Sciences 109(35): 13915–13921. Available at: www.pnas.org/content/109/35/13915.full.pdf+html (accessed 19 May 2016).

Sevcikova, H., L. Alkema, and A. Raftery. 2011. "bayesTFR: An R Package for Probabilistic Projections of the Total Fertility Rate." *Journal of Statistical Software* 43: 1–29. Doi: http://dx.doi.org/10.18637/jss.v043.i01.

Statistics NZ. 2012a. "National Labour Force Projections: 2006(base)–2061 (August 2012 update)." Available at: http://www.stats.govt.nz/browse_for_stats/population/estimates_and_projections/NationalLabourForceProjections_HOTP06-61Augupdate.aspx (accessed 19 May 2016).

Statistics NZ. 2012b. "National Population Projections: 2011(base)–2061.2." Available at: http://www.stats.govt.nz/browse_for_stats/population/estimates_and_projections/NationalPopulationProjections_HOTP2011.aspx (accessed 19 May 2016).

Statistics NZ. 2012c. "NZ.Stat user guide." Available at: http://nzdotstat.stats.govt.nz/wbos/ModalHelp/OECD/WBOS%20User%20Guide%20(EN).PDF (accessed 19 May 2016).

Statistics NZ. 2014. "National Population Projections: 2014(base)–2068." Available at: http://www.stats.govt.nz/browse_for_stats/population/estimates_and_projections/NationalPopulationProjections_HOTP2014.aspx (accessed 19 May 2016).

Statistics NZ. 2015a. "National Ethnic Population Projections: 2013(base)–2038." Available at: http://www.stats.govt.nz/browse_for_stats/population/estimates_and_projections/NationalEthnicPopulationProjections_HOTP2013-38.aspx (accessed 19 May 2016).

Statistics NZ. 2015b. "National Labour Force Projections: 2015(base)–2068". Available at: http://www.stats.govt.nz/browse_for_stats/population/estimates_and_projections/NationalLabourForceProjections_HOTP15-68.aspx (accessed 19 May 2016).

Statistics NZ. 2015c. "Statement of Strategic Intentions for Statistics New Zealand 2015–19." Available at: http://www.stats.govt.nz/about_us/what-we-do/our-publications/statement-of-intent/statement-of-intent-2015.aspx (accessed 25 October 2016).

Statistics NZ. 2016. "National Population Projections: 2016(base)–2068." Available at: http://www.stats.govt.nz/browse_for_stats/population/estimates_and_projections/NationalPopulationProjections_HOTP2016.aspx (accessed 25 October 2016).

Stoto, M. 1983. "The Accuracy of Population Projections." *Journal of the American Statistical Association* 78: 13–20. Doi: http://dx.doi.org/10.1080/01621459.1983.10477916.

United Nations. 2014. "World Population Prospects: The 2012 Revision, Methodology of the United Nations Population Estimates and Projections." Working Paper No. ESA/P/Wp. 235. Available at: https://esa.un.org/unpd/wpp/publications/Files/WPP2012_Methodology.pdf (accessed 25 October 2016).

Wisniowski, A., P. Smith, J. Bijak, J. Raymer, and J. Forster. 2015. "Bayesian Population Forecasting: Extending the Lee-Carter Method." *Demography* 53: 1035–1059. Doi: http://dx.doi.org/10.1007/s13524-015-0389-y.

Woods, C. and K. Dunstan. 2014. "Forecasting Mortality in New Zealand: A New Approach for Population Projections Using a Coherent Functional Demographic Model." (Statistics New Zealand Working Paper No 14–01). Available at: http://www.stats.govt.nz/methods/research-papers/working-papers-original/forecasting-mortality-14-01.aspx (accessed 19 May 2016).

Received May 2015
Revised June 2016
Accepted August 2016

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 963–986, http://dx.doi.org/10.1515/JOS-2016-0051

# Small-Area Estimation with Zero-Inflated Data – a Simulation Study

*Sabine Krieg*[1], *Harm Jan Boonstra*[1], *and Marc Smeets*[1]

Many target variables in official statistics follow a semicontinuous distribution with a mixture of zeros and continuously distributed positive values. Such variables are called zero inflated. When reliable estimates for subpopulations with small sample sizes are required, model-based small-area estimators can be used, which improve the accuracy of the estimates by borrowing information from other subpopulations. In this article, three small-area estimators are investigated. The first estimator is the EBLUP, which can be considered the most common small-area estimator and is based on a linear mixed model that assumes normal distributions. Therefore, the EBLUP is model misspecified in the case of zero-inflated variables. The other two small-area estimators are based on a model that takes zero inflation explicitly into account. Both the Bayesian and the frequentist approach are considered. These small-area estimators are compared with each other and with design-based estimation in a simulation study with zero-inflated target variables. Both a simulation with artificial data and a simulation with real data from the Dutch Household Budget Survey are carried out. It is found that the small-area estimators improve the accuracy compared to the design-based estimator. The amount of improvement strongly depends on the properties of the population and the subpopulations of interest.

*Key words:* Generalized linear mixed model; EBLUP; MCMC; Logit; Dutch Household Budget Survey.

## 1. Introduction

Traditionally, national statistical institutes (NSIs) such as Statistics Netherlands prefer design-based estimation methods, since these methods lead to approximately design-unbiased estimates. However, the demand for detailed estimates for subpopulations is increasing, while at the same time budgets are under continuous pressure. Therefore, several NSIs started to investigate the possibilities of small-area estimation (SAE), see, for example Eurarea (2004) and Boonstra et al. (2008). This model-based methodology is developed for situations where the sample sizes of the subpopulations (often called domains or areas in the SAE context) or time periods are too small to compute reliable estimates based on design-based methods. An SAE method borrows information from other domains or from other time periods to improve the accuracy of the domain estimates.

The most common SAE estimator is the Empirical Best Linear Unbiased Predictor (EBLUP) (Battese et al. 1988; Rao 2003). The EBLUP is based on a linear mixed model and assumes normal distributions. However, NSIs often have to deal with non-normally distributed data, for which the EBLUP may yield seriously biased estimates. For such

[1] Statistics Netherlands, Postbus 4481, 6401CZ Heerlen, The Netherlands. Emails: skrg@cbs.nl, hbta@cbs.nl, and mset@cbs.nl

situations, different adjustments of the EBLUP and some new SAE methods have been developed in recent years. For example, the robust EBLUP (Sinha and Rao 2009) reduces the influence of outliers in the data. Chandra and Chambers (2011b) developed an estimator for skewly distributed data, and the M-quantile estimator (Chambers and Tzavidis 2006) does not make any assumptions about the distribution.

This article deals with the estimation for variables that are zero for a substantial part of the population. This type of data is also called zero-inflated data. Pfeffermann et al. (2008) and Chandra and Sud (2012) developed an estimator for such kinds of data, the first using a Bayesian approach and the second a frequentist approach. The estimator is based on two models, the first being a linear mixed model for the nonzero values and the second a generalized linear mixed model for the binary zero indicator. Both the Bayesian and the frequentist approaches are used in this article, with a small simplification of the method used in Pfeffermann et al. (2008). The SAE method for zero-inflated data is compared with the EBLUP and with a design-based method (the survey regression estimator). In the first part of the article, a simulation with artificial data is carried out in which different populations are created to investigate the properties of the considered estimators in different situations. This simulation shows to what extent the model misspecification of the EBLUP increases the bias of the estimates and to what extent the accuracy of the estimates is improved when the estimators of Pfeffermann et al. (2008) and Chandra and Sud (2012) are applied instead. In a second simulation, the estimators are applied to real zero-inflated data of the Dutch Household Budget Survey (HBS). The HBS measures the consumption expenditures of Dutch households. Many target variables which describe the expenditures for different products, are zero inflated.

In Section 2 the considered methods are described. Then the results of the simulation with artificial populations are discussed in Section 3. The results of the simulation for the HBS follow in Section 4. In Section 5 the conclusions are given.

## 2. Methods

### 2.1. *Notation*

The finite population $U$ with $N$ elements is divided into $m$ subpopulations or domains. A sample with $n$ elements is drawn using simple random sampling without replacement. The observed value of the target variable for unit $i$ in domain $j$ is given by $y_{ij}$. The total sample and population size in domain $j$ are denoted by $n_j$ and $N_j$, respectively. The total sample is called $S$ and the sample in domain $j$ is called $S_j$.

The explanatory variables for unit $i$ in domain $j$ are given by the vector $\mathbf{x}_{ij} = \left(x_{ij}^1, \ldots, x_{ij}^p\right)^{\mathrm{t}}$. An intercept is always included, that is, it can be assumed that $x_{ij}^1 = 1$. Population means $Y_j^{\text{mean}} = \frac{1}{N_j}\sum_{i=1}^{N_j} y_{ij}$ for target variable $y$ for all domains $j = 1, \ldots, m$ have to be estimated.

The target variable $y_{ij}$ is equal to zero for a substantial part of the population. We define

$$\delta_{ij} = \begin{cases} 1 & \text{if} \quad y_{ij} \neq 0 \\ 0 & \text{if} \quad y_{ij} = 0. \end{cases} \tag{1}$$

The subscript nz is used to denote the nonzero part of the population or sample.

### 2.2. *Survey Regression*

Survey regression (SR) is a design-based model-assisted estimator which is approximately design unbiased (Woodruff 1966; Battese et al. 1988; Särndal et al. 1992). In this article the SR is considered to be the reference estimator; the model-based methods are expected to be more accurate than the SR. The SR of the unknown population mean $Y_j^{\text{mean}}$ for domain $j$ is given by

$$\hat{Y}_j^{\text{SR}} = \hat{Y}_j^{\text{HT}} + \left(\mathbf{X}_j^{\text{mean}} - \hat{\mathbf{X}}_j^{\text{HT}}\right)^{\text{t}} \hat{\beta}, \tag{2}$$

where

$$\hat{\beta} = (\mathbf{X}^{\text{t}}\mathbf{X})^{-1}\mathbf{X}^{\text{t}}\mathbf{y}.$$

Here the Horvitz-Thompson estimators are given by $\hat{Y}_j^{\text{HT}} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ and $\hat{\mathbf{X}}_j^{\text{HT}} = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{x}_{ij}$. Furthermore, $\mathbf{X}_j^{\text{mean}} = \frac{1}{N_j}\sum_{i=1}^{N_j}\mathbf{x}_{ij}$ is the $p$-vector of population means of the auxiliary information in domain $j$, $\mathbf{y} = (y_{11}, \ldots, y_{n_1 1}, y_{12} \ldots, y_{n_m m})^{\text{t}}$ and $\mathbf{X} = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{n_1 1}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{n_m m})^{\text{t}}$.

### 2.3. *Empirical Best Linear Unbiased Predictor (EBLUP)*

Consider the linear mixed model given by

$$y_{ij} = \mathbf{x}_{ij}^{\text{t}}\beta + \vartheta_j + e_{ij}, \quad \text{for} \quad j = 1, \ldots, m \quad \text{and} \quad i = 1, \ldots, N_j, \tag{3}$$

where

$$\vartheta_j \sim \mathcal{N}\left(0, \sigma_r^2\right), \; e_{ij} \sim \mathcal{N}\left(0, \sigma_e^2\right).$$

Here $\sigma_e^2$ is the within-area variance parameter, whereas $\sigma_r^2$ is the between-domain variance.

Based on Model (3), the EBLUP (Rao 2003) is considered to estimate the population means $Y_j^{\text{mean}}$ for the domains $j = 1, \ldots, m$. The estimator for $Y_j^{\text{mean}}$ is then given by

$$\hat{Y}_j^{\text{EBLUP}} = \mathbf{X}_j^{\text{mean}}\hat{\beta} + \hat{\vartheta}_j. \tag{4}$$

Expressions for $\hat{\beta}$ and $\hat{\vartheta}_j$ can be found in Rao 2003, sec. 7.2. The variance parameters $\sigma_r^2$ and $\sigma_e^2$ are estimated by the method of Restricted Maximum Likelihood (REML).

A refined version of (4) would use predicted values only for the nonsampled part of the population, and the observed values for themselves. However, when sampling fractions are small, the difference is negligible and for that reason (4) is used in this article. The EBLUP estimator is computed with R (R Development Core Team 2009), where the function `lmer` of package `lme4` (Bates et al. 2015) is used to fit the linear mixed model.

### 2.4. *A Small-Area Estimator for Zero-Inflated Data*

In this section, an estimator is described that takes the zero inflation into account. There are two approaches to estimate the models: the frequentist approach (Subsection 2.4.1), described by Chandra and Sud (2012), and the Bayesian approach (Subsection 2.4.2), described by Pfeffermann et al. (2008). For both approaches we use the abbreviation

ZERO in the rest of the article, or ZERO-F or ZERO-B to make clear which approach is used. The theoretical properties of the estimators are discussed in Pfeffermann et al. (2008) and Chandra and Sud (2012).

Note that an important disadvantage of ZERO compared with the EBLUP is that ZERO can only be applied if the auxiliary information is known for all elements in the population.

### 2.4.1. The Frequentist Approach

The target variable $y_{ij}$ is assumed to be the product of an underlying normally distributed variable $y_{ij}^*$ and $\delta_{ij}$, that is $y_{ij} = y_{ij}^*\delta_{ij}$. These two variables are modelled in two different (generalized) linear mixed models. The first model describes the distribution of $y_{ij}^*$:

$$y_{ij}^* = \mathbf{x}_{\mathrm{nz},ij}^{\mathrm{t}}\beta_{\mathrm{nz}} + \vartheta_{\mathrm{nz},j} + e_{ij}, \quad \text{for} \quad j = 1, \ldots, m \quad \text{and} \quad i = 1, \ldots, N_j, \tag{5}$$

where

$$\vartheta_{\mathrm{nz},j} \sim \mathcal{N}\left(0, \sigma_{r,\mathrm{nz}}^2\right), \; e_{ij} \sim \mathcal{N}\left(0, \sigma_{e,\mathrm{nz}}^2\right).$$

The second model describes the probabilities $p_{ij} = P(\delta_{ij} = 1) = P(y_{ij} \neq 0)$ of the target variable to be nonzero:

$$\mathrm{logit}(p_{ij}) = \ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \mathbf{x}_{\mathrm{z},ij}^{\mathrm{t}}\beta_{\mathrm{z}} + \vartheta_{\mathrm{z},j}, \quad \text{for} \quad j = 1, \ldots, m \quad \text{and} \quad i = 1, \ldots, N_j, \tag{6}$$

with

$$\vartheta_{\mathrm{z},j} \sim \mathcal{N}\left(0, \sigma_{r,\mathrm{z}}^2\right).$$

Model (5) is estimated based on the nonzero part of the sample, Model (6) is estimated based on the complete sample, resulting in the estimates $\hat{\beta}_{\mathrm{nz}}$, $\hat{\vartheta}_{\mathrm{nz},j}$, $\hat{\beta}_{\mathrm{z}}$, $\hat{\vartheta}_{\mathrm{z},j}$ for the location parameters and in estimates $\hat{\sigma}_{r,\mathrm{nz}}$, $\hat{\sigma}_{e,\mathrm{nz}}$, $\hat{\sigma}_{r,\mathrm{z}}$ for the variance parameters.

Based on these estimates, $y_{ij}^*$ and $p_{ij}$ are estimated for all elements in the population:

$$\hat{y}_{ij}^* = \mathbf{x}_{\mathrm{nz},ij}^{\mathrm{t}}\hat{\beta}_{\mathrm{nz}} + \hat{\vartheta}_{\mathrm{nz},j}, \tag{7}$$

$$\hat{p}_{ij} = \frac{\exp\left(\mathbf{x}_{\mathrm{z},ij}^{\mathrm{t}}\hat{\beta}_{\mathrm{z}} + \hat{\vartheta}_{\mathrm{z},j}\right)}{1 + \exp\left(\mathbf{x}_{\mathrm{z},ij}^{\mathrm{t}}\hat{\beta}_{\mathrm{z}} + \hat{\vartheta}_{\mathrm{z},j}\right)}. \tag{8}$$

The estimate for $y_{ij}$ is then taken to be the product $\hat{y}_{ij} = \hat{y}_{ij}^*\hat{p}_{ij}$, and the mean for domain $j$ can be estimated as

$$\hat{Y}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} \hat{y}_{ij}^*\hat{p}_{ij}. \tag{9}$$

Note that the model for $y^*$ can only be fitted using the nonzero observations, whereas it is applied to predict all population elements, zero or nonzero. In order to reduce the risk of bias, it is therefore important to also include variables that predict $\delta_{ij}$ in the model for $y^*$. In this article we always use the same predictors $\mathbf{x}$ in both models.

Again, for convenience, the prediction in (9) is used for all population elements, including the ones observed. The mixed models can be estimated using the function `lmer` of R-package `lme4`. Within this function, the `family` parameter is taken to be `binomial(link = "logit")` for Model (6) and gaussian for Model (5).

Chandra and Sud (2012) proposed parametric bootstrapping for the estimation of the mean squared error.

### 2.4.2. The Bayesian Approach

The two Models (5) and (6) can also be estimated in a Bayesian fashion using a Markov Chain Monte Carlo (MCMC) simulation. Such a simulation results in a series of draws of parameters from their joint posterior distribution given the data. An important advantage of the Bayesian MCMC approach is that the draws can be used both for computing point estimates and for measures of accuracy, including interval estimates. Parametric bootstrapping, as proposed by Chandra and Sud (2012) for the frequentist approach, is less easily available in R software packages.

The MCMC simulation is carried out over $R$ runs. The first part of the MCMC simulation (burnin) is not used, as it depends too strongly on the starting values. Moreover, only every $l$th run is retained to save memory and increase the effective number of independent draws. In the end, $r$ runs are retained for further analysis. Both $R$ and $r$ have to be chosen sufficiently large so that the Markov chain can converge and explore the entire distribution. There is no reason that the number of retained runs $r_{\mathrm{z}}$ and $r_{\mathrm{nz}}$ has to be equal for the two Models (5) and (6) to achieve this goal. Equality $r = r_{\mathrm{z}} = r_{\mathrm{nz}}$ is necessary for the computation of model estimates for $Y_j$. In all MCMC simulations carried out for this article we have taken $R = 40{,}000$ runs with a burnin of 20,000 and thinning by retaining each 20th iteration, so that $r = 1{,}000$ draws are retained for posterior analysis. From inspection of trace plots and autocorrelations, these numbers were seen to be adequate.

From the parameter draws obtained for both MCMC simulations, posterior draws for the small-area quantities of interest can be computed by simulating from the posterior predictive distributions:

1. Draw residuals $e_{ij,\rho} \sim \mathcal{N}\left(0, \sigma^2_{e,\mathrm{nz},\rho}\right)$ independently for all population units $i, j$ and for each MCMC iteration $\rho = 1, \ldots, r$, and form posterior predictions

$$y^*_{ij,\rho} = \mathbf{x}^{\mathrm{t}}_{\mathrm{nz},ij}\beta_{\mathrm{nz},\rho} + \vartheta_{\mathrm{nz},j,\rho} + e_{ij,\rho}.$$

   All parameter draws $\sigma_{e,\mathrm{nz},\rho}$, $\beta_{\mathrm{nz},\rho}$, $\vartheta_{\mathrm{nz},j,\rho}$ are part of the MCMC simulation output.
2. Similarly, draw zero indicators independently from the Bernoulli distribution according to

$$\delta^*_{ij,\rho} \sim \mathrm{Be}(p_{ij,\rho}),$$

with probability of a nonzero response value

$$p_{ij,\rho} = \frac{\exp\left(\mathbf{x}^{t}_{z,ij}\beta_{z,\rho} + \vartheta_{z,j,\rho}\right)}{1 + \exp\left(\mathbf{x}^{t}_{z,ij}\beta_{z,\rho} + \vartheta_{z,j,\rho}\right)}.$$

3. Combine the posterior predictive draws to obtain posterior draws for the small-area estimands,

$$Y^{*}_{j,\rho} = \frac{1}{N_j}\sum_{i=1}^{N_j} y^{*}_{ij,\rho}\,\delta^{*}_{ij,\rho}. \tag{10}$$

Estimates for the domain means of interest are now obtained as MCMC approximations of the posterior means, that is,

$$\hat{Y}_{j,\text{mcmc}} = \frac{1}{r}\sum_{\rho=1}^{r} Y^{*}_{j,\rho}.$$

The mean squared error of $\hat{Y}_{j,\text{mcmc}}$ under the model, that is, the posterior variance, is approximated by

$$\text{mse}(\hat{Y}_{j,\text{mcmc}}) = \frac{1}{r}\sum_{\rho=1}^{r}\left(Y^{*}_{j,\rho} - \hat{Y}_{j,\text{mcmc}}\right)^{2}. \tag{11}$$

Credible intervals are also considered. In particular, highest posterior 95% intervals have been computed using the R package `coda` (Plummer et al. 2006).

The MCMC simulations have been carried out using the function `MCMCglmm` from the R package of the same name (Hadfield 2010), which supports both models by way of Gibbs sampling (Geman and Geman 1984; Gelfand and Smith 1990). We use weakly informative default priors as implemented in `MCMCglmm` for the coefficients and variance parameters in both models. In particular, the regression coefficients in both models are assigned normal priors with zero mean and very large variance. Following Gelman (2006), we use parameter-expanded inverse-chi-squared priors for the random effect variances in both models, implying half-Cauchy priors on the standard-deviation parameters. The scales of the half-Cauchy priors are taken to be 25, larger than the scale of the response variable in both models. The half-Cauchy priors are more robust than inverse chi-squared priors and their parameter-expansion representation also improves convergence and mixing of the Gibbs sampler, especially in situations with relatively small random effect variances (Gelman et al. 2008). For the residual variance of Model (5), a default noninformative prior $p\left(\sigma^2_{e,\text{nz}}\right) \propto 1/\sigma^2_{e,\text{nz}}$ is used.

### 2.4.3. Correlated Random Effects

In Pfeffermann et al. (2008), a single two-part model is used that allows for correlations between the random effects of the two submodels. It is possible that such a model would better fit the data. For this article we have chosen to use the somewhat simpler model in which components are treated independently. The main reason for this simplification is that the separate models can be fit using relatively fast and standard functions in R. In an example, Pfeffermann et al. (2008) showed that taking the correlation into account only slightly improved the accuracy of the estimates.

## 3. Simulation with Artificial Populations

### 3.1. Lay-Out of the Simulation

To investigate the properties of the ZERO and to compare it with the SR and the EBLUP, a simulation with artificial populations is carried out. From the artificial populations, samples are drawn repeatedly. Based on these samples, the SR, EBLUP, and ZERO are computed. In most cases, only the frequentist approach (ZERO-F) is used because the MCMC simulation (ZERO-B) takes much more computation time. This choice makes it possible to simulate many different situations. In a small part of the investigated situations, the MCMC approach is also applied and both approaches are compared.

We start with the description of the main part of the simulations with artificial populations. The artificial populations consist of $m = 50$ domains with $N = 60{,}000$ elements. The domains are not equally sized. The domain size increases from 30 for the first five domains up to 3,250 for the last domain.

The creation of the artificial populations starts with drawing an auxiliary variable $x$ from the normal distribution $\mathcal{N}(2,2.25)$. The mean of the auxiliary variable is then more or less equal for all domains. This is not realistic. To get an idea of the consequences of unequal means of the auxiliary variable, the value of the 0.9-quantile of the vector $x$ is added for one randomly chosen domain. This is not realistic either, but it makes it easier to analyze the effects of such a deviation. The random effects $\vartheta_{\mathrm{nz},j}$ and $\vartheta_{\mathrm{z},j}$ for the domains $j = 1, \ldots, m$ are independently distributed following $\mathcal{N}\left(0, \sigma^2_{r,\mathrm{nz}}\right)$ and $\mathcal{N}\left(0, \sigma^2_{r,\mathrm{z}}\right)$. The target variable is then computed as $y_{ij} = y^*_{ij}\delta_{ij}$, where $y^*$ and $\delta$ are generated according to Models (5) and (6) and $\delta_{ij} \sim \mathrm{Be}(p_{ij})$ is Bernoulli distributed taking value 1 with probability $p_{ij}$. Model (6) is extended with residuals $e_{ij,\mathrm{z}} \sim \mathcal{N}\left(0, \sigma^2_{e,\mathrm{z}}\right)$. In both models the vector of covariates consists of two components, the intercept and the generated auxiliary variable $x$. The corresponding coefficients will be referred to as $\beta_{0,\mathrm{nz}}$, $\beta_{1,\mathrm{nz}}$, $\beta_{0,\mathrm{z}}$, $\beta_{1,\mathrm{z}}$ with subscripts 0 and 1 corresponding to the intercept and $x$, respectively.

With different choices for $\beta_{0,\mathrm{nz}}$, $\beta_{1,\mathrm{nz}}$, $\beta_{0,\mathrm{z}}$, $\beta_{1,\mathrm{z}}$, $\sigma^2_{r,\mathrm{nz}}$, $\sigma^2_{r,\mathrm{z}}$, $\sigma^2_{e,\mathrm{nz}}$, $\sigma^2_{e,\mathrm{z}}$ different types of populations can be created. In total, 48 situations based on different parameter sets are investigated. The parameters are chosen in such a way that populations with a wide range of properties are included in the study, with

- a small (around 0.1), medium (around 0.5), or large proportion (around 0.85) of nonzeros by an appropriate choice of $\beta_{0,\mathrm{nz}}$, $\beta_{1,\mathrm{nz}}$, $\beta_{0,\mathrm{z}}$, $\beta_{1,\mathrm{z}}$,
- a small (around 0.2) or large (around 0.7) correlation between the auxiliary variable $x$ and $p$, by an appropriate choice of $\sigma^2_{e,\mathrm{z}}$,
- a small (around 0.3) or large (around 0.7) correlation between the auxiliary variable $x$ and $y^*$, by an appropriate choice of $\sigma^2_{e,\mathrm{nz}}$,
- small or large random effects $\vartheta_{\mathrm{z},j}$ and $\vartheta_{\mathrm{nz},j}$ by an appropriate choice of $\sigma^2_{r,\mathrm{z}}$ and $\sigma^2_{r,\mathrm{nz}}$. In the case of small random effect variances, their frequentist estimates are often zero.

The considered sets of parameters and corresponding types of populations are shown in Table 1.

For each set of parameters, ten different populations are created, and with each population, a simulation with 500 runs is carried out. In each run, a sample of size $n = 2{,}000$ using simple random sampling without replacement is drawn. By creating different populations of each type, coincidences in the populations have less influence. The number of ten populations per set of parameters turns out to be adequate, for the generation of different sets of ten populations consistently gives almost the same properties. At the same time, with 500 runs for each population it is possible to analyze the results for different domains, for example domains with large random effects.

In addition to the simulation with 48 different parameter sets, a few special cases are investigated. First, the simulations with the first four parameter sets are repeated with population and sample sizes that are three times as large for all domains. Second, a correlation of 0.5 and 0.9 between the random effects of the two model parts is added. This is also investigated with the first four parameter sets, with the original population and sample sizes. Third, the simulations with the first four parameter sets are repeated using a Bayesian approach (with independent random effects). Here, only a single population is created, for which a simulation with 1,000 runs is carried out. The frequentist approach is applied to the same 1,000 samples.

In SAE it is sometimes useful to include the domain mean $\bar{x}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_{ij}$ as auxiliary information (Bafumi and Gelman 2006; Neuhaus and McCulloch 2006). It appears that this is also the case for the EBLUP in this application, especially for the domain where the 0.9-quantile of the vector $\mathbf{x}$ is added. Therefore, for the EBLUP $\mathbf{x}_{ij} = (1, x_{ij}, \bar{x}_j)^t$. For the other estimators $\mathbf{x}_{ij} = (1, x_{ij})^t$ is used, as the additional area-level covariate would slightly deteriorate the accuracy of these estimates (results not presented).

### 3.2. Evaluation Measures

The most important quality measure of the estimators is the accuracy measured by the mean squared error (mse). We use the root mse (rmse), computed as

$$\mathrm{rmse}_j = \sqrt{\sum_{q=1}^{\nu} \left(\hat{Y}_{j,q} - Y_j^{\mathrm{mean}}\right)^2 / \nu}, \tag{12}$$

with $Y_j^{\mathrm{mean}}$ the population mean of domain $j$ for the target variable $y$, $\hat{Y}_{j,q}$ the estimate for this population mean based on one of the methods in the $q$th run of the simulation, and $\nu$ the number of runs in the simulation. In the simulation with artificial populations, $\nu = 500$.

The mse is the sum of the variance and the squared bias. In order to further analyze the accuracy of the methods, the standard deviation (root of the variance, sd) and the bias are also discussed. These measures are computed as

$$\mathrm{sd}_j = \sqrt{\sum_{q=1}^{\nu} (\hat{Y}_{j,q} - \bar{Y}_j)^2 / \nu}, \quad \mathrm{bias}_j = \sum_{q=1}^{\nu} \left(Y_j^{\mathrm{mean}} - \hat{Y}_{j,q}\right) / \nu \tag{13}$$

where $\bar{Y}_j = \sum_{q=1}^{\nu} \hat{Y}_{j,q} / \nu$ is the mean of the estimates.

Table 1. Description of the types of populations, model parameters, fractions nonzeros and population means.

| No. | $\beta_{0,z}$ | $\beta_{1,z}$ | $\beta_{0,nz}$ | $\beta_{1,nz}$ | $\sigma_{r,z}$ | $\sigma_{r,nz}$ | $\sigma_{e,z}$ | $\sigma_{e,nz}$ | Fraction nonzeros | Popmean |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | −4 | 2.0 | 10 | 1 | 0.2 | 0.08 | 1.0 | 1 | 0.51 | 6.70 |
| 2 | −4 | 2.0 | 10 | 1 | 2.0 | 0.08 | 1.0 | 1 | 0.50 | 6.63 |
| 3 | −4 | 2.0 | 10 | 1 | 0.2 | 0.80 | 1.0 | 1 | 0.51 | 6.70 |
| 4 | −4 | 2.0 | 10 | 1 | 2.0 | 0.80 | 1.0 | 1 | 0.50 | 6.64 |
| 5 | −4 | 2.0 | 30 | 1 | 0.2 | 0.08 | 1.0 | 5 | 0.51 | 16.87 |
| 6 | −4 | 2.0 | 30 | 1 | 2.0 | 0.08 | 1.0 | 5 | 0.51 | 16.86 |
| 7 | −4 | 2.0 | 30 | 1 | 0.2 | 0.80 | 1.0 | 5 | 0.51 | 16.78 |
| 8 | −4 | 2.0 | 30 | 1 | 2.0 | 0.80 | 1.0 | 5 | 0.51 | 16.80 |
| 9 | −1 | 0.5 | 10 | 1 | 0.2 | 0.08 | 2.0 | 1 | 0.50 | 6.31 |
| 10 | −1 | 0.5 | 10 | 1 | 2.0 | 0.08 | 2.0 | 1 | 0.52 | 6.26 |
| 11 | −1 | 0.5 | 10 | 1 | 0.2 | 0.80 | 2.0 | 1 | 0.51 | 6.27 |
| 12 | −1 | 0.5 | 10 | 1 | 2.0 | 0.80 | 2.0 | 1 | 0.50 | 6.21 |
| 13 | −1 | 0.5 | 30 | 1 | 0.2 | 0.08 | 2.0 | 5 | 0.51 | 16.39 |
| 14 | −1 | 0.5 | 30 | 1 | 2.0 | 0.08 | 2.0 | 5 | 0.51 | 16.35 |
| 15 | −1 | 0.5 | 30 | 1 | 0.2 | 0.80 | 2.0 | 5 | 0.50 | 16.25 |
| 16 | −1 | 0.5 | 30 | 1 | 2.0 | 0.80 | 2.0 | 5 | 0.50 | 16.27 |
| 17 | 0 | 2.0 | 10 | 1 | 0.2 | 0.08 | 0.2 | 1 | 0.88 | 10.86 |
| 18 | 0 | 2.0 | 10 | 1 | 2.0 | 0.08 | 0.2 | 1 | 0.83 | 10.45 |
| 19 | 0 | 2.0 | 10 | 1 | 0.2 | 0.80 | 0.2 | 1 | 0.88 | 10.87 |
| 20 | 0 | 2.0 | 10 | 1 | 2.0 | 0.80 | 0.2 | 1 | 0.85 | 10.47 |
| 21 | 0 | 2.0 | 30 | 1 | 0.2 | 0.08 | 0.2 | 5 | 0.88 | 28.33 |
| 22 | 0 | 2.0 | 30 | 1 | 2.0 | 0.08 | 0.2 | 5 | 0.85 | 27.34 |
| 23 | 0 | 2.0 | 30 | 1 | 0.2 | 0.80 | 0.2 | 5 | 0.88 | 28.42 |
| 24 | 0 | 2.0 | 30 | 1 | 2.0 | 0.80 | 0.2 | 5 | 0.84 | 27.41 |
| 25 | 2 | 0.5 | 10 | 1 | 0.2 | 0.08 | 2.0 | 1 | 0.86 | 10.50 |
| 26 | 2 | 0.5 | 10 | 1 | 2.0 | 0.08 | 2.0 | 1 | 0.81 | 9.90 |
| 27 | 2 | 0.5 | 10 | 1 | 0.2 | 0.80 | 2.0 | 1 | 0.86 | 10.51 |
| 28 | 2 | 0.5 | 10 | 1 | 2.0 | 0.80 | 2.0 | 1 | 0.82 | 9.99 |
| 29 | 2 | 0.5 | 30 | 1 | 0.2 | 0.08 | 2.0 | 5 | 0.86 | 27.83 |
| 30 | 2 | 0.5 | 30 | 1 | 2.0 | 0.08 | 2.0 | 5 | 0.80 | 26.17 |
| 31 | 2 | 0.5 | 30 | 1 | 0.2 | 0.80 | 2.0 | 5 | 0.86 | 27.72 |
| 32 | 2 | 0.5 | 30 | 1 | 2.0 | 0.80 | 2.0 | 5 | 0.81 | 26.05 |
| 33 | −9 | 2.0 | 10 | 1 | 0.3 | 0.10 | 0.5 | 1 | 0.09 | 1.36 |
| 34 | −9 | 2.0 | 10 | 1 | 3.0 | 0.10 | 0.5 | 1 | 0.16 | 2.15 |
| 35 | −9 | 2.0 | 10 | 1 | 0.3 | 1.00 | 0.5 | 1 | 0.10 | 1.37 |
| 36 | −9 | 2.0 | 10 | 1 | 3.0 | 1.00 | 0.5 | 1 | 0.15 | 2.04 |
| 37 | −9 | 2.0 | 30 | 1 | 0.3 | 0.10 | 0.5 | 5 | 0.09 | 3.24 |
| 38 | −9 | 2.0 | 30 | 1 | 3.0 | 0.10 | 0.5 | 5 | 0.15 | 5.09 |
| 39 | −9 | 2.0 | 30 | 1 | 0.3 | 3.00 | 0.5 | 5 | 0.09 | 3.27 |
| 40 | −9 | 2.0 | 30 | 1 | 3.0 | 3.00 | 0.5 | 5 | 0.16 | 5.09 |
| 41 | −6 | 0.6 | 10 | 1 | 0.3 | 0.10 | 2.5 | 1 | 0.07 | 0.95 |
| 42 | −6 | 0.6 | 10 | 1 | 3.0 | 0.10 | 2.5 | 1 | 0.14 | 1.80 |
| 43 | −6 | 0.6 | 10 | 1 | 0.3 | 1.00 | 2.5 | 1 | 0.07 | 0.94 |
| 44 | −6 | 0.6 | 10 | 1 | 3.0 | 1.00 | 2.5 | 1 | 0.15 | 1.83 |
| 45 | −6 | 0.6 | 30 | 1 | 0.3 | 0.10 | 2.5 | 5 | 0.07 | 2.34 |
| 46 | −6 | 0.6 | 30 | 1 | 3.0 | 0.10 | 2.5 | 5 | 0.14 | 4.52 |
| 47 | −6 | 0.6 | 30 | 1 | 0.3 | 3.00 | 2.5 | 5 | 0.07 | 2.37 |
| 48 | −6 | 0.6 | 30 | 1 | 3.0 | 3.00 | 2.5 | 5 | 0.14 | 4.57 |

Since the population size for the first five domains is only 30 and the inclusion probability is $\frac{1}{30}$, empty samples occur regularly for these domains in the simulation. In these runs the SR cannot be computed. In the comparison of the accuracy of the SR with the EBLUP and ZERO-F, the first five domains are therefore ignored. In the other domains, empty samples are very rare but not impossible in the simulation. These runs are ignored in the computation of the abovementioned measures $\mathrm{rmse}_j$, $\mathrm{bias}_j$, and $\mathrm{sd}_j$ for the SR. Since these cases are very rare, this does not disturb the results.

In the simulation with ten populations with the same parameters, the mean of these measures over the ten populations is computed.

### 3.3. *Results*

Table 2 shows the mean absolute bias and mean rmse over the domains and over the ten created populations for the SR, the EBLUP and ZERO-F. In the first six columns of the table, where the SR is compared with the EBLUP and ZERO-F, only domains 6–50 are included, as mentioned in the end of Subsection 3.2. The table shows that in all cases considered, both SAE methods are more accurate than the SR, and ZERO-F is more accurate than the EBLUP. The gain in accuracy strongly depends on the properties of the population. The following points are noticed:

- the SR is generally approximately design unbiased. Small nonzero values are due to the approximate nature of SR's design unbiasedness and to the finite number of simulation runs.
- Both model-based SAE methods are biased. The bias of the EBLUP is generally only slightly larger than the bias of ZERO-F. The model misspecification does not cause a serious bias of the EBLUP.
- Generally, the improvement in accuracy of both SAE methods with respect to the SR is very large in the cases with small $\sigma_{r,\mathrm{z}}$ (odd numbers). In those cases, the rmse is often more than halved by the SAE methods. In the case of large $\sigma_{r,\mathrm{z}}$, the rmse of the SAE methods is usually around ten percent smaller than the rmse of the SR.
- In some cases, the gain in accuracy of ZERO-F with respect to the EBLUP in the five smallest domains is substantially larger than in the other domains. Therefore, it is important to compare the EBLUP and ZERO-F with and without these domains included.
- In many cases, the additional gain in accuracy by using the ZERO-F instead of the EBLUP is only five percent to ten percent.
- Larger gains with ZERO-F instead of the EBLUP are possible in the case of large $\sigma_{r,\mathrm{nz}}$, small $\sigma_{r,\mathrm{z}}$ and a small residual variance $\sigma^2_{e,\mathrm{nz}}$, especially if the nonzero fraction is around 0.5 or 0.85 (number 3,11, 19, 27).
- Larger gains with ZERO-F instead of the EBLUP are also possible in the case of a small residual variance $\sigma^2_{e,\mathrm{z}}$ if the nonzero fraction is around 0.1 or 0.85 (number 17–24, 33–40). This is not surprising as small $\sigma_{e,\mathrm{z}}$ means that Model (6) is almost the true model used to simulate the data. The gain is somewhat larger if the nonzero fraction is around 0.1 than if it is 0.85.
- Altogether, the possible gain with ZERO-F instead of the EBLUP depends only slightly on the nonzero fraction.

*Table 2. Mean absolute bias and mean rmse.*

| | Domains | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| No. | bias 6–50 SR | bias 6–50 EBLUP | bias 6–50 ZERO-F | rmse 6–50 SR | rmse 6–50 EBLUP | rmse 6–50 ZERO-F | rmse 1–50 EBLUP | rmse 1–50 ZERO-F |
| 1 | 0.03 | 0.21 | 0.20 | 0.78 | 0.29 | 0.27 | 0.38 | 0.33 |
| 2 | 0.03 | 0.18 | 0.16 | 0.74 | 0.70 | 0.65 | 0.82 | 0.77 |
| 3 | 0.03 | 0.28 | 0.20 | 0.78 | 0.41 | 0.30 | 0.49 | 0.36 |
| 4 | 0.03 | 0.18 | 0.16 | 0.74 | 0.70 | 0.65 | 0.83 | 0.78 |
| 5 | 0.08 | 0.54 | 0.53 | 2.14 | 0.79 | 0.74 | 0.97 | 0.83 |
| 6 | 0.07 | 0.52 | 0.44 | 2.04 | 1.92 | 1.73 | 2.25 | 2.06 |
| 7 | 0.07 | 0.64 | 0.62 | 2.13 | 0.91 | 0.85 | 1.08 | 0.93 |
| 8 | 0.07 | 0.53 | 0.50 | 2.05 | 1.92 | 1.76 | 2.39 | 2.11 |
| 9 | 0.04 | 0.26 | 0.26 | 1.01 | 0.38 | 0.36 | 0.43 | 0.40 |
| 10 | 0.03 | 0.23 | 0.22 | 0.89 | 0.85 | 0.83 | 1.03 | 0.99 |
| 11 | 0.04 | 0.34 | 0.27 | 1.01 | 0.49 | 0.38 | 0.57 | 0.44 |
| 12 | 0.03 | 0.25 | 0.24 | 0.90 | 0.86 | 0.85 | 1.07 | 1.02 |
| 13 | 0.09 | 0.73 | 0.74 | 2.73 | 1.03 | 0.99 | 1.17 | 1.12 |
| 14 | 0.09 | 0.60 | 0.55 | 2.38 | 2.28 | 2.20 | 2.69 | 2.63 |
| 15 | 0.09 | 0.73 | 0.72 | 2.72 | 1.06 | 0.99 | 1.22 | 1.16 |
| 16 | 0.09 | 0.58 | 0.58 | 2.42 | 2.29 | 2.24 | 2.68 | 2.63 |
| 17 | 0.02 | 0.13 | 0.12 | 0.51 | 0.18 | 0.16 | 0.22 | 0.18 |
| 18 | 0.02 | 0.14 | 0.12 | 0.53 | 0.48 | 0.41 | 0.54 | 0.46 |
| 19 | 0.02 | 0.19 | 0.11 | 0.51 | 0.40 | 0.22 | 0.43 | 0.26 |
| 20 | 0.02 | 0.13 | 0.12 | 0.53 | 0.49 | 0.44 | 0.61 | 0.51 |
| 21 | 0.06 | 0.32 | 0.31 | 1.64 | 0.50 | 0.44 | 0.65 | 0.51 |
| 22 | 0.06 | 0.49 | 0.38 | 1.68 | 1.50 | 1.19 | 1.75 | 1.35 |
| 23 | 0.06 | 0.50 | 0.42 | 1.63 | 0.74 | 0.65 | 0.86 | 0.72 |
| 24 | 0.06 | 0.48 | 0.45 | 1.67 | 1.48 | 1.26 | 1.64 | 1.42 |
| 25 | 0.02 | 0.17 | 0.17 | 0.68 | 0.25 | 0.24 | 0.31 | 0.28 |
| 26 | 0.02 | 0.20 | 0.19 | 0.67 | 0.63 | 0.61 | 0.75 | 0.70 |
| 27 | 0.03 | 0.25 | 0.16 | 0.68 | 0.46 | 0.27 | 0.49 | 0.31 |
| 28 | 0.02 | 0.19 | 0.20 | 0.67 | 0.63 | 0.62 | 0.74 | 0.71 |
| 29 | 0.07 | 0.45 | 0.46 | 1.98 | 0.67 | 0.64 | 0.77 | 0.74 |
| 30 | 0.07 | 0.56 | 0.52 | 1.94 | 1.80 | 1.64 | 2.27 | 1.88 |
| 31 | 0.07 | 0.61 | 0.54 | 1.98 | 0.86 | 0.79 | 0.99 | 0.92 |
| 32 | 0.07 | 0.55 | 0.57 | 1.96 | 1.82 | 1.72 | 2.12 | 1.95 |
| 33 | 0.02 | 0.16 | 0.14 | 0.59 | 0.23 | 0.19 | 0.32 | 0.21 |
| 34 | 0.02 | 0.16 | 0.12 | 0.63 | 0.59 | 0.47 | 0.69 | 0.56 |
| 35 | 0.02 | 0.17 | 0.14 | 0.59 | 0.24 | 0.20 | 0.29 | 0.22 |
| 36 | 0.02 | 0.16 | 0.13 | 0.62 | 0.58 | 0.47 | 0.70 | 0.56 |
| 37 | 0.05 | 0.36 | 0.32 | 1.43 | 0.52 | 0.44 | 0.66 | 0.52 |
| 38 | 0.05 | 0.38 | 0.31 | 1.54 | 1.44 | 1.17 | 1.77 | 1.43 |
| 39 | 0.06 | 0.40 | 0.35 | 1.45 | 0.57 | 0.49 | 0.70 | 0.54 |
| 40 | 0.05 | 0.38 | 0.30 | 1.53 | 1.43 | 1.17 | 1.78 | 1.40 |
| 41 | 0.02 | 0.13 | 0.13 | 0.55 | 0.20 | 0.18 | 0.26 | 0.22 |
| 42 | 0.02 | 0.14 | 0.13 | 0.56 | 0.54 | 0.51 | 0.74 | 0.62 |
| 43 | 0.02 | 0.14 | 0.14 | 0.56 | 0.21 | 0.19 | 0.27 | 0.22 |
| 44 | 0.02 | 0.15 | 0.13 | 0.59 | 0.56 | 0.53 | 0.68 | 0.60 |
| 45 | 0.05 | 0.35 | 0.34 | 1.41 | 0.52 | 0.49 | 0.59 | 0.56 |

*Table 2. Continued.*

| | Domains | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| No. | bias 6–50 SR | bias 6–50 EBLUP | bias 6–50 ZERO-F | rmse 6–50 SR | rmse 6–50 EBLUP | rmse 6–50 ZERO-F | rmse 1–50 EBLUP | rmse 1–50 ZERO-F |
| 46 | 0.06 | 0.43 | 0.38 | 1.52 | 1.43 | 1.36 | 1.77 | 1.63 |
| 47 | 0.05 | 0.35 | 0.34 | 1.43 | 0.52 | 0.49 | 0.63 | 0.60 |
| 48 | 0.05 | 0.36 | 0.37 | 1.48 | 1.42 | 1.37 | 1.69 | 1.62 |

Another way to summarize the results about the rmse is to compute the ratios $rmse_{\mathrm{EBLUP}}/rmse_{\mathrm{ZERO-F}}$ for all domains and the ten populations and compute quantiles of these ratios. The results are shown in Table 3. Since the focus of this article is the comparison of the EBLUP and ZERO-F, such a comparison is not carried out between SR and SAE methods. We see the following results:

- In all cases there are at least some domains where the EBLUP is more accurate than ZERO-F.
- In almost all cases, the 35% quantile is larger than 1, so ZERO-F is more accurate than the EBLUP in at least 65% of the domains.
- In the cases with large $\sigma_{r,\mathrm{z}}$ (even numbers) and a nonzero fraction of around 0.5, the differences between the domains are relatively small with a ten percent quantile of between 0.96 and 1.03 and a 90% quantile between 1.05 and 1.2.
- In the cases of large residual variances $\sigma^2_{e,\mathrm{z}}$ and $\sigma^2_{e,\mathrm{nz}}$ and a nonzero fraction of around 0.5 (number 13 and 15), the differences between the domains are also relatively small.
- For a nonzero fraction around 0.85 or 0.1, the differences between the domains are generally larger, with two exceptions (small random effects $\vartheta_{\mathrm{z},j}$ and $\vartheta_{\mathrm{nz},j}$ and large residual variance $\sigma^2_{e,\mathrm{z}}$, nonzero fraction of around 0.85 (number 25 and 29).
- In many cases with small $\sigma_{r,\mathrm{z}}$ (odd numbers), the EBLUP is substantially more accurate than ZERO-F for quite a large fraction of the domains (10% quantile smaller than 0.9). These are often the cases where the mean gain of ZERO-F with respect to the EBLUP over all domains is relatively large. This means that the gain in accuracy in many domains has to be paid for with some substantial loss in accuracy in some other domains.

### *3.4. Results for Domains*

Table 3 shows that the gain in accuracy of ZERO-F with respect to the EBLUP sometimes differs strongly between the domains. An analysis of the results for the domains shows that in the situations with large $\sigma_{r,\mathrm{z}}$ (even numbered rows), the gain in accuracy of ZERO-F generally depends strongly on the size of the random effects $\vartheta_{\mathrm{z},j}$. The gain is larger in the domains with the smallest (most negative) and/or the largest random effects. This gain is

Table 3. Quantiles, minimum and maximum of ratios rmse EBLUP and ZERO-F.

| No. | Min | 10% | 25% | 35% | 50% | 65% | 75% | 90% | Max |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.34 | 0.85 | 0.97 | 1.01 | 1.06 | 1.11 | 1.15 | 1.34 | 27.79 |
| 2 | 0.78 | 1.02 | 1.04 | 1.05 | 1.06 | 1.08 | 1.10 | 1.16 | 4.36 |
| 3 | 0.34 | 0.78 | 1.07 | 1.19 | 1.33 | 1.55 | 1.76 | 2.25 | 21.54 |
| 4 | 0.76 | 1.01 | 1.03 | 1.04 | 1.06 | 1.07 | 1.09 | 1.15 | 2.88 |
| 5 | 0.41 | 0.87 | 0.97 | 1.01 | 1.06 | 1.11 | 1.15 | 1.33 | 34.43 |
| 6 | 0.77 | 1.03 | 1.06 | 1.07 | 1.09 | 1.11 | 1.13 | 1.20 | 5.18 |
| 7 | 0.26 | 0.86 | 0.97 | 1.00 | 1.05 | 1.09 | 1.14 | 1.29 | 45.14 |
| 8 | 0.72 | 0.99 | 1.03 | 1.05 | 1.07 | 1.09 | 1.12 | 1.18 | 56.46 |
| 9 | 0.75 | 0.98 | 1.00 | 1.01 | 1.02 | 1.03 | 1.04 | 1.08 | 6.32 |
| 10 | 0.85 | 0.98 | 1.00 | 1.00 | 1.01 | 1.02 | 1.03 | 1.06 | 11.52 |
| 11 | 0.42 | 0.81 | 0.98 | 1.06 | 1.20 | 1.39 | 1.57 | 1.97 | 8.67 |
| 12 | 0.55 | 0.98 | 0.99 | 1.00 | 1.01 | 1.01 | 1.02 | 1.05 | 4.93 |
| 13 | 0.85 | 0.99 | 1.01 | 1.01 | 1.02 | 1.03 | 1.04 | 1.06 | 3.33 |
| 14 | 0.83 | 0.99 | 1.01 | 1.01 | 1.02 | 1.04 | 1.05 | 1.10 | 1.31 |
| 15 | 0.77 | 0.94 | 0.98 | 1.00 | 1.02 | 1.05 | 1.07 | 1.12 | 4.68 |
| 16 | 0.81 | 0.96 | 1.00 | 1.00 | 1.01 | 1.03 | 1.04 | 1.08 | 1.47 |
| 17 | 0.28 | 0.84 | 0.96 | 1.01 | 1.08 | 1.14 | 1.23 | 1.44 | 43.97 |
| 18 | 0.53 | 1.03 | 1.08 | 1.11 | 1.15 | 1.22 | 1.31 | 1.56 | 6.15 |
| 19 | 0.36 | 1.21 | 1.50 | 1.65 | 1.84 | 1.98 | 2.11 | 2.49 | 5.42 |
| 20 | 0.62 | 1.03 | 1.08 | 1.09 | 1.12 | 1.16 | 1.19 | 1.28 | 19.79 |
| 21 | 0.18 | 0.85 | 0.97 | 1.04 | 1.12 | 1.19 | 1.25 | 1.56 | 40.88 |
| 22 | 0.50 | 1.06 | 1.14 | 1.19 | 1.25 | 1.33 | 1.44 | 1.69 | 62.60 |
| 23 | 0.12 | 0.82 | 0.96 | 1.04 | 1.13 | 1.21 | 1.28 | 1.49 | 31.22 |
| 24 | 0.50 | 0.98 | 1.07 | 1.11 | 1.15 | 1.21 | 1.29 | 1.51 | 5.04 |
| 25 | 0.91 | 0.97 | 0.99 | 1.01 | 1.02 | 1.05 | 1.06 | 1.10 | 17.24 |
| 26 | 0.64 | 0.95 | 0.99 | 1.00 | 1.04 | 1.08 | 1.13 | 1.28 | 11.82 |
| 27 | 0.40 | 0.94 | 1.27 | 1.42 | 1.68 | 1.93 | 2.10 | 2.56 | 5.27 |
| 28 | 0.65 | 0.94 | 0.98 | 1.00 | 1.02 | 1.05 | 1.07 | 1.14 | 12.87 |
| 29 | 0.71 | 0.97 | 0.99 | 1.01 | 1.03 | 1.05 | 1.06 | 1.09 | 6.46 |
| 30 | 0.61 | 0.98 | 1.02 | 1.04 | 1.10 | 1.19 | 1.25 | 1.37 | 16.10 |
| 31 | 0.64 | 0.88 | 0.97 | 1.00 | 1.06 | 1.11 | 1.16 | 1.30 | 3.43 |
| 32 | 0.62 | 0.94 | 1.00 | 1.02 | 1.05 | 1.10 | 1.16 | 1.33 | 12.37 |
| 33 | 0.39 | 0.79 | 0.99 | 1.10 | 1.21 | 1.35 | 1.49 | 1.99 | 18.16 |
| 34 | 0.48 | 1.09 | 1.16 | 1.18 | 1.24 | 1.33 | 1.46 | 1.80 | 8.12 |
| 35 | 0.16 | 0.80 | 0.96 | 1.08 | 1.20 | 1.34 | 1.48 | 1.97 | 18.84 |
| 36 | 0.52 | 1.07 | 1.14 | 1.17 | 1.23 | 1.33 | 1.42 | 1.75 | 4.21 |
| 37 | 0.13 | 0.76 | 0.95 | 1.03 | 1.14 | 1.25 | 1.37 | 1.90 | 11.24 |
| 38 | 0.51 | 1.07 | 1.13 | 1.16 | 1.23 | 1.34 | 1.51 | 1.85 | 5.12 |
| 39 | 0.11 | 0.76 | 0.94 | 1.02 | 1.13 | 1.27 | 1.38 | 1.85 | 9.72 |
| 40 | 0.43 | 1.06 | 1.13 | 1.15 | 1.21 | 1.34 | 1.52 | 1.76 | 5.97 |
| 41 | 0.45 | 0.88 | 0.94 | 0.98 | 1.01 | 1.06 | 1.09 | 1.22 | 6.51 |
| 42 | 0.60 | 0.95 | 1.00 | 1.02 | 1.08 | 1.17 | 1.22 | 1.35 | 6.02 |
| 43 | 0.59 | 0.85 | 0.94 | 1.00 | 1.06 | 1.11 | 1.18 | 1.37 | 8.00 |
| 44 | 0.61 | 0.96 | 1.00 | 1.02 | 1.08 | 1.15 | 1.21 | 1.34 | 4.13 |
| 45 | 0.64 | 0.90 | 0.96 | 0.98 | 1.01 | 1.05 | 1.08 | 1.17 | 3.00 |
| 46 | 0.55 | 0.95 | 0.99 | 1.01 | 1.06 | 1.12 | 1.18 | 1.31 | 4.81 |
| 47 | 0.61 | 0.85 | 0.92 | 0.96 | 1.01 | 1.07 | 1.12 | 1.27 | 2.26 |
| 48 | 0.54 | 0.95 | 0.99 | 1.01 | 1.05 | 1.11 | 1.16 | 1.27 | 3.16 |

*Table 4. Ratio mean rmse EBLUP and ZERO-F over the ten created populations and over groups of domains, ordered by size of random effects $\vartheta_{z,j}$.*

| No. | Fraction nonzeros | Ratio domains 1–10 | Ratio domains 11–40 | Ratio domains 41–50 |
|---|---|---|---|---|
| 2 | 0.50 | 1.10 | 1.05 | 1.08 |
| 18 | 0.83 | 1.09 | 1.15 | 1.43 |
| 34 | 0.16 | 1.60 | 1.23 | 1.12 |

caused by both a smaller bias and a smaller standard deviation of ZERO-F in these domains. In situations with around 50% nonzero target variables, the gain in accuracy is similar in the domains with the smallest and the largest random effects. In situations with around 85% nonzero target variables, this gain is larger in the domains with the largest random effects. In situations with around ten percent nonzero target variables, it is the opposite. This is demonstrated for three situations in Table 4. There, for three groups of domains, the mean rmse is computed over the selected domains and over the ten created populations for each situation. This is done for both the EBLUP and the ZERO-F. The column 'Ratio domains 1–10' shows the ratio of both values for the ten domains with the smallest (most negative) random effects $\vartheta_{z,j}$. The same ratio for the ten domains with the largest random effects is given in the column 'Ratio domains 41–50' and the ratio for the other 30 domains is computed in column 'Ratio domains 11–40'. For the other situations with large $\sigma_{r,z}$, similar results are found. However, sometimes the pattern is disturbed due to coincidences in the domains.

In the situations with small $\sigma_{r,z}$ (odd numbers), there is no visible influence of the size of the random effects $\vartheta_{z,j}$ on the gain in accuracy in the domains of ZERO-F with respect to the EBLUP. In a few cases, a similar dependency on the size of the random effects $\vartheta_{nz,j}$ is visible. The gain in accuracy of ZERO-F with respect to the EBLUP does not depend strongly on the domain size. The gain in accuracy of both SAE methods with respect to the design-based SR decreases with increasing sample size, a rather general phenomenon in small-area estimation.

In many situations with small $\sigma_{r,z}$, the differences between the domains cannot be explained by domain size or the size of the random effects.

The results for the domain where the 0.9 quantile of the vector $\mathbf{x}$ is added are special in many cases. There, the rmse of the EBLUP and the SR are similar, and the rmse of the ZERO-F is smaller, whereas in most of the other domains, the rmse of the EBLUP is smaller than the one of the SR.

### 3.5. Results for Larger Populations and for Correlated Random Effects

The simulations for the first four situations are repeated for larger populations ($N = 180{,}000$) and larger samples ($n = 6{,}000$). The results for these simulations are similar to those for the smaller populations and samples discussed in the previous subsection and are therefore not included in detail. As expected, the possible gain in accuracy by using SAE methods instead of the SR is smaller when the sample size increases. The gain in accuracy of ZERO-F with respect to the EBLUP is more or less equal to that with smaller sample sizes.

*Table 5. Mean absolute bias and mean rmse, correlation 0.5.*

| | Domains | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| No. | bias 6–50 SR | bias 6–50 EBLUP | bias 6–50 ZERO-F | rmse 6–50 SR | rmse 6–50 EBLUP | rmse 6–50 ZERO-F | rmse 1–50 EBLUP | rmse 1–50 ZERO-F |
| 1 | 0.03 | 0.21 | 0.21 | 0.78 | 0.30 | 0.28 | 0.35 | 0.32 |
| 2 | 0.03 | 0.19 | 0.18 | 0.74 | 0.70 | 0.65 | 0.80 | 0.76 |
| 3 | 0.03 | 0.29 | 0.22 | 0.78 | 0.47 | 0.31 | 0.53 | 0.36 |
| 4 | 0.03 | 0.18 | 0.19 | 0.75 | 0.71 | 0.67 | 0.89 | 0.81 |

Furthermore, the simulations for the first four situations are repeated with correlated random effects $\vartheta_{z,j}$ and $\vartheta_{nz,j}$. The results are shown in Table 5 (for correlation 0.5) and Table 6 (for correlation 0.9). The accuracy of the SR is, as expected, not affected by this correlation. The effect on the accuracy of the EBLUP and ZERO-F is also small. Only for the EBLUP in Situation 3 is there some loss in accuracy, compared with the situation with uncorrelated random effects (Table 2). Despite the model misspecification of ZERO-F (by ignoring the correlation), the improvement of the accuracy by ZERO-F instead of the SR is of the same order as in the situation where the correlation is zero. Nevertheless, it can be useful to investigate ZERO with modelling the correlation in order to achieve an additional gain in accuracy. However, in the example of Pfeffermann et al. (2008) the improvement in accuracy by using this more complex model is very small.

### *3.6. Results for Bayesian Approach*

Finally, the simulations for the first four situations are repeated with ZERO-B. For these simulations, a single population is created for each situation, the number of runs in the simulations being 1,000. The mean absolute bias, mean sd and mean rmse over the domains of ZERO-F and ZERO-B are shown in Table 7. The general conclusion is that the accuracy of both approaches is very similar. The bias is slightly reduced with the Bayesian approach, whereas the sd is slightly increased.

Figures 1 and 2 show boxplots of the model-based rmse based on the MCMC simulations for all 1,000 runs of the simulation for Situations 1 and 2. For Situations 3 and 4 similar results were obtained, so these results are omitted. The simulation rmses,

*Table 6. Mean absolute bias and mean rmse, correlation 0.9.*

| | Domains | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| No. | bias 6–50 SR | bias 6–50 EBLUP | bias 6–50 ZERO-F | rmse 6–50 SR | rmse 6–50 EBLUP | rmse 6–50 ZERO-F | rmse 1–50 EBLUP | rmse 1–50 ZERO-F |
| 1 | 0.03 | 0.24 | 0.24 | 0.78 | 0.33 | 0.30 | 0.40 | 0.35 |
| 2 | 0.03 | 0.18 | 0.19 | 0.74 | 0.70 | 0.66 | 0.86 | 0.77 |
| 3 | 0.03 | 0.29 | 0.22 | 0.78 | 0.48 | 0.31 | 0.52 | 0.35 |
| 4 | 0.03 | 0.16 | 0.19 | 0.75 | 0.71 | 0.67 | 0.91 | 0.84 |

Table 7. Bias, sd and rmse of ZERO-F and ZERO-B, mean over the domains, for bias mean of absolute values.

| No. | freq bias | mcmc bias | freq sd | mcmc sd | freq rmse | mcmc rmse |
|---|---|---|---|---|---|---|
| 1 | 0.239 | 0.231 | 0.162 | 0.172 | 0.307 | 0.308 |
| 2 | 0.222 | 0.217 | 0.644 | 0.648 | 0.714 | 0.715 |
| 3 | 0.255 | 0.248 | 0.208 | 0.215 | 0.350 | 0.351 |
| 4 | 0.234 | 0.229 | 0.660 | 0.664 | 0.735 | 0.735 |

computed with (12), are added to the figures. Again, there is a large difference between the situations with large and small $\sigma_{r,z}$. For large $\sigma_{r,z}$ (Situation 2, Figure 2), the model-based rmse tracks the simulation rmse very well. In those cases, the variation of the model-based rmse is quite small (except for the smallest domains) and the bulk of the distribution is positioned closely around the simulation rmse. If $\sigma_{r,z}$ is small (Situation 1, Figure 1), the bulk of the distribution of the model-based rmse often deviates from the simulation rmse. The model-based rmses do not vary much over the domains in these cases, whereas the simulation rmses do. Nevertheless, the model-based rmses are of the same order of magnitude as the simulation rmses and can therefore be useful as an indication for the accuracy of the estimates, even in a repeated sampling sense.

## 4. Simulation with Dutch HBS Data

### *4.1. Design of HBS*

The aim of the Dutch Household Budget Survey (HBS) is to measure the expenditures of households. Some of these expenses are on a regular basis, for example often the same amount of money is paid every month for rent and insurance premiums. Other expenses are quite regular, although with varying amounts of money spent. This often concerns cheaper products; for example, food is bought almost every week. Finally, there are also

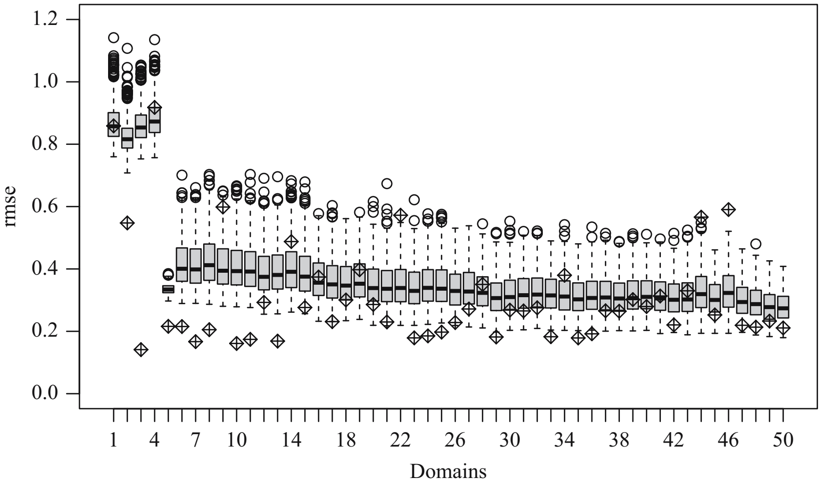


*Fig. 1. Boxplots of MCMC estimates for rmse of ZERO-B from 1,000 simulation runs and rmse based on simulation (diamonds), Situation 1.*

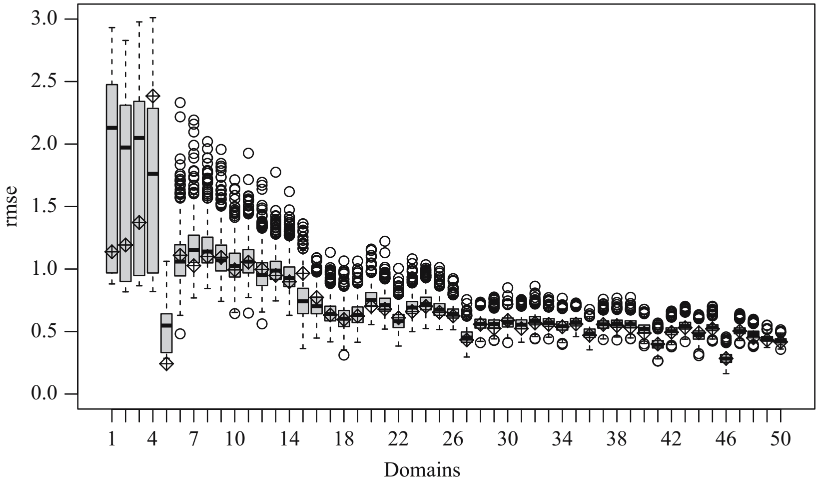


*Fig. 2. Boxplots of MCMC estimates for rmse of ZERO-B from 1,000 simulation runs and rmse based on simulation (diamonds), Situation 2.*

expenses that are more rare, for example furniture or clothes. These products are often, but not always, relatively expensive. Therefore, the HBS considers three kinds of expenditures which are measured in different parts of the survey. In this simulation we consider three kinds of expenditures, mainly measured in the part "large expenditures". In this part, the responding households keep a diary of their expenditures over €20.

The HBS has been redesigned repeatedly with the aim to increase response rates and decrease costs. Since 2012, the diary for large expenditures has been kept for a period of four weeks. For the simulation, data from the period 2005–2010 is used. In those years, the diary for large expenditures was kept for three months. To approximate the current design as far as possible, we use periods of one month in the simulation, in which each original sample household with expenditures over three months is considered as three independent sample households with expenditures over one month.

Data from 2005–2010 are combined in a single dataset of $N = 100{,}000$ households with expenditures for one month. The expenditures are corrected for inflation to have comparable prices over the years. This artificial population can be considered a representative sample from the population of Dutch households. The complete Dutch population consists of more than seven million households. The artificial population is chosen to be smaller for computational reasons.

Based on the HBS, household expenditures are published for the entire country and for different classifications in subpopulations. In this article, we consider a classification in $m = 11$ types of households. Table 8 shows these domains and their sizes in the artificial population. In the simulation, samples of size $n = 5{,}000$ are drawn by simple random sampling without replacement. Complications caused by different response probabilities which occur in practice are avoided. In the simulation 3,000 samples are drawn.

In the simulation, the expenditures for clothes, men's clothes and motor fuel are used as target variables. All three variables contain substantial amounts of zeros. This is partly because the households had no expenditures of this kind in the considered month, and

Table 8. *Population size per type of household in artificial population of 100,000 households.*

| No. | Description | Population size |
|---|---|---|
| 1 | single man, younger than 65 years | 12,976 |
| 2 | single man, 65 years or older | 2,985 |
| 3 | single woman, younger than 65 years | 11,176 |
| 4 | single woman, 65 years or older | 8,141 |
| 5 | couple, main wage earner younger than 65 years | 18,781 |
| 6 | couple, main wage earner 65 years or older | 10,514 |
| 7 | couple with child(ren), all children younger than 18 years | 19,803 |
| 8 | couple with child(ren), at least one child 18 years or older | 8,020 |
| 9 | one-parent family, all children younger than 18 years | 4,006 |
| 10 | one-parent family, at least one child 18 years or older | 2,291 |
| 11 | other households | 1,307 |

partly because they do not have expenditures of this kind at all. For example, households with only female members generally do not buy men's clothes and households without a car or a motorcycle do not buy motor fuel.

Table 9 shows the percentages of nonzero expenditures, the means of the nonzero expenditures, and the overall expenditure means for the three target variables and the eleven household types. There are substantial differences between the domains. These differences suggest that substantial random effects can be expected. However, part of the differences may be explained by other auxiliary variables used in the models.

For the considered estimators (SR, EBLUP, and ZERO), the same auxiliary information is used. This is a combination of different socio-economic variables. Income is the only continuous auxiliary variable; furthermore, categorical variables about the source of income of the main wage earner, the housing situation (owner or tenant) are used. Since the expenditures vary over the course of the year, quarter is also added.

Table 9. *Percentage nonzero expenditures, mean of nonzero expenditures and overall mean for three target variables and eleven household types.*

| | Percentage | | | Mean of nonzeros | | | Mean expenditure | | |
|---|---|---|---|---|---|---|---|---|---|
| No. | Clothes | Motor fuel | Men's clothes | Clothes | Motor fuel | Men's clothes | Clothes | Motor fuel | Men's clothes |
| 1 | 23 | 46 | 22 | 134 | 99 | 136 | 31 | 45 | 29 |
| 2 | 21 | 57 | 20 | 111 | 71 | 111 | 24 | 40 | 22 |
| 3 | 40 | 41 | 0.7 | 109 | 75 | 89 | 43 | 30 | 0.6 |
| 4 | 35 | 28 | 0.6 | 114 | 55 | 75 | 39 | 15 | 0.5 |
| 5 | 47 | 73 | 22 | 164 | 109 | 136 | 76 | 79 | 30 |
| 6 | 40 | 70 | 17 | 142 | 78 | 111 | 57 | 54 | 19 |
| 7 | 56 | 73 | 20 | 160 | 113 | 130 | 89 | 82 | 26 |
| 8 | 56 | 76 | 27 | 164 | 121 | 126 | 92 | 92 | 33 |
| 9 | 44 | 53 | 3.5 | 105 | 86 | 94 | 46 | 45 | 3.3 |
| 10 | 42 | 61 | 9.5 | 124 | 89 | 110 | 52 | 54 | 10 |
| 11 | 44 | 64 | 18 | 158 | 110 | 128 | 69 | 71 | 23 |

The accuracy of the estimates slightly depends on the auxiliary information, as was investigated in a preliminary study (details not presented). However, the main results do not change, as long as the model is not overfitted.

As in Section 3, the point estimates of ZERO-F and ZERO-B are very similar. Therefore, only the results under ZERO-B are presented. In Subsection 4.2 the point estimates are discussed, and in Subsection 4.3 results for the mse estimates as well as credible intervals are presented.

### 4.2. Point Estimates

The rmses for the different estimators for the target variable clothes are shown for the eleven domains in Figure 3. The results are mixed: for each estimator, there is at least one domain where this estimator has the largest rmse. On the other hand, the SAE methods are more accurate than the SR for a majority of the domains. This is most clear for ZERO-B which is more accurate than the SR for all but one domain. The Domains 1 and 2 (single men, younger than 65 years/65 years and older) are special for clothes, since these households do not buy many clothes (compare Table 9). ZERO-B can handle these domains better than the EBLUP. The results for men's clothes and motor fuel are similar and therefore not shown in figures. There, other domains are special due to very small expenditures in these domains. For men's clothes, these are Domains 3, 4, and 9 (single women, younger than 65 years/65 years and older, one-parent family, all children younger than 18 years), where again ZERO is more accurate than the EBLUP. For motor fuel, it is

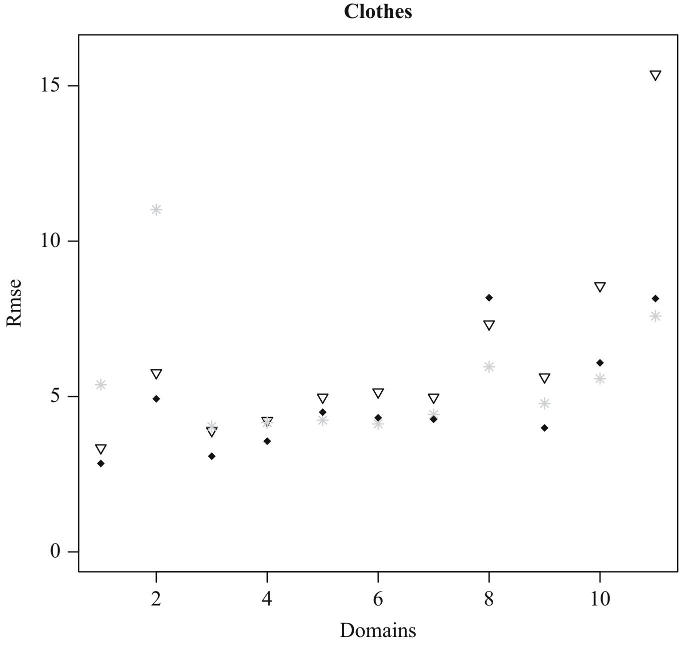


*Fig. 3. Root mse of three estimators with four different fixed effects for clothes (triangle: SR, star: EBLUP, diamond: ZERO-B).*

*Table 10. Mean rmse (first three columns), mean relative rmse with all domains included (Columns 5–7) and mean relative rmse with Domains 3 and 4 excluded (Columns 5–7, between brackets) for three variables.*

| | rmse | | | rel. rmse | | |
|---|---|---|---|---|---|---|
| Fixed | SR | EBLUP | ZERO-B | SR | EBLUP | ZERO-B |
| Clothes | 6.294 | 5.570 | 4.900 | 0.125 (−) | 0.130 (−) | 0.097 (−) |
| Men's clothes | 3.530 | 3.295 | 2.945 | 0.372 (0.239) | 0.675 (0.238) | 0.369 (0.183) |
| Motor fuel | 3.834 | 3.432 | 3.284 | 0.074 (−) | 0.072 (−) | 0.068 (−) |

Domain 4 (single women, 65 years and older). There, however, the EBLUP and ZERO-B have a similar rmse, which is much larger than the rmse of the SR.

To summarize the results, the mean of the rmse and the mean of the relative rmse over the domains are computed. Since the relative rmse for Domain 3 and 4 is extremely large for men's clothes, for this variable the mean relative rmse is also computed with these domains excluded. The results are shown in Table 10. Based on this table, it can be concluded that ZERO-B is the most accurate estimator. The fact that the mean relative rmse of ZERO-B is almost equal to the one of the SR for men's clothes (Column 5 and 7) is caused by the extremely large values for Domain 3 and 4 (compare the numbers between brackets). The EBLUP is also more accurate than the SR, but the possible gain is smaller than the one achieved by using ZERO-B. The possible gain in accuracy also varies between the target variables.

### 4.3. mse Estimates and Coverage

Figure 4 compares the model-based rmse obtained from the MCMC simulation computed using (11) with the design-based rmse based on the simulation for the variable clothes. Even though the two rmse concepts are quite different, it can still be useful to have good frequentist properties of the model-based rmse and of model-based intervals. In most domains the model-based rmse is on average somewhat larger than the simulation rmse.

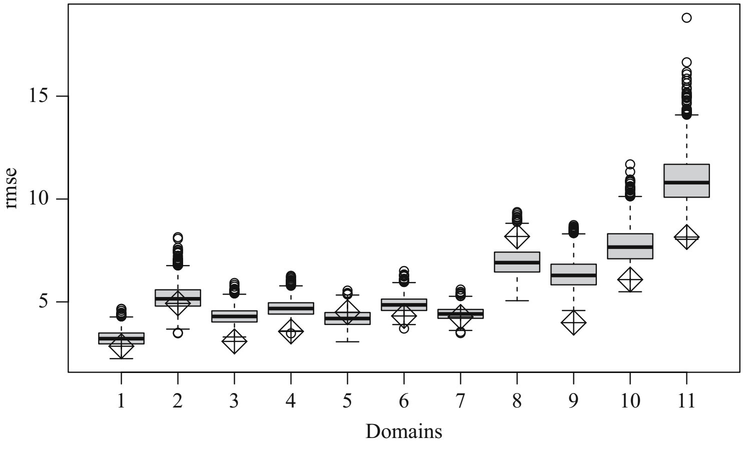


*Fig. 4. Boxplot of MCMC estimates for root mse and rmse based on simulation (diamonds), clothes.*

Table 11. Coverage for 95% highest posterior density intervals.

| | Domain | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Clothes | 96.4 | 97.7 | 99.3 | 98.4 | 91.6 | 96.1 | 95.3 | 89.6 | 99.5 | 99.3 | 99.2 |
| Men's clothes | 90.0 | 94.4 | 97.4 | 96.9 | 92.7 | 98.5 | 93.4 | 96.6 | 98.2 | 98.6 | 96.3 |
| Motor fuel | 95.4 | 99.5 | 98.6 | 97.0 | 95.4 | 98.8 | 94.9 | 93.5 | 96.3 | 98.5 | 97.0 |

For Domain 8 it is smaller on average. Nevertheless the model-based rmse seems a useful measure of accuracy for the domain estimates, even in a repeated sampling sense. For the other two variables, no figures are shown because the results are similar.

The coverages of 95% highest posterior density intervals are shown in Table 11 for all three variables. Most coverages are not very far from 95%, although some undercoverage and overcoverage occurs depending on the domain and the variable of interest. Intervals based on the normal approximation using the model-based mse have also been computed. They are quite similar to the highest posterior density intervals, although slightly wider, and in almost all cases their coverages are close to those for the highest posterior density intervals.

## 5. Conclusion

Model-based small-area estimation (SAE) can be considered as an alternative to approximately design-unbiased estimation if the sample size is too small for producing reliable design-based estimates. Zero-inflated target variables occur in many surveys by national statistical institutets. Therefore, in this article three SAE methods are compared with each other and with a design-based estimator in a simulation study using zero-inflated variables. The first SAE method is the EBLUP (Rao 2003), which is the most common SAE method but ignores zero inflation. The second and third SAE method, developed by Pfeffermann et al. (2008) and Chandra and Sud (2012), take the zero inflation explicitly into account. They are based on the same models but use the Bayesian and the frequentist approach respectively. They result in similar point estimates and are referred to in abbreviation as ZERO. The general conclusion is that in the case of zero-inflated variables, an improvement of accuracy can be achieved with all SAE estimators compared with design-based methods. So the performance of the EBLUP is often satisfactory even though the model of the EBLUP is misspecified since it ignores the zero inflation. Generally ZERO is more accurate than the EBLUP. In a simulation with artificial populations, the properties of the populations can be controlled. There, ZERO is less model misspecified. The amount of improvement in accuracy of ZERO compared with the EBLUP depends on the properties of the entire population and the domains. In some populations, the improvement is negligible; in others, it is substantial. In all considered simulations, there are also some domains where the EBLUP is more accurate than ZERO.

The accuracy of the point estimates of ZERO under the frequentist approach or under the Bayesian approach is almost equal, which means that the statistician's taste can be the deciding factor. A disadvantage of the Bayesian approach is that the computation time is higher, while an advantage is that information about the accuracy of the estimates follows directly. For the frequentist approach, no formula for the mean squared error has been developed so far. Parametric bootstrapping can be applied as proposed by Chandra and Sud (2012), which is also computationally intensive. The mean squared error estimates under the Bayesian approach do not always track the simulation error accurately. However, the mean squared error estimates seem to be useful as an indication of accuracy.

In a second simulation, real data of the Household Budget Survey (HBS) of Statistics Netherlands are used. The considered target variables, expenditures for three products, are zero inflated. In this simulation, the properties of the population cannot be controlled. Model misspecification is now more pronounced for ZERO since this estimator takes only one particular deviation from normality (zero inflation) into account, but no other possible deviations. Nevertheless, ZERO is the most accurate estimator for the majority of the domains. Contrary to the first simulation, in the simulation with HBS data there are some domains where the design-based estimator is substantially more accurate than ZERO. Such domains are very rare in the simulation with artificial data. Further results of both simulations are similar.

ZERO as used in this paper assumes a normal distribution of the nonzero part of the population. This assumption is not quite met in our application to the HBS. A suitable transformation applied to the target variable, as described in Dreassi et al. (2012) and Chandra and Chambers (2011a), could improve the model and the estimates. Furthermore, a model that replaces the normal distribution for random effects by one with wider tails might be able to better accommodate outlying random effects to prevent overshrinkage. In the continuation of this research, these potential improvements can be implemented and the results can be compared with those in this paper. Other research questions are how the estimators work if a complex design of the survey and different response probabilities must be taken into account.

## 6. References

Bafumi, J. and A. Gelman. 2006. "Fitting Multilevel Models When Predictors and Group Effects Correlate." Manuscript prepared for the 2006 Annual Meeting of the Midwest Political Science Association, Chicago. DOI: http://dx.doi.org/10.2139/ssrn.1010095.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67(1). Doi: http://dx.doi.org/10.18637/jss.v067.i01.

Battese, G., R. Harter, and W. Fuller. 1988. "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data." *Journal of the American Statistical Association* 83: 28–36. Doi: http://dx.doi.org/10.2307/2288915.

Boonstra, H., J. van den Brakel, B. Buelens, S. Krieg, and M. Smeets. 2008. "Towards Small Area Estimation at Statistics Netherlands." *Metron* LXVI (1): 21–49. Available at: https://www.researchgate.net/profile/Jan_Brakel/publication/227458249_Towards_small_area_estimation_at_Statistics_Netherlands/links/0c96052f8fdda8aedd000000.pdf.

Chambers, R. and N. Tzavidis. 2006. "M-Quantile Models for Small Area Estimation." *Biometrika* 93: 255–268. Doi: http://dx.doi.org/10.1093/biomet/93.2.255.

Chandra, H. and R. Chambers. 2011a. "Small Area Estimation for Skewed Data in Presence of Zeros." *The Bulletin of Calcutta Statistical Association* 63: 249–252.

Chandra, H. and R. Chambers. 2011b. "Small Area Estimation under Transformation to Linearity." *Survey Methodology* 37: 39–51.

Chandra, H. and U. Sud. 2012. "Small Area Estimation for Zero-Inflated Data." *Communications in Statistics - Simulation and Computation* 41: 632–642. Doi: http://dx.doi.org/10.1080/03610918.2011.598991.

Dreassi, E., A. Petrucci, and E. Rocco. 2012. "Small Area Estimation for Semicontinuous Skewed Georeferenced Data." Technical Report Working paper 2012/05, Dipartimento di Statistica "G. Parenti", Florence. Available at: http://local.disia.unifi.it/ricerca/pubblicazioni/working_papers/2012/wp2012_05.pdf.

Eurarea. 2004. *Project Reference Volume, deliverable d7.1.4*. Technical report, EURAREA consortium.

Gelfand, A. and A. Smith. 1990. "Sampling Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85: 398–409. Doi: http://dx.doi.org/10.1080/01621459.1990.10476213.

Gelman, A. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1: 515–534. Doi: http://dx.doi.org/10.1214/06-BA117A.

Gelman, A., D.V. Dyk, Z. Huang, and W. Boscardin. 2008. "Using Redundant Parameterizations to Fit Hierarchical Models." *Journal of Computational and Graphical Statistics* 17: 95–122. Doi: http://dx.doi.org/10.1198/106186008X287337.

Geman, S. and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741. Doi: http://dx.doi.org/10.1109/TPAMI.1984.4767596.

Hadfield, J.D. 2010. "Mcmc Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package." *Journal of Statistical Software* 33: 1–22. Doi: http://dx.doi.org/10.18637/jss.v033.i02.

Neuhaus, J. and C. McCulloch. 2006. "Separating Between- and Within-Cluster Covariate Effects by Using Conditional and Partitioning Methods." *Journal of the Royal Statistical Society B* 68: 859–872. Doi: http://dx.doi.org/10.1111/j.1467-9868.2006.00570.x.

Pfeffermann, D., B. Terryn, and F. Moura. 2008. "Small Area Estimation under a Two Part Random Effects Model with Application to Estimation of Literacy in Developing Countries." *Survey Methodology* 34: 67–72.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. "Coda: Convergence Diagnosis and Output Analysis for mcmc." *R News* 6: 7–11. Available at: http://oro.open.ac.uk/id/eprint/22547 (accessed 13 October, 2016).

R Development Core Team. 2009. *R: A Language and Environment for Statistical Computing*. Technical Report, R Foundation for Statistical Computing, Vienna. Available at: http://www.R-project.org. (accessed 13 October, 2016).

Rao, J.N.K. 2003. *Small Area Estimation*. New York: John Wiley.

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

Sinha, S.K. and J.N.K. Rao. 2009. "Robust Methods for Small Area Estimation." *Canadian Journal of Statistics* 37: 381–399. Doi: http://dx.doi.org/10.1002/cjs.10029.

Woodruff, R. 1966. "Use of a Regression Technique to Produce Area Breakdowns of the Monthly National Estimates of Retail Trade." *Journal of the American Statistical Association* 61: 496–504. Doi: http://dx.doi.org/10.2307/2282839.

Received March 2015
Revised April 2016
Accepted August 2016

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 987–1010, http://dx.doi.org/10.1515/JOS-2016-0052

# Dead or Alive? Dealing with Unknown Eligibility in Longitudinal Surveys

*Nicole Watson*[1]

Longitudinal surveys follow people over time and some of these people will die during the life of the panel. Through fieldwork effort, some deaths will be reported or known, but others will be unobserved due to sample members no longer being issued to field or having inconclusive fieldwork outcomes (such as a noncontact that is not followed by a contact at a later wave). The coverage of deaths identified among sample members has flow-on implications to nonresponse correction. Using the Household, Income and Labour Dynamics in Australia (HILDA) Survey, four methods are used to examine the extent of missing death reports. The first method matches the sample to the national death register. The second method uses life-expectancy tables to extrapolate the expected number of deaths among the sample with unknown eligibility. The third method is similar but models deaths from data internal to the survey. The fourth method models deaths as part of the attrition process of a longitudinal survey. The last three methods are compared to the first method and the implications for the construction of balanced panel weights and subsequent population inference are explored.

*Key words:* HILDA Survey; death register; life-expectancy tables; survival model; weighting methods.

## 1. Introduction

Both cross-sectional and longitudinal surveys have some sample units with unknown eligibility that need to be addressed in some way. In the population frame used to select the sample in a cross-sectional survey, or the first wave of a longitudinal survey, some units will be out of scope. For example, in an area-based frame some dwellings may be empty or may contain people that are not in scope of the survey. Sometimes it may not be possible to make contact with the occupants of a dwelling to ascertain whether they are in scope or not, or indeed if there is anyone resident at that address at all. Survey practitioners tend to focus on high contact rates to ensure the group of units with unknown eligibility is reasonably small.

In a longitudinal survey, the situation is more complex. A sample is selected at one point in time and then is periodically interviewed over a number of years or even decades. While

[1] Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Level 5, 111 Barry St, University of Melbourne Vic 3010, Australia. Email: n.watson@unimelb.edu.au
**Acknowledgments:** This research makes use of unit record data from the HILDA Survey (Release 13). The HILDA Project was initiated and is funded by the Australian Government Department of Social Services (DSS) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this article, however, are those of the author and should not be attributed to either DSS or the Melbourne Institute. The author thanks John Henstridge, Robert Breunig, an Associate Editor of the Journal and three anonymous referees for their helpful comments on an earlier version of this article.

eligible to be interviewed at the start of the panel, some sample members will die or move abroad during the life of the panel and thus become ineligible to be interviewed at a later wave. Through fieldwork effort, the eligibility of some sample members will be reported or known while others will be unobserved due to sample members no longer being issued to field or having inconclusive fieldwork outcomes (such as a noncontact not followed by a contact in a later wave). The total number of noncontacts grows each wave, mainly due to cases being lost or not being issued to field, and some of these sample members will be out of scope. To illustrate, the Household, Income and Labour Dynamics in Australia (HILDA) Survey is a household-based longitudinal survey that began in 2001. By Wave 13 (conducted in 2013), 25% of the 13,969 Wave 1 respondents were not contacted: four percent were lost following a move (the majority of whom were lost by Wave 3), 20% were not issued to field as a result of a prior refusal or permanent health condition that makes them unable to participate, and 0.4% were issued to field but the interviewer could not make contact. Figure 1 shows the development of these different noncontact categories over time.

The coverage of deaths and moves abroad identified among sample members has important implications for nonresponse correction in constructing sample weights. Many longitudinal surveys provide data users with a range of weights, both cross-sectional and longitudinal, reflecting the different populations to be represented for different analytic objectives and the sample cases that can contribute to these estimates (Lynn 2009). For example, many longitudinal analyses for the period from Wave 1 to Wave $t$ are concerned with individuals living in the population throughout that time, so would exclude people who have died or moved abroad since Wave 1. Some analyses may be concerned with individuals living in the population in Wave 1 who subsequently died by Wave $t$. Other analyses may focus on individuals that move abroad after Wave 1. Without appropriately accounting for whether sample members with unknown eligibility have died or moved abroad after Wave 1, longitudinal estimates (and cross-sectional estimates post Wave 1) may be biased.

This article examines what impact unknown deaths have on estimates relating to the population in Wave 1 who die by Wave $t$ as well as estimates for individuals living in the

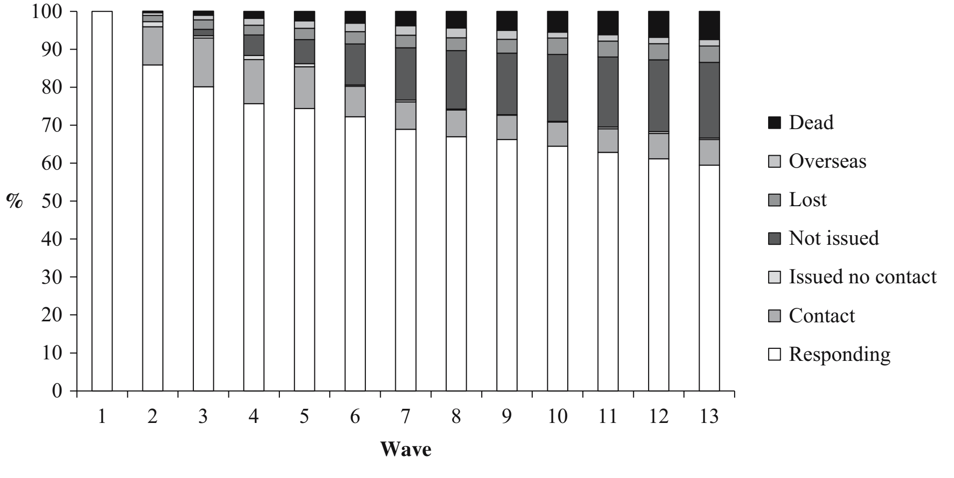


*Fig. 1. Outcomes of Wave 1 respondents (n = 13,969).*

population in Wave 1 through to Wave *t*. The focus is on deaths, rather than moves abroad, for several reasons:

i) it is the larger of the two groups (there were 243 Wave 1 respondents reported to have moved overseas by Wave 13 compared to 1,038 Wave 1 respondents who had died),
ii) it is an absorbing state so can only grow over time, and
iii) linked survey and administrative death data can provide a benchmark to which alternative weighting methods can be compared.

This analysis of unknown deaths will help inform how other types of unknown eligibility can best be handled in the weighting process.

In the construction of weights for responding sample members in cross-sectional surveys, cases with unknown eligibility may be addressed by a separate (usually simple) step in the weighting process to adjust for different eligibility rates among certain classes of the sample (Valliant et al. 2013, 315). Alternatively they may be handled in the response-propensity adjustments, or implicitly by benchmarking the responding sample to known population totals relating to eligible members in the population. In a similar way, multiple methods exist for calculating response rates in the presence of unknown eligibility. The proportion of eligible cases among those with unknown eligibility is assumed to be 0, 1 or some fraction based on known cases or external data sources. As yet there are no recommendations as to the preferred approach for dealing with unknown eligibility (Smith 2009).

A variety of practices have also been adopted by panel studies to adjust for unknown deaths in the construction of weights for the balanced panel from Wave 1 to Wave *t*. Typically general population studies make no adjustment for unknown deaths, as appears to be the case for the German Socio-Economic Panel (Kroh 2013), the British Household Panel Survey (BHPS) (Taylor 2010), and the Swiss Household Panel (Weaver 2011). There are three notable exceptions, though each have taken a different approach. First, the US Panel Study of Income Dynamics (up to 1993) used life tables from the US Census Bureau to predict the deaths of noncontacted sample members (Gouskova et al. 2008). From 1993, they assumed that all new nonrespondents were alive, presumably due to very high contact rates. Second, the Canadian Survey of Labour and Income Dynamics applies a nonresponse adjustment to known deaths, although they are not included in the nonresponse modelling to calculate the size of the adjustment (LaRoche 2007). Third, up until 2013 the HILDA Survey counted known deaths as a response in the nonresponse modelling process in constructing the weights (Watson 2012), an approach which is suggested by Kalton and Brick (2000). In 2014, the HILDA Survey sample was matched to the national death register (Watson and Summerfield 2014). Longitudinal studies of older individuals have more to gain from matching their samples to national death registers, as this provides the opportunity to gather additional information about the cause of death along with the exact timing. This is done for the US Health and Retirement Study (Hayward undated; HRS undated). Some other ageing studies take a different approach. The Survey of Health, Ageing and Retirement in Europe benchmarks the longitudinal sample to mortality-adjusted benchmarks (SHARE 2013). The English Longitudinal Study of Ageing does not appear to make any correction for unknown deaths (Scholes et al.

2008). Longitudinal studies of children are likely to ignore the issue of unknown deaths as the number of missing deaths is likely to be very small (see, for example, Hawkes and Plewis 2006, for an analysis of the UK National Child Development Study).

There appears to only be one study that has looked explicitly at the impact of adjusting for unknown eligibility on weighted estimates in a longitudinal survey. Sadig (2014) estimated and applied an adjustment factor to BHPS response-propensity-corrected weights. This adjustment factor was calculated as the ratio of the survival rate in the population (obtained from official statistics) to the survival rate in the (unweighted) sample for each sex and age group (with nine age categories). He found that the adjusted weights reduced both the mean and standard deviation of the weights (as older respondents were receiving higher weights than they ought prior to the adjustment). The adjustment also resulted in sizeable differences to population estimates of subjective health status and some differences to regression coefficients of a model of self-reported health. This approach makes two assumptions that may be violated to some degree. The first is that all individuals in the sample are equal. Despite the initial sample having an approximately equal probability of selection design, there were certainly differences in the probabilities of response in the first wave (Taylor 2010). The second assumption is that the sample has the same survival probabilities as the population, which may not be accurate in early waves because the sample was selected from private dwellings (and excludes people living in institutions such as nursing homes). It will be more accurate the longer the timeframe involved (Sadig's paper reported on the 18-year panel).

How many deaths are missing from the part of the sample that is no longer contacted? What is the best way to adjust for these deaths in constructing weights for the balanced panel from Wave 1 to Wave $t$? How do these missing deaths and weighting methods affect estimates for the population that die between Wave 1 and Wave $t$? How do they affect the estimates for individuals living in the population in Waves 1 through to Wave $t$? To answer these questions, data from the HILDA Survey is used. The HILDA Survey began with an approximately equal probability sample selected from private dwellings across Australia. The sample is clustered to allow for face-to-face interviewing. The Wave 1 household-level response rate was 66% (96% contact rate and 68% response rate given contact). Sample members are followed and those aged 15 and older are interviewed annually, collecting a broad range of socioeconomic data. While people living in nonprivate dwellings (i.e., institutions such as nursing homes) are excluded from the Wave 1 sample, sample members who move into nonprivate dwellings in Wave 2 or later are followed and interviewed. The sample is extended to include people who join a sample member's household for as long as they are co-resident, though some of these people are converted to permanent sample members (such as new births to permanent sample members). Four different methods to adjust for unknown deaths in a longitudinal survey are examined in this article. The first method matches the sample to the national death registry, and as a high match rate is achieved, this method serves as the gold standard to which the other methods are compared. The second method uses life-expectancy tables to extrapolate the expected number of deaths in the original sample over time. The third method models survival time from the observed sample and applies this to the nonresponding sample. The fourth method includes known deaths as responses in the nonresponse adjustment process over time. To assess the performance of each method against the death-register matching method, the number of deaths, the timing of these deaths, and the sociodemographic characteristics of those who die are compared. Various population estimates and model estimates are also compared for individuals living in the population from Wave 1 to 13, as these estimates may be impacted by differences in the portion of the Wave 1 population that is removed due to death by Wave 13.

The article proceeds as follows. Section 2 describes the four methods in detail. Section 3 compares the number and composition of actual and estimated deaths under each method. Section 4 examines the balanced panel weights produced and the implications for estimation are outlined in Section 5. Section 6 concludes the article.

## 2. Methods to Account for Deaths

This section describes four alternative methods for estimating the number of unknown deaths in the sample. The first three methods seek to identify or impute deaths in the part of the sample that is no longer contacted. The fourth method weights up the known deaths to allow for unknown deaths.

### *2.1. Method 1: Matching to Death Register*

Following the twelfth wave of data collection, the HILDA Survey sample was matched to the national death register. A name-based probabilistic linking strategy was used which matched on name (first name and surname), date of birth and sex (Australian Institute of Health and Welfare 2011). A series of passes were made over the data to allow for variations in the recorded data. Match weights were assigned to linked records based on how closely the records matched: high weights indicate the matched pairs have exact or very similar data and low weights indicate they have very different data. The main contribution to the match weight is made by the names, based on how common each name is and how close the names are in the matched pair. An agreement in day, month, year, or sex increases the weight by 1 for each component and a disagreement in day, month, or sex decreases the weight by 1 for each component. The weight is decreased by 1 for each year the date of birth disagrees. The weights are used to decide between multiple matches and the higher-weighted match is taken. The weights are also used to decide when a match is likely. To determine the cutoff point to use, the distribution of the match weights for the sample known to be alive in Wave 12 (i.e., they were part of a responding household) is compared to the known dead sample (i.e., reported to be dead in Wave 12 or earlier). Figure 2 shows the distribution of the match weights for the known live sample (grey line), the known dead sample (black line), and also the entire sample (dashed grey line). As an example, there were 1,191 matched records with a match weight of 11. This is made up of 863 cases among the known live sample at Wave 12, four cases among the known dead sample, and 324 cases where it was unknown whether they were dead or alive (this last group is not shown as a separate line on the graph). Almost all (98%) sample members whose status is known that have a match weight of 29 or more are known deaths. By comparison, for those sample members with a known status that have a match weight of 21 or less, only 0.5% are known deaths. As a result, the matches with a match weight above 21 were clerically reviewed for consistency with the date of last contact. Overall, 95% of the known (or reported) deaths could be matched to a death-register record. Included in

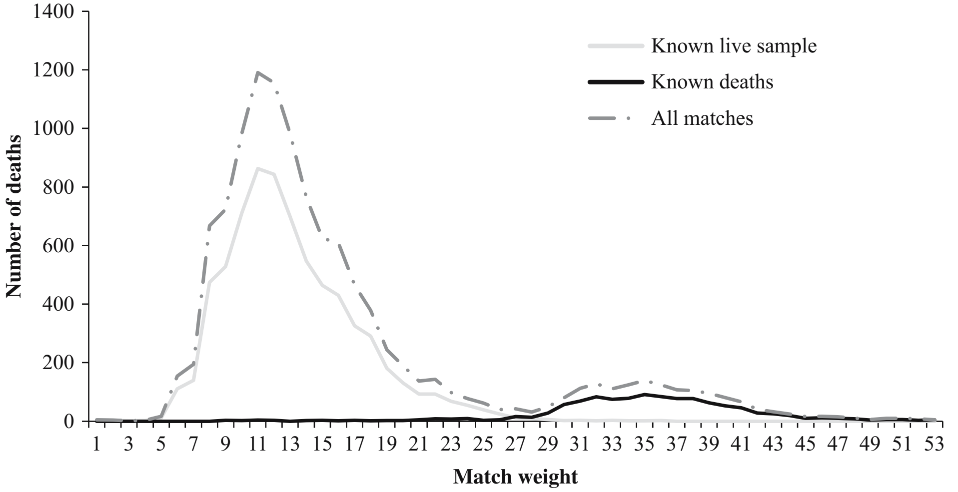


*Fig. 2. Death-registry match weights, all match pairs.*

this number were eight known deaths that had a match weight below 22 (amounting to 0.7% of all known deaths). Of the five percent of known deaths that could not be matched, the main reasons were as follows:

- they had died overseas,
- the surname did not match (and was not close), and
- the date of birth was not close.

This method adds a total of 265 (unweighted) deaths of Wave 1 respondents to the datasets by Wave 13, representing a 26% increase in the number of deaths recorded. This also increases the response rate of Wave 1 respondents re-interviewed in Wave 13 (excluding those considered out of scope in Wave 13) from 65% to 67%.

Matching to the death register will be superior to the other methods if the match rate is high. As the match rate of reported deaths to the death register is high for the HILDA sample, this method is used as the gold standard against which the other methods are compared. This method is more costly and time-consuming (in terms of the clerical work to review matches) than the other methods and may not be able to be undertaken every wave.

### 2.2. *Method 2: Using Life Tables*

The second method employs age- and sex-specific death rates from life tables produced annually by the Australian Bureau of Statistics (ABS 2014). These death rates are the proportion of deaths that occur between exact age $x$ and $x + 1$, and is denoted as $p_x$. To apply these death rates to the HILDA sample where a sample member is aged $x$ years and $m$ months at the date of last contact, a weighted combination of and $p_x$ and $p_{x+1}$ is used. That is, the probability that a person aged $x$ years and $m$ months at Wave $t$ dies between Wave $t$ and $t + 1$ is:

$$p_{txm} = \frac{(12 - m)}{12} p_x + \frac{m}{12} p_{x+1}$$

These rates are used to calculate the probability of a death between Wave $t$ and $t + n$ for a person aged $x$ years and $m$ months at Wave $t$ as:

$$p_{tnxm} = p_{txm} + \sum_{j=1}^{n-1} p_{t+j,x+j,m} \prod_{k=0}^{j-1} \left(1 - p_{t+k,x+k,m}\right)$$

That is, the probability of a death between Wave $t$ and $t + n$ is the sum of the probability of death between Wave $t$ and $t + 1$, the probability of death between Wave $t + 1$ to $t + 2$ for those who survive to Wave $t + 1$, the probability of death between Wave $t + 2$ to $t + 3$ for those who survive to Wave $t + 2$, and so on.

These rates of death are applied from the last wave in which a sample member is part of a *responding* household (where the household roster has been completed and at least one person has been interviewed). Each person is assigned a random number from the uniform distribution and is flagged as an imputed death in Wave $n$ if the random number is below $p_{tnxm}$. An alternative would be to assign death from the wave of last *contact* with a sample member (i.e., those lost to tracking, not issued, and issued but not contacted), but the interviewer may not speak with the sample member directly as this contact could have been with another member of the household or with some other person and the eligibility of the sample member may not have been ascertained. This method adds 279 deaths of Wave 1 respondents to the datasets by Wave 13, representing a 27% increase in the number of deaths recorded.

The life-table method is a relatively simple method to apply; however, life tables usually reflect survival rates in the population as a whole irrespective of whether the individuals live in institutions or not. If the sample has been selected in such a way as to exclude people living in institutions (as is often the case) or if the following rules preclude following people into institutions, then the general population survival rates may be too low (and the death rates too high) when applied to the sample in the short term.

### 2.3. *Method 3: Model Deaths Explicitly*

The third method uses the deaths observed through the fieldwork to model the survival curve for sample members. By using the available survey variables, it may be possible to predict deaths more accurately than by just using age and sex alone as in Method 2.

Survival time of Wave 1 sample members is modelled as a Weibull proportional hazards model. Sample members are designated to be alive, dead, or right censored (when the sample is not known to be either alive or dead). The survival time $t$ for individual $j$ is modelled as:

$$S_j(t) = \exp\left\{-\exp(\mathbf{X}_j\boldsymbol{\beta})t_j^p\right\}$$

where $\boldsymbol{X}$ is the matrix of covariates, $\boldsymbol{\beta}$ is the vector of fitted coefficients and $p$ is the Weibull shape parameter estimated from the data. Models for males and females are built separately. For each, a model for those belonging to a Wave 1 responding household (where only some basic information is available) is constructed and another for Wave 1 respondents (for whom more detailed information is available). The covariates considered for sample members belonging to a responding household are age, presence of long-term health condition, whether married/de facto married, and employment status. The covariates for Wave 1 respondents includes similar variables from the individual interview together with education and occupation. The models are unweighted, as the aim is to predict deaths in the sample rather than the population. If certain segments of the population had been oversampled (for example, immigrants), then variables could be included in the model or separate models could be developed to reflect this. This is not the case for the HILDA Survey and including weights make little difference to the results.

The fitted models are used to estimate the time to death for sample members in the last wave in which they are part of a responding household and were not known to have died at a later wave. The characteristics used for these sample members are those collected in the last responding wave. Each person is given a random number from the uniform distribution and when that number exceeds the predicted probability of survival in Wave $t + n$, the sample member is assumed to have died in Wave $t + n$. This method adds 262 deaths of Wave 1 respondents to the datasets by Wave 13, representing a 24% increase in the number of deaths recorded.

It is expected that by modelling deaths from the data internal to the survey, the limitations of the life-table method (Method 2) will be overcome. It should more accurately predict deaths in samples that do not align with the overall population (to which the life tables relate). It should also more accurately reflect the sociodemographic characteristics of those that die as it uses other characteristics besides age and sex to model the probability of death. This method assumes that there is reasonably good coverage of reported deaths within the portion of the sample recently contacted.

### 2.4. *Method 4: Model Deaths Implicitly*

The fourth method implicitly models deaths when adjusting for attrition in the weighting process. In the ideal world of 100% response rates, the eligibility of all sample members would be known in all waves. In reality, with response and contact rates less than 100%, sample members are known to be alive (as they are part of a responding household), known to be dead, or are not known to be either dead or alive (as they are part of a nonresponding household). When adjusting for nonresponse in the construction of the balanced panel of responding persons in Waves 1 through 13, unknown deaths can *implicitly* be taken account of by classifying the sample into two groups:

- Response – interviewed in Wave 1, and then interviewed, overseas, or dead in Waves 2 to 13, or
- Nonresponse – all other individuals interviewed in Wave 1.

This method of counting sample members who die or are not eligible to be interviewed (because they move overseas or, in some surveys, move into institutions) as responses was suggested by Kalton and Brick (2000).

Note that there are a number of different ways that weights can be developed for longitudinal surveys and Kalton and Brick (2000) provide an overview of the various choices that can be made. Different methods can be used to calculate nonresponse adjustments (via weighting classes, classification trees, logistic regression, or generalised raking), or how these adjustments are applied over time (wave by wave, in a single step, or in segments of waves). There are also different choices regarding poststratification adjustments, including whether or not they are made, and if made, to which population they refer (the population as at Wave 1, or some adjustment to this population to exclude deaths, emigration and immigration to reflect the live longitudinal population from Wave 1 to Wave *t*). It is beyond the scope of this article to consider how all of these choices interact with the ways in which unknown eligibility may be handled. Rather, the method for dealing with unknown eligibility has been varied in this article while holding the remaining aspects of the weighting strategy constant. It is not expected that these choices in other aspects of the weighting strategy would greatly affect the results.

The particular weighting strategy adopted in the HILDA Survey is as follows. The probability of response is modelled on numerous Wave 1 characteristics including age, sex, marital status, ability to speak English, employment status, hours worked, education level, health status, number of times moved in last ten years, interviewer-rated cooperation, whether the self-completion component was returned, geographical location, remoteness area, number of people in the household, household type, condition of dwelling, number of calls made, and whether the household was partially responding. Some post-Wave 1 mobility information was included when available. The balanced panel weight $w_{r1\_13}$ for responding persons in Waves 1 through 13 is taken to be the Wave 1 responding person weight $w_{r1}$ divided by this probability of response in Waves 2 through 13 $p_{r2\_13}$:

$$w_{r1\_13} = \frac{w_{r1}}{p_{r2\_13}}$$

As a result, those respondents who share similar characteristics with nonrespondents receive a higher adjusted weight than those who do not. These weights are then calibrated to the Wave 1 benchmarks for age, sex, labour-force status, social marital status, household composition, and geographic area.

To obtain a weighted estimate of the sample deaths, this balanced panel population weight is rescaled to the Wave 1 responding sample size. This method estimates an additional 317 deaths by Wave 13, representing a 30% increase in the weighted number of known deaths (weighted by the Wave 1 responding person weight rescaled to the sample size).

It is expected that this method would produce similar estimates to the death-modelling method (Method 3) as they both use data internal to the survey.

## 3. Comparison of Methods

A comparison of the number of deaths reported, matched, or estimated by the different methods is given in Table 1. Part 1 of the table provides the weighted estimate of the number of deaths cumulated from Wave 2 to each subsequent wave. The weight used to create the estimate for all methods except Method 4 (weighting) is the Wave 1 responding person weight rescaled to the sample size. The weight used for Method 4 is the balanced panel weight for Waves 1 through to Wave *t* calculated via that method (as described in Subsection 2.4). Part 2 of the table shows the difference in the number of deaths estimated via Method 1 to the other methods. Part 3 provides the standard errors for these differences; these are calculated using jackknife estimation from the 45 replicates

*Table 1. Cumulative number of deaths by wave and method.*

| | Wave | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| *1. Estimated number of deaths* | | | | | | | | | | | | |
| Reported deaths | 56 | 159 | 270 | 368 | 455 | 542 | 629 | 710 | 782 | 867 | 963 | 1036 |
| Method 1: Death register | 71 | 178 | 291 | 403 | 506 | 610 | 728 | 829 | 940 | 1060 | 1204 | 1325 |
| Method 2: Life tables | 66 | 179 | 309 | 433 | 544 | 656 | 757 | 874 | 971 | 1089 | 1232 | 1342 |
| Method 3: Model deaths | 63 | 170 | 294 | 402 | 516 | 630 | 745 | 861 | 966 | 1078 | 1210 | 1333 |
| Method 4: Weighting | 64 | 166 | 284 | 377 | 483 | 597 | 703 | 831 | 953 | 1077 | 1231 | 1353 |
| *2. Difference from Method 1* | | | | | | | | | | | | |
| Reported deaths | −15 | −18 | −21 | −35 | −51 | −68 | −100 | −119 | −158 | −193 | −242 | −289 |
| Method 2: Life tables | −5 | 1 | 17 | 30 | 38 | 46 | 29 | 44 | 30 | 29 | 27 | 17 |
| Method 3: Model deaths | −8 | −7 | 3 | −1 | 10 | 19 | 17 | 31 | 25 | 18 | 6 | 8 |
| Method 4: Weighting | −8 | −11 | −8 | −26 | −23 | −14 | −25 | 2 | 13 | 16 | 27 | 28 |
| *3. Standard error of difference* | | | | | | | | | | | | |
| Reported deaths | 4.2 | 4.5 | 4.4 | 5.8 | 6.8 | 9.0 | 9.8 | 12.5 | 12.4 | 15.5 | 19.0 | 21.0 |
| Method 2: Life tables | 5.0 | 7.0 | 8.8 | 11.1 | 13.2 | 16.3 | 15.4 | 18.3 | 16.4 | 18.2 | 19.6 | 22.4 |
| Method 3: Model deaths | 6.2 | 6.0 | 6.9 | 7.3 | 9.4 | 11.3 | 12.4 | 12.1 | 15.4 | 17.5 | 18.4 | 19.2 |
| Method 4: Weighting | 3.9 | 10.0 | 12.4 | 14.2 | 16.1 | 18.6 | 19.6 | 22.3 | 24.0 | 27.1 | 35.7 | 33.8 |
| *4. P(difference = 0)* | | | | | | | | | | | | |
| Reported deaths | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** |
| Method 2: Life tables | 0.310 | 0.903 | **0.048** | **0.008** | **0.004** | **0.005** | *0.062* | **0.015** | *0.066* | 0.116 | 0.163 | 0.441 |
| Method 3: Model deaths | 0.201 | 0.218 | 0.692 | 0.883 | 0.287 | *0.084* | 0.171 | **0.009** | 0.102 | 0.305 | 0.754 | 0.672 |
| Method 4: Weighting | *0.051* | 0.254 | 0.532 | *0.069* | 0.155 | 0.469 | 0.200 | 0.933 | 0.601 | 0.543 | 0.452 | 0.404 |

Note: Numbers are weighted by the population weight rescaled to number of Wave 1 respondents ($n = 13{,}969$). For all except Method 4, the weight is the Wave 1 responding person population weight. For Method 4, the weight is the balanced panel weight for Wave 1 to Wave $t$, where $t$ is the wave number for the column. P-values in bold and italics indicate the estimate is significantly different from Method 1 at the 5% and 10% level respectively.

routinely used to construct the HILDA weights. Part 4 of the table reports the probability that these differences are equal to zero.

The death-register matching approach (Method 1) adds 68 new deaths (weighted) by Wave 7 and 289 by Wave 13 to the number of deaths already reported in the sample. As not all of the known deaths were matched (five percent were unmatched), it would be reasonable to expect there to be a similar portion of deaths not matched to the death register among sample members with unknown eligibility. This would translate to about four further deaths by Wave 7 and 15 by Wave 13, suggesting the true number of missing deaths is about 72 in Wave 7 and 304 in Wave 13. Nevertheless, as the death-register matching was of reasonably high quality, the other methods are compared to this method. The life-table method (Method 2) significantly overestimates the number of missing deaths by Wave 4 and does not come back into line with the death-register results until Wave 10. This is likely to be a result of selecting the initial sample from private dwellings (thus excluding nursing homes and other institutions), whereas the life tables relate to all people living in Australia. This healthy sample bias disappears after a decade as the next cohort of sample members age and are followed into these institutions prior to death. Modelling the deaths observed in the data (Method 3) or implicitly modelling the deaths through the attrition adjustment (Method 4) produce very similar results to those obtained by matching to the death register. Only by Wave 9 is the cumulative number of deaths higher when modelling the deaths (Method 3) compared to the death-register matching method (at the 5% significance level).

The four methods differ somewhat in the distribution of deaths over time. Figure 3 shows the year of death for reported deaths and the total number of deaths for the four methods. As more of the sample is withheld from fieldwork over time, a greater portion of deaths are missed in later waves. This is seen in the light grey dotted line, which has the highest number of deaths in 2003 and then generally falls over time after that. The number of deaths in 2002 is lower than in 2003 as the initial sample is selected from private dwellings and the respondents are well enough to take part in an interview in 2001. Matching to the death register helps correct the number of deaths to a generally more stable number over time from 2004. Note that the death-register matching results in a lower number of deaths recorded for 2003 than for the reported deaths. This is because the *exact* timing of these deaths is not always reported. For about one third of the deaths identified in the field, all that is known about the timing of the death is that the death was not reported at the last wave and is reported at the current wave. A quarter of those deaths are assumed to have occurred in the previous year and three quarters in the current year (as approximately 50% of the interviews are completed by September). Of the cases where a year of death was reported, only three percent did not match the year of death recorded in the death register. In terms of the other methods, the life-table method (Method 2) and the death-modelling method (Method 3) distributes the deaths in a fairly similar manner to the death-register matching, though there are significantly more deaths in 2003 with the life-table method. The weighting method (Method 4), where the balanced panel weight for Waves 1 through to 13 is applied to the known deaths, shows substantially more deaths in early years and fewer in the later years. This is a result of greater coverage of deaths (i.e., known deaths compared to all deaths) in the early waves and less in the later waves. This may be ameliorated by modelling attrition on a wave-by-wave basis rather than in one step

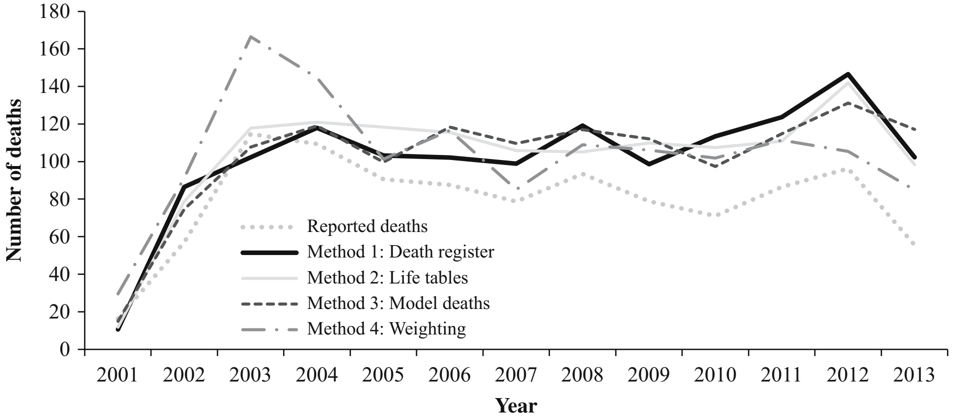


*Fig. 3. Year of death of Wave 1 respondents (weighted).*

from Wave 1, however this may also increase the variability of the weights. Alternatively, a benchmark for the number of additional deaths estimated via each subsequent balanced panel could be included in the poststratification adjustment step of the weighting process as the cumulative number of deaths reported in Table 1 are consistent with the number of deaths identified from matching to the death register.

Aside from the total number and distribution of the deaths over time, how well do the methods reflect the sociodemographic characteristics and health conditions (as reported at Wave 1) of those who subsequently die? Table 2 provides a summary of the (weighted) Wave 1 characteristics of the reported deaths compared to the deaths identified via the four methods by Wave 13. Compared to the death-register matching (which is expected to be the closest to the truth), the people reported to have died are more likely to be younger, male, live in a two-adult household, have a long-term health condition, or have a higher mental health score. These characteristics to some extent reflect the household situations in which a death is more likely to be reported, as another sample member can report on the death. Deaths of older females living alone are less likely to be reported. For the two methods that impute missing deaths based on respondent characteristics (Methods 2 and 3), the additional deaths are likely to be (at Wave 1) younger, born in countries where the main language spoken is not English, residing in a capital city, have higher physical functioning scores, and higher mental health scores, and be less likely to live in single-person households than those identified via the death-register matching. This suggests that, while these two methods achieve the same number of deaths by Wave 13 as the death-register matching, they are not sufficiently discriminating in the types of people imputed as deaths over the life of the panel. For the twelve characteristics considered in Table 2, Method 4 (weighting) is indistinguishable from Method 1.

## 4. Weights

To explore the implications of these different methods of dealing with unknown deaths in a longitudinal survey, a set of weights for the balanced panel from Wave 1 through to Wave 13 is produced for each method. The weighting process for Method 4 is described

*Table 2. Mean characteristics of deaths.*

| | Known deaths | | Method 1: Death register | Method 2: Life tables | | Method 3: Model deaths | | Method 4: Weighting | |
|---|---|---|---|---|---|---|---|---|---|
| Wave 1 characteristic | Est | P(Rpt-M1 = 0) | Est | Est | P(M2-M1 = 0) | Est | P(M3-M1 = 0) | Est | P(M4-M1 = 0) |
| Age (years) | 69.4 | **0.006** | 70.0 | 69.1 | **0.000** | 68.8 | **0.000** | 69.6 | 0.266 |
| Female (%) | 43.2 | **0.008** | 45.6 | 45.9 | 0.777 | 44.4 | 0.229 | 44.9 | 0.537 |
| Born abroad in main English-speaking country (%) | 13.5 | 0.645 | 13.3 | 13.4 | 0.923 | 13.5 | 0.720 | 13.2 | 0.848 |
| Born abroad not in main English-speaking country (%) | 13.3 | 0.109 | 14.5 | 16.1 | *0.082* | 17.1 | **0.003** | 14.2 | 0.848 |
| Single-adult household (%) | 30.3 | 0.351 | 31.1 | 29.6 | **0.025** | 29.0 | **0.002** | 31.1 | 0.967 |
| Two-adult household (%) | 55.0 | **0.003** | 52.4 | 52.4 | 0.968 | 53.7 | 0.151 | 52.4 | 0.989 |
| Three or more adult household (%) | 14.6 | **0.033** | 16.5 | 17.9 | 0.174 | 17.3 | 0.397 | 16.6 | 0.990 |
| Married/defacto (%) | 60.2 | 0.295 | 59.2 | 60.3 | 0.215 | 61.0 | *0.051* | 58.1 | 0.459 |
| Lives in capital city (%) | 56.3 | 0.602 | 55.9 | 58.2 | **0.006** | 58.6 | **0.002** | 55.8 | 0.953 |
| Has long-term health condition (%) | 58.8 | **0.009** | 56.6 | 55.4 | 0.216 | 55.7 | 0.366 | 58.5 | 0.180 |
| SF-36 Physical functioning score (0–100) | 56.4 | 0.411 | 56.0 | 57.5 | **0.002** | 57.7 | **0.001** | 56.3 | 0.716 |
| SF-36 Mental health score (0–100) | 73.7 | **0.040** | 73.1 | 73.9 | **0.007** | 73.8 | **0.013** | 74.0 | *0.084* |

Note: Bold and italicised *p*-values indicate the estimate is significantly different from the Method 1 estimate at 1% level and 5% level respectively.

*Table 3. Summary of weights by method.*

| | All respondents (aged 15+ in Wave 1) | | | All respondents (aged 65+ in Wave 1) | | |
|---|---|---|---|---|---|---|
| Weight description | Mean | StdDev | CV | Mean | StdDev | CV |
| Method 0: Known death as out of scope, unknown death as nonresponse | 1898.2 | 1226.2 | 0.646 | 2014.0 | 1294.8 | 0.643 |
| Method 1: Death register | 1852.9 | 1199.3 | 0.647 | 1731.4 | 1113.8 | 0.643 |
| Method 2: Life table | 1851.3 | 1196.1 | 0.646 | 1704.6 | 1081.5 | 0.634 |
| Method 3: Model death | 1856.3 | 1201.9 | 0.647 | 1760.4 | 1132.4 | 0.643 |
| Method 4: Known death as response | 1848.2 | 1191.9 | 0.645 | 1699.7 | 1097.2 | 0.646 |

earlier in Subsection 2.4 and the other three methods follow the same process, with the exception of how deaths are treated. For these three methods, deaths that are reported, matched (as in Method 1), or imputed (as in Methods 2 and 3) do not have an attrition adjustment applied. They do, however, have a separate benchmark applied and this is calculated as the weighted sum of deaths (using the Wave 1 cross-sectional weight). The small portion of reported deaths (five percent) not matched to the death register in Method 1 are weighted according to the procedure adopted for Method 4. Another method is also considered here which treats known deaths as out of scope and unknown deaths as being alive nonrespondents (denoted as Method 0). This method is included as it is explicitly or implicitly adopted by a number of longitudinal studies, as mentioned earlier.

Table 3 provides the mean, standard deviation, and coefficient of variation (= standard deviation/mean) of the weights for the individuals living in the population from Wave 1 to 13. Note that the mean of the weights is not the same across all methods, as different numbers of deaths are assumed under each method. The coefficient of variation is a useful measure to compare across weights with different means as it is not unit specific. Weights with a lower coefficient of variation are preferred over those with a higher one (assuming no difference in the bias of resulting estimates) as the estimates will be more efficient. These summary statistics are provided for all adults and for those aged 65 and over.

Under Method 0, the mean of the weights for the older members of the population are substantially higher as it is assumed that all unknown deaths are alive nonrespondents and respondents with similar characteristics (such as being of a similar age) receive higher weights. All methods are very similar in terms of their coefficient of variation.

## 5. Impact on Analysis

In this section, various population estimates and model estimates are compared across the five weighting schemes. Table 4 provides weighted estimates for Wave 13 (2013) of the proportion of people with a long-term health condition, self-rated health status (“excellent”, “very good”, “good”, “fair”, and “poor”), health satisfaction, life satisfaction, and a rating of how much their health limits the work they can do (including paid work, work around the house or garden). The weights used here are the balanced panel weights for the

Table 4. Weighted estimates for various health measures at Wave 13, by age.

| | Method 0 | Method 1 | Method 2 | Method 3 | Method 4 |
|---|---|---|---|---|---|
| Aged 15+ | | | | | |
| Has long-term health condition (%) | 37.3 | 36.5 | 36.5 | 36.6 | 36.4 |
| Health status (%) | | | | | |
| Excellent | 8.5 | 8.6 | 8.6 | 8.6 | 8.6 |
| Very good | 33.8 | 34.1 | 34.1 | 34.1 | 34.2 |
| Good | 37.0 | 37.0 | 36.9 | 37.0 | 37.0 |
| Fair | 16.7 | 16.3 | 16.3 | 16.4 | 16.2 |
| Poor | 4.1 | 4.0 | 4.0 | 4.1 | 4.0 |
| Health satisfaction (0–10) | 7.0 | 7.0 | 7.0 | 7.0 | 7.0 |
| Life satisfaction (0–10) | 7.9 | 7.9 | 7.8 | 7.9 | 7.9 |
| Health limitation (0–10) | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| Aged 65+ | | | | | |
| Has long-term health condition (%) | 70.9 | 70.4 | 70.3 | 70.5 | 70.3 |
| Health status (%) | | | | | |
| Excellent | 2.2 | 2.0 | 2.0 | 2.1 | 2.0 |
| Very good | 20.3 | 20.7 | 20.8 | 20.6 | 20.9 |
| Good | 38.8 | 38.7 | 38.6 | 38.7 | 38.7 |
| Fair | 32.1 | 31.8 | 31.9 | 32.0 | 31.8 |
| Poor | 6.6 | 6.7 | 6.7 | 6.7 | 6.6 |
| Health satisfaction (0–10) | 6.8 | 6.8 | 6.8 | 6.8 | 6.8 |
| Life satisfaction (0–10) | 8.2 | 8.2 | 8.2 | 8.2 | 8.2 |
| Health limitation (0–10) | 3.4 | 3.4 | 3.4 | 3.4 | 3.4 |

Note: Method 0 weights unknown death as nonresponse. Method 1 obtains extra deaths from death register. Method 2 imputes extra deaths based on life tables. Method 3 imputes extra deaths from model of deaths in sample. Method 4 weights known deaths as response.

13-wave panel. Estimates are provided for all people aged 15 and over, and for those aged 65 and over.

There are no significant differences between the means calculated under the death-register method and any of the other methods. The estimate that is closest to being significantly different is the proportion of people aged 65 and over reporting 'very good' health for weights calculated with no adjustment versus those with the death-register matching (Method 0 vs Method 1; difference=0.4%, $p$-value=0.149).

Next the effect that the various death-adjustment methods have on two weighted random-effect logistic models is considered; one model is focused on life satisfaction and the other model examines the extent of health limitations. It is possible for the marginal distribution of these outcome variables (i.e., conditional on covariates in the model) to be affected by the weighting method even though the unconditional means are unaffected. Both models considered here have been restricted to older adults (aged 65+) as the impact of how unknown deaths are handled in the weights (if there is any) is likely to be greatest among this group due to their higher rate of death. The models were fitted in Stata using the *gllamm* command which allows for probability weights in the calculation of the random effects logistic model. The dependent variable for the first model is high life satisfaction (i.e., a score of eight or higher out of ten), and the dependent variable for the second model is a rating of eight or higher out of ten in terms of how limiting the respondent's health condition is on the work they can do. The covariates include age, sex, broad country of birth (indicators of whether born overseas in a mainly English-speaking country or not), real equivalised household income, marital status, education level, region, whether another adult was present during the interview, and (for the first model only) presence of a long-term health condition. These variables were included following other studies of life satisfaction and health limitations (Green 2011; Au and Johnston 2014). The weights used here are the balanced panel weights for the 13-wave panel. The standard errors are calculated using the replicate weights and thus take account of the complex sample design. Table 5 presents the regression coefficients for the life-satisfaction model. The significant coefficients differ by at most seven percent from the baseline model (Model 0). Which weighting method is adopted is irrelevant to this model: the coefficients for the significant variables for Models 2, 3, and 4 differ by less than two percent compared to those for the death-register method (Method 1). None of these differences are significant. The coefficient closest to being significantly different from the Method 1 coefficients is the indicator variable for remote Australia for Method 0 (diff = 0.187, $p$-value = 0.119).

In the model of health limitations presented in Table 6, the significant coefficients differ by at most 13% from the baseline model (Method 0). Again, there is little difference resulting from the method used to adjust for unknown deaths, with significant coefficients changing by no more than five percent. For the covariate sex, the coefficient becomes marginally significant under Method 4 (weighting), reflecting only a small change in the $p$-value for the coefficient which is close to the ten-percent significance level in the other models. None of the differences in the coefficients of these models are significantly different from zero. The coefficient closest to being significantly different from Method 1 is for age in Method 0 (diff = −0.079, $p$-value = 0.180).

An alternative multilevel modelling approach proposed by Skinner and Holmes (2003) could have been used here. This method incorporates all balanced panels from Wave 1, not just the balanced panel from Wave 1 to 13, and thus increases the number of person-year observations used in the models by nearly 80%. Under this modelling approach, the Wave 1 weight is used as the individual (Level 2) weight and a rescaled version of the balanced panel weights is used for the wave observations (Level 1). The wave-observation (Level 1) weights are rescaled so that the sum of these weights equals the number of wave observations for an individual (Scaling Method 2 in Rabe-Hesketh and Skrondal 2006). The model parameters under this approach are more robust to the particular choice of weighting method (Method 0–4) adopted. The parameter estimates differ by at most two percent in the model of life satisfaction from the baseline method (Method 0) and at most five percent in the model of health limitations (results not shown).

## 6. Summary and Conclusions

This article examined four different methods for adjusting for unknown deaths in the construction of weights for a longitudinal sample. Matching the sample to the death register will produce the most accurate results assuming the matching process is of high quality. It will also provide the opportunity to add further information to the dataset, such

*Table 5. Random effects logistic regression coefficients for high life satisfaction (>=8 out of 10), aged 65+.*

| | Method 0 b (se) | Method 1 b (se) | Method 2 b (se) | Method 3 b (se) | Method 4 b (se) |
|---|---|---|---|---|---|
| Age/10 | 1.309 | 0.627 | 0.860 | 0.886 | 0.633 |
| | (1.857) | (1.848) | (1.959) | (1.890) | (1.817) |
| Age squared/100 | −0.097 | −0.053 | −0.069 | −0.070 | −0.054 |
| | (0.121) | (0.121) | (0.129) | (0.124) | (0.119) |
| Sex | 0.153 | 0.120 | 0.119 | 0.127 | 0.102 |
| | (0.258) | (0.261) | (0.266) | (0.267) | (0.224) |
| Born overseas, main English-speaking country | −0.099 | −0.087 | −0.088 | −0.094 | −0.070 |
| | (0.340) | (0.360) | (0.349) | (0.362) | (0.313) |
| Born overseas, not main English-speaking country | −1.146*** | −1.192*** | −1.196*** | −1.187*** | −1.183*** |
| | (0.337) | (0.348) | (0.342) | (0.360) | (0.334) |
| Real equiv. household income/100,000 | 1.740** | 1.820** | 1.840** | 1.800** | 1.810** |
| | (0.774) | (0.785) | (0.791) | (0.783) | (0.774) |
| Real equiv. household income squared/100,000$^2$ | −0.695 | −0.728 | −0.744 | −0.711 | −0.701 |
| | (0.622) | (0.624) | (0.629) | (0.623) | (0.616) |
| Separated / divorced | −0.622** | −0.661** | −0.658** | −0.641** | −0.666** |
| | (0.280) | (0.304) | (0.328) | (0.313) | (0.279) |
| Widowed | −0.420** | −0.421** | −0.428** | −0.419** | −0.415** |
| | (0.181) | (0.196) | (0.195) | (0.197) | (0.174) |
| Single | −0.748 | −0.675 | −0.690 | −0.717 | −0.712 |
| | (0.654) | (0.661) | (0.672) | (0.677) | (0.550) |
| Graduate | −0.905 | −0.903 | −0.889 | −0.901 | −0.887 |
| | (0.774) | (0.811) | (0.828) | (0.833) | (0.728) |
| Year 11 and below | −0.292 | −0.265 | −0.271 | −0.276 | −0.275 |
| | (0.209) | (0.224) | (0.228) | (0.233) | (0.209) |
| Has long-term health condition | −0.601*** | −0.591*** | −0.594*** | −0.593*** | −0.593*** |
| | (0.140) | (0.153) | (0.146) | (0.148) | (0.125) |

*Table 5. Continued.*

| | Method 0 b (se) | Method 1 b (se) | Method 2 b (se) | Method 3 b (se) | Method 4 b (se) |
|---|---|---|---|---|---|
| Regional Australia | −0.066 | −0.069 | −0.064 | −0.068 | −0.084 |
| | (0.210) | (0.232) | (0.229) | (0.233) | (0.227) |
| Remote Australia | −0.143 | −0.331 | −0.324 | −0.281 | −0.304 |
| | (0.412) | (0.441) | (0.487) | (0.446) | (0.433) |
| Other adult present during interview | 0.076 | 0.099 | 0.104 | 0.099 | 0.110 |
| | (0.140) | (0.137) | (0.134) | (0.138) | (0.135) |
| Constant | −1.460 | 1.143 | 0.272 | 0.161 | 1.173 |
| | (6.950) | (6.982) | (7.405) | (7.154) | (6.844) |
| Number of person-year observations | 8796 | 8796 | 8796 | 8796 | 8796 |

Notes: 1. Method 0 weights unknown death as nonresponse. Method 1 obtains extra deaths from death register. Method 2 imputes extra deaths based on life tables. Method 3 imputes extra deaths from model of deaths in sample. Method 4 weights known deaths as response.
2. ***, **, and * represent significance at 1% level, 5% level and 10% level respectively.

*Table 6. Random effects logistic regression coefficients for work-limiting health condition (>=8 out of 10). aged 65+.*

| | Method 0 b (se) | Method 1 b (se) | Method 2 b (se) | Method 3 b (se) | Method 4 b (se) |
|---|---|---|---|---|---|
| Age/10 | 3.562 | 4.352 | 4.321 | 4.123 | 4.371 |
| | (2.999) | (2.971) | (3.230) | (3.177) | (2.756) |
| Age squared/100 | −0.119 | −0.168 | −0.167 | −0.155 | −0.169 |
| | (0.191) | (0.189) | (0.206) | (0.203) | (0.174) |
| Sex | −0.435 | −0.434 | −0.441 | −0.442 | −0.414* |
| | (0.288) | (0.269) | (0.289) | (0.273) | (0.226) |
| Born overseas, main English-speaking country | −0.262 | −0.270 | −0.268 | −0.265 | −0.283 |
| | (0.333) | (0.360) | (0.354) | (0.348) | (0.329) |
| Born overseas, not main English-speaking country | 1.145*** | 1.193*** | 1.194*** | 1.191*** | 1.201*** |
| | (0.314) | (0.333) | (0.328) | (0.334) | (0.300) |
| Real equiv. household income/100,000 | 0.519 | 0.224 | 0.183 | 0.302 | 0.150 |
| | (1.220) | (1.190) | (1.180) | (1.190) | (1.180) |
| Real equiv. household income squared/100,000$^2$ | −1.050 | −0.865 | −0.831 | −0.919 | −0.827 |
| | (1.430) | (1.410) | (1.400) | (1.410) | (1.420) |
| Separated/divorced | −0.668 | −0.656 | −0.662 | −0.660 | −0.666 |
| | (0.420) | (0.416) | (0.434) | (0.420) | (0.414) |
| Widowed | 0.005 | 0.021 | 0.052 | 0.033 | 0.026 |
| | (0.320) | (0.317) | (0.312) | (0.310) | (0.296) |
| Single | 0.171 | 0.196 | 0.211 | 0.210 | 0.265 |
| | (0.788) | (0.841) | (0.847) | (0.834) | (0.673) |
| Graduate | 0.359 | 0.393 | 0.395 | 0.387 | 0.404 |
| | (0.504) | (0.510) | (0.486) | (0.503) | (0.486) |
| Year 11 and below | 0.771*** | 0.774*** | 0.781*** | 0.785*** | 0.806*** |
| | (0.253) | (0.276) | (0.269) | (0.283) | (0.244) |
| Regional Australia | 0.254 | 0.279 | 0.264 | 0.271 | 0.292 |
| | (0.326) | (0.311) | (0.328) | (0.322) | (0.317) |

*Table 6. Continued.*

| | Method 0 b (se) | Method 1 b (se) | Method 2 b (se) | Method 3 b (se) | Method 4 b (se) |
|---|---|---|---|---|---|
| Remote Australia | −0.079 | 0.048 | 0.060 | 0.010 | 0.034 |
| | (1.065) | (1.148) | (1.346) | (1.215) | (1.255) |
| Other adult present during interview | 0.174 | 0.129 | 0.125 | 0.141 | 0.127 |
| | (0.192) | (0.209) | (0.202) | (0.214) | (0.202) |
| Constant | −24.064** | −27.140** | −26.936** | −26.215** | −27.290** |
| | (11.92) | (11.85) | (12.79) | (12.56) | (10.97) |
| Number of person-year observations | 8784 | 8784 | 8784 | 8784 | 8784 |

Notes: 1. Method 0 weights unknown death as nonresponse. Method 1 obtains extra deaths from death register. Method 2 imputes extra deaths based on life tables. Method 3 imputes extra deaths from model of deaths in sample. Method 4 weights known deaths as response.
2. ***, **, and * represent significance at 1% level, 5% level and 10% level respectively.

as a more accurate date of death and cause of death. Nonetheless, there are some difficulties in using the death register. In the HILDA Survey sample, five percent of the known (or reported) deaths were unmatched, suggesting that among the unknown deaths there would be a similar percentage of deaths not identified via this method. By Wave 13, this would amount to approximately 15 missed deaths and these would most likely be recently separated or divorced women who change their surname or people who died after moving overseas. Some adjustment in the construction of the weights would need to be made for these types of people.

Where matching to the death register is not an option or is of low quality, an alternative method will need to be adopted to improve estimates of the number of deaths, the timing of the deaths, and the sociodemographic characteristics of those that die since the start of the panel. The life-table method (Method 2) tends to overestimate deaths in the HILDA Survey sample in the short term (up to around Wave 9) for two reasons. First, the initial sample was selected from private dwellings only, so excludes institutions such as nursing homes. Second, the sample members also needed to be well enough to participate in the Wave 1 interview. For both of these reasons, the sample would be in better health on average than the overall population to which the life tables relate. Modelling the deaths from data internal to the survey (Method 3) proves to be a reasonably good short-term solution, at least in terms of the number of deaths. Unfortunately, the life-table method and the modelling-death method do not adequately reflect the sociodemographic characteristics of people who die according to the death register. Weighting the known deaths as responses provides a good solution for both the number of deaths and the sociodemographic characteristics of those that die, though the standard errors for these estimates are a little higher than the other methods (as a smaller number of cases is being given a higher weight under this method than under the other methods). Method 4, with additional benchmarks to improve the timing of the deaths, is recommended over Methods 2 or 3.

Estimates for the live population from Wave 1 to 13 were unaffected by the method chosen to adjust for the unknown deaths, at least among the variables and the time frame considered in this article. It is plausible that over a longer time frame, or for a survey with a higher rate of unknown eligibility than the HILDA Survey, the estimates for the live population would begin to be affected (as was observed by Sadig's (2014) analysis of an 18-wave panel). This is because the Wave 1 population is split into several population segments by Wave *t*:

i) the part of the population who die by Wave *t*;
ii) the part of the population who move overseas in at least one wave by Wave *t*; and
iii) the remaining portion of the Wave 1 population who continuously live in the population to Wave *t*.

Differences in the characteristics of one of these groups (as seen for those who die) would eventually result in differences in the other groups as the overall size of the Wave 1 population is fixed.

What is clear from this analysis is that assuming sample members with unknown eligibility are all eligible leads to inaccurate estimates of the total number of people who die, the timing of their death, and their sociodemographic characteristics. Similar findings are expected for other longitudinal surveys and could be confirmed through replication of these analyses. The impact may be greater in another study if a smaller proportion of deaths is reported or if these deaths are more dissimilar to the unknown deaths than found in the HILDA Survey. This is a problem that only grows over time as more of the sample is withheld from fieldwork for various reasons. Some adjustment to address this issue, either implicitly or explicitly, should be made.

While this article considered deaths specifically, this same issue of unknown eligibility also applies to people who move out of scope in other ways. For the HILDA Survey, this applies to people who move abroad, but for other surveys this could include people who move into institutions. A national registry for these transitions out of scope is unlikely to be available for matching purposes, and the findings presented here may help guide decisions on alternative methods. Aggregate statistics on transitions may be available (such as from Census or migration statistics) which could be used in a similar way to life tables in Method 2. It would certainly be feasible to use the data internal to the survey to model moves abroad or to institutions (as in Method 3) or to weight known out of scopes as responses (as in Method 4). Indeed, for those surveys of specific subgroups in the population it may not be possible to use any external data sources at all, so any adjustment for unknown eligibility would be limited to methods that use internal data (such as Methods 3 and 4). Further work could include a similar assessment for other types of unknown eligibility as presented here for unknown deaths.

## 7. References

Au, N. and D.W. Johnston. 2014. "Self-Assessed Health: What Does it Mean and What Does it Hide?" *Social Science and Medicine* 121: 21–28. Doi: http://dx.doi.org/10.1016/j.socscimed.2014.10.007.

Australian Bureau of Statistics (ABS). 2014. *Life Tables, States, Territories and Australia, 2011-2013*. Canberra: ABS. Available at: http://www.abs.gov.au (accessed January 2015).

Australian Institute of Health and Welfare. 2011. *Comparing an SLK-Based and a Name-Based Data Linkage Strategy: An Investigation into the PIAC Linkage*. Canberra: AIHW. Available at: http://www.aihw.gov.au/publication-detail/?id=10737418279 (accessed November 2015).

Gouskova, E., S. Heeringa, K. McGonagle, R. Schoeni, and F. Stafford. 2008. "Panel Study of Income Dynamics Revised Longitudinal Weights 1993-2005." PSID Technical Series Paper 08-05. Ann Arbor: University of Michigan. Available at: https://psidonline.isr.umich.edu/publications/Papers/tsp/2008-05_PSID_Revised_Longitudinal_Weights_1993-2005%20.pdf (accessed 30 June 2014).

Green, F. 2011. "Unpacking the Misery Multiplier: How Employability Modifies the Impacts of Unemployment and Job Insecurity on Life Satisfaction and Mental Health." *Journal of Health Economics* 30: 265–276. Doi: http://dx.doi.org/10.1016/j.jhealeco.2010.12.005.

Hawkes, D. and I. Plewis. 2006. "Modelling Non-Response in the National Child Development Study." *Journal of the Royal Statistical Society, Series A* 169: 479–491. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00401.x.

Hayward, M.D. Undated. "Using the Health and Retirement Survey to investigate health discrepancies." Health and Retirement Survey documentation. Ann Arbor: University of Michigan. Available at: http://hrsonline.isr.umich.edu/sitedocs/dmc/hrs_healthdisparities_hayward.pdf (accessed February 2015).

Health and Retirement Survey (HRS). Undated. "Sampling Weights: Revised for Tracker 2.0 and Beyond." Health and Retirement Survey documentation. Ann Arbor: University of Michigan. Available at: http://hrsonline.isr.umich.edu/sitedocs/wghtdoc.pdf (accessed February 2015).

Kalton, G. and M. Brick. 2000. "Weighting in Household Panel Surveys." In *Researching Social and Economic Change: the Uses of Household Panel Studies*, edited by D. Rose, 96–112. London: Routledge.

Kroh, M. 2013. *Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2012)*. DIW Data Documentation 71. Berlin: DIW.

LaRoche, S. 2007. *Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics*. Income Research Paper Series, 75F0002MIE. Ottawa: Statistics Canada.

Lynn, P. 2009. "Methods for Longitudinal Surveys." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 1–19. Chichester: Wiley.

Rabe-Hesketh, S. and A. Skrondal. 2006. "Multilevel Modelling of Complex Survey Data." *Journal of the Royal Statistical Society, Series A* 169: 805–827. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00426.x.

Sadig, H. 2014. *Unknown Eligibility Whilst Weighting for Non-Response: The Puzzle of Who has Died and Who is Still Alive?* ISER Working Paper 2014-35. Colchester: Institute for Social and Economic Research, University of Essex.

Scholes, S., R. Taylor, H. Cheshire, K. Cox, and C. Lessof. 2008. *Retirement, Health and Relationships of the Older Population in England: The 2004 English Longitudinal Study of Ageing*, NatCen Technical Report P2808. London: National Centre for Social Research.

SHARE. 2013. *Release Guide 1.1.1 Wave 4*. Munich: Munich Centre for the Economics of Ageing. Available at: http://www.share-project.org/fileadmin/pdf_documentation/SHARE_wave_4_release_guide_1.1.1.pdf (accessed February 2015).

Skinner, C.J. and D.J. Holmes. 2003. "Random Effects Models for Longitudinal Survey Data." In *Analysis of Survey Data*, edited by R. Chambers and C.J. Skinner, 205–219. Chichester: Wiley.

Smith, T.W. 2009. *A Revised Review of Methods to Estimate the Status of Cases with Unknown Eligibility*. Report of the Standard Definitions Committee for the American Association for Public Opinion Research. Available at: https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/FindingE.pdf (accessed October 2016).

Taylor, M.F., ed. 2010. *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices*. Colchester: University of Essex.

Valliant, R., J.A. Dever, and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Watson, N. 2012. *Longitudinal and Cross-Sectional Weighting Methodology for the HILDA Survey*. HILDA Technical Paper Series 2/12. Melbourne: Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Watson, N. and M. Summerfield. 2014. *Outcomes from Matching the HILDA Survey Sample to the Death Register*, HILDA Technical Paper Series 2/14. Melbourne: Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Weaver, B. 2011. *Combined 2009 and 2010 Technical Weighting Report for the Swiss Household Panel*. Lausanne: Swiss Foundation for Research in Social Sciences. Available at: http://forscenter.ch/en/our-surveys/swiss-household-panel/documentationfaq-2/methods/weighting/ (accessed February 2015).

Received March 2015
Revised August 2016
Accepted August 2016

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 1011–1014, http://dx.doi.org/10.1515/JOS-2016-0053

# Book Review

*Jessica M. E. Herzing*[1]

**Callegaro, Mario, Katja Lozar Manfreda, and Vasja Vehovar.** 2015.*Web Survey Methodology*. London, UK: SAGE Publications Ltd, ISBN 978-0-85702-860-0, 318 pp., £22.51.

The book *Web Survey Methodology* by Mario Callegaro, Katja Lozar Manfreda, and Vasja Vehovar introduces readers to the fundamental concepts of web surveys. It covers key concepts and findings in the literature on questionnaire design, sampling, recruitment, fielding, nonresponse, data preparation, post-survey adjustments, paradata, survey software, and cost issues. The book gives practical advice on conducting web surveys and points out critical aspects and new developments in the field of web survey methodology.

The book's approach differs from that of other books on web surveys, such as Tourangeau et al. (2013), for example; Callegaro et al. do not structure their book around the Total Survey Error (TSE) framework (Groves et al. 2011), but base the chapters on the web survey process. Therefore, the book *Web Survey Methodology* goes beyond the topics of error sources or design effects (also see Couper 2008) by including subjects such as field management, cost, and legal issues. Consequently, the book *Web Survey Methodology* is a valuable addition to the literature on web survey methodology and should satisfy readers who are interested in carrying out their own web survey.

The book contains eight chapters. The first chapter is a general introduction on the topic of web survey methodology. This is followed by three chapters on the key steps of the web survey process: prefielding, fielding, postfielding. The last part consists of three chapters on selected advanced topics and future developments in the field of web surveys. The book ends with a conclusion on how web surveys have changed and will continue to change the field of survey methodology.

Chapter 1 sets the scene for the entire book by providing definitions and typologies. Despite the fact that it is common practice to use the terms online survey, Internet survey and web survey interchangeably, the authors do a good job of clarifying and differentiating these terms from each other. For example, the concept of web survey is described here as HTML content with a unique web address which enables users to enter data on a web form. The generated content is then transferred to a researcher's server (p. 12). This is followed by an overview of advantages and limitations of the web survey method from a researcher's, practitioner's and do-it-yourself user's perspective. While web surveys have variable cost and measurement advantages, they are also limited by the absence of a sampling frame of email addresses or IP addresses (which makes contacting the general population demanding). The chapter concludes with a description of different types of web

[1] Collaborative Research Center "Political Economy of Reforms", University of Mannheim, Germany. Email: jessica.herzing@uni-mannheim.de

surveys, such as business surveys, online panels, website evaluations, and so on. In this context, the authors highlight the relationship between the different types and purposes of web surveys and the characteristics of the sample population. For instance, website evaluations measure the characteristics of the websites' visitors and the analysis focus is on website usage. Thus the sample consists of respondents who have visited the website, and this is the reason why the sample is not representative of the general population. The issue of sample composition becomes apparent in several chapters and sections throughout the book where differences between nonprobability and probability-based surveys are emphasized (see, for example Chapter 2).

Chapter 2, the most elaborate chapter, looks into the prefielding stage which covers issues related to survey mode choice, sampling, questionnaire design, technical implementation, nonresponse strategies, and general field management. In this chapter, the authors provide practical guidance to aid researchers in their decision in favor of or against collecting data in the web mode. The guiding questions at the beginning of the chapter are particularly helpful in this regard. Because a decision to conduct a web survey has consequences for the sampling procedure, the chapter also describes the availability of sampling frames and options for mixed-mode approaches. In many ways, therefore, this chapter draws on the book by Dillman et al. (2014). In this context, the authors take a very thoughtful and critical view of similarities and differences between nonprobability and probability-based sampling frames and their consequences for statistical inferences. The authors delineate the problem of noncoverage and nonresponse to web surveys and potential solutions. Furthermore, this chapter offers a general introduction to basic principles of questionnaire design and possible pitfalls in the technical implementation. While not reaching the level of detail on web-questionnaire design as achieved in Couper (2008), they cover all important design issues pertinent to modern web survey implementation. Various methods for questionnaire testing are described and guiding questions help practitioners considering decisions regarding data capturing (for example capturing paradata, which might be useful for field monitoring), data security, and privacy issues. Building on the current research of nonresponse theory, the authors specify influencing factors of nonresponse which are unique to web surveys. In addition, the authors give an introduction to related concepts, such as nonresponse rates, nonresponse bias, incentives, and invitations. Chapter 2 concludes with further guiding questions for practitioners which point to critical time points in the general management of the prefielding phase.

Compared with the chapter on prefielding, the chapter on the fielding of a web survey (Chapter 3) is more concise. The recruitment of respondents and the measurement process are briefly summarized. The chapter provides helpful advice on strategies for launching the survey. The authors suggest a soft launch which involves a small subset of respondents. This soft launch offers the opportunity to conduct first quality checks and detect technical problems. Furthermore, this chapter describes performance indicators and actions that can be taken in case of unexpected events during the survey fielding.

Chapter 4 looks into the postfielding period, which covers data preparation, preliminary results, data exporting and documentation. Aspects of data preparation such as defining the response status, data validations and data editing, imputations and weighting are discussed. The section on imputations contributes to the understanding that imputation strategies depend on the type of missing data and hence cannot be used in every case. In addition, the chapter discusses the coding of open answers, as well as short and long-term accessibility and solutions for the anonymization of the data.

The emergence of new technologies and thus the use of mobile devices, such as smartphones, challenge the field of survey methodology, in particular in the web survey context. Nowadays respondents are able to use multiple devices, such as computers, smartphones, and tablets, to fill out web questionnaires. Therefore, the authors propose strategies for handling respondents' use of multiple devices in Chapter 5, as the display of web questionnaires can change by browser. Furthermore, the type of device may differ for certain respondent groups and hence device effects can occur (for example, participation rates differ by device). Moreover, the authors discuss the purpose of online panels as a source of web survey respondents and data quality issues regarding both nonprobability and probability-based panels. This is a contribution few other books on web surveys make, yet it is very valuable in current debates concerning the quality of probability and nonprobability online panels (Goel et al. 2015 or Kennedy et al. 2016). The authors further describe what is available on the web survey software market, software characteristics, and their effect during the web survey process. They also point to stages in the web survey process which are not yet or scarcely supported by software. A specific section on finding the right software tool is particularly useful to future practitioners.

Chapter 6 discusses the role of web surveys in the context of general survey methodology. It adds an extended definition of the term web survey mode, as well as its role in the context of the TSE framework (Groves et al. 2011). Beyond that, the authors explain how the TSE framework can be enriched by the concept of survey data quality (Lyberg 2012). In this regard, the authors briefly address the important question of how the survey mode relates to survey data quality. This chapter offers further insights for survey practitioners into variations of the interactive fieldwork design and the project management framework. Once again, questions guide the reader through important legal and ethical issues concerning web surveys.

Chapter 7 covers new trends in technological development, web survey software, methodology, and broader business and societal issues. Among other trends, the authors expect an increasing amount of do-it-yourself research with web surveys, due to the advancement of web survey software. The authors' thoughtful elaboration on the rise and decline of mixed-mode data collection and the integration of survey data with big data is particularly interesting.

In Chapter 8 the authors elaborate on how web surveys have changed and will continue to change the field of survey methodology. They conclude that the discussion on sampling and recruiting respondents in web surveys will persist. Furthermore, they point out topics where further research is needed, such as the development of standards for nonprobability web surveys.

In summary, the chapters provide a rich collection of the current state-of-the-art literature on web survey methodology. In addition to citing the relevant published literature, the authors provide further information at the accompanied WebSM website (http://www.websm.org). However, a "further readings" section at the end of each chapter would have been desirable, as it would have facilitated finding literature on specific topics.

While reading the book *Web Survey Methodology*, I got a strong sense that it was written with students and practitioners in mind who have different levels of knowledge and come from different fields, such as survey or psychological researchers, official statistics, and market or customer-satisfaction researchers. Some chapters are pitched at an introductory level and are useful for practitioners or students who are new to the field, while other chapters target practitioners who want to update their knowledge on web surveys. Even though the book does not offer a recipe on how to conduct web surveys, the authors succeed in instructing readers on the different decisions one has to take throughout the web survey process. Besides offering practical guidance, this book is an excellent, comprehensive, and rich introduction to the field of web survey methodology.

In conclusion, I can highly recommend this book to students, practitioners, and researchers who are planning their own web surveys and wishing to gain theoretical and practical knowledge on web survey methodology.

## References

Couper, M.P. 2008. *Designing Effective Web Surveys*. New York: Cambridge University Press.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons.

Goel, S., A. Obeng, and D. Rothschild. 2015. "Non-Representative Surveys: Fast, Cheap, and Mostly Accurate." Available at: http://researchdmr.com/FastCheapAccurate (accessed July 2016).

Groves, R.M., F.J. Fowler, Jr., M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2011. *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.

Kennedy, C., A. Mercer, S. Keeter, N. Hatley, K. McGeeney, and A. Gimenez. 2016. "Evaluating Online Nonprobability Surveys." Pew Research Center. Available at: http://www.pewresearch.org/2016/05/02/evaluating-online-nonprobability-surveys/ (accessed September 2016).

Lyberg, L.E. 2012. "Survey Quality." *Survey Methodology* 38: 107–130.

Tourangeau, R., F. Conrad, and M.P. Couper. 2013. *The Science of Web Surveys*. London: Oxford University Press.

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 1015–1017, http://dx.doi.org/10.1515/JOS-2016-0054

# Book Review

*Kristen Olson*[1]

**Uwe Engel, Ben Jann, Peter Lynn, Annette Scherpenzeel, and Patrick Sturgis, eds.** *Improving Survey Methods: Lessons from Recent Research*. NY: Routledge, 2015, ISBN 978-0-415-81762-2, 430pps, $165.

*Improving Survey Methods: Lessons from Recent Research* is a compilation of research by survey methodological leaders across Europe. The book is organized into eight sections – modes, interviewers, sensitive questions, web surveys, access panels, nonsurvey data collection, nonresponse, and missing data. Each section starts with a brief overview chapter followed by three or four (generally) empirical chapters. The chapters themselves vary in approach, with some being simple literature reviews, others reporting the results of a simple 2 × 2 experiment, and still others conducting extensive observational analyses. This volume is a clear indication that survey methodological research is strong in Europe.

The first section, "Survey Modes and Response Effects," contains two chapters that make use of the German Priority Programme on Survey Methodology (PPSM) panel. First, there is an experimental comparison of scale labels and 'don't know' options in telephone and web modes by De Leeuw and Hox (Chapter 3), showing differences across modes in the effects of full versus endpoint labeling and in the use of 'don't know' options. Engel and Koster (Chapter 4) then look at measurement quality within one mode – telephone – as related to the effects of a variety of question features, such as response-option order, question wording, number of scale points, and variation by interviewer's ratings of the respondent's attention to the task. In the third mode chapter, Busse and Fuchs (Chapter 5) take a high-level look at the error structure of mobile phone surveys, providing an overview of mobile surveys in both Europe and the US with some empirical data on errors of nonobservation in these types of surveys. The empirical results from these chapters are interesting, although the section feels connected more by the "response effects" part of the section title than "survey modes."

The second section, "Interviewers and Survey Error," contains two chapters – one by Turner et al. (Chapter 7) examining interviewer variance effects in the National Travel Survey of Great Britain, and one by Menold et al. (Chapter 8) on a case-control study for a method to detect interviewer falsification (summarizing prior extended reports and papers). Both of these brief chapters were interesting and informative, and contribute to our understanding of interviewers and their effects on survey measurement. Less attention was given in this section to the role of interviewers in other survey error sources, although studying error sources other than measurement is challenging.

[1] University of Nebraska-Lincoln, Department of Sociology, 703 Oldfather Hall, Lincoln, NE 68506 U.S.A. Email: kolson5@unl.edu

Section III focuses on “Asking Sensitive Questions.” This section takes a different approach to sensitive questions than the traditional method of examining modes or question wording (for example, following Tourangeau and Yan 2007), with two chapters that focus on randomized response and similar techniques (Diekmann and Hoglinger, Chapter 10; Krumpal et al., Chapter 11) and one using vignettes (Auspurg et al., Chapter 12). The chapter by Krumpal et al. (Chapter 11) provides a particularly useful overview of three groups of these methods, worth assigning in a graduate class.

The three strong contributions of Section IV turn to web surveys. The first chapter in this section by Bethlehem (Chapter 14) provides a succinct overview of nonobservation errors – coverage, sampling, and nonresponse – to consider when conducting web surveys. Chapter 15 by Vehovar, Petrovcic, and Slavec is a particularly novel contribution on the information-communication technology software tools for questionnaire drafting and development. Although the design of the study in this chapter is somewhat limited (students in a graduate class; two single-topic questionnaires), this chapter provides important insights both into the time required for questionnaire development and into how computerized tools may or may not assist with this development. The final chapter in this section by Braun, Behr, Kaczmirek, and Bandilla (Chapter 16) demonstrates the strength of web surveys in cross-national surveys for using open-ended probes to understand answers to closed-ended questions. Overall, the web survey section is one to read.

The next section examines the set-up, participation in, and estimation from probability-based panels of willing respondents, called “access panels” in the volume – the German PPSM Access Panel (Engel, Chapter 18), the Longitudinal Internet Studies for the Social Sciences panel (LISS; Scherpenzeel, Chapter 19), and the Microcensus access panel for the German Federal Statistical Office (Rendtel and Amarov, Chapter 20). Chapter 21 focuses on estimation and weighting for a variety of estimates calculated from access panels (Enderle and Munnich, Chapter 21). With a growing number of access panels being developed in the US (e.g., NORC’s AmeriSpeak panel (https://www.amerispeak.org/), Pew Research Center’s American Trends Panel (http://www.pewresearch.org/methodology/u-s-survey-research/american- trends-panel/)) and elsewhere, the insights from these chapters are practical and informative for many researchers who are setting up their own panels or who use a probability-based access panel.

Section VI provides overview chapters about the future of survey research, with data linkage to administrative records (Schnell, Chapter 23), data collected by observers, environmental, or biological measurements (Schnell, Chapter 24), and data about the survey data collection process, also known as paradata (Kreuter, Chapter 25). These chapters do not provide new empirical findings, but can be a quick read for a graduate student or researcher wanting a description and initial set of references for research in these areas.

Nonresponse is the focus of Section VII. Lynn (Chapter 27) argues that the wealth of data collected in earlier waves of longitudinal surveys should be used to specifically tailor recruitment protocol features in later waves of longitudinal surveys to increase response rates and potentially reduce nonresponse bias, providing four brief examples from studies in the United Kingdom. Göritz (Chapter 28) provides a handy summary of existing meta-analyses of incentive experiments, with an extended discussion of the effectiveness of different types of incentives in web surveys. A particularly interesting finding in this review is that offering participants results from the study had no effect or *decreased* response rates relative to no incentive. Stoop's (Chapter 29) discussion of the European Social Survey (ESS) is a nice summary of the extensive work done on the ESS for researchers who do not have time to read the more extensive volumes (e.g., Stoop et al. 2010). Overall, these chapters, and especially the contributions by Lynn and Göritz, provide interesting insights into different factors related to ways to increase participation rates in different types of surveys.

The final section of the book focuses on missing data, with a particular focus on multiple imputation. The applications of multiple imputation to split questionnaire designs and matrix sampling (Bahrami et al., Chapter 31) and zero-inflated and multilevel count data (Kleinke and Reinecke, Chapter 32) are practical solutions to common data issues encountered by analysts. De Jong and Spiess (Chapter 33) provide a new semiparametric imputation model in the most statistically technical chapter of the volume. These chapters will appeal to the more statistically oriented readers, although their findings and approaches could be used by many scholars as they become implemented in software packages.

As with all edited volumes, different chapters in this book have different audiences. Some are new empirical findings, while others are overviews of existing literature. The preface (pp. xi-xii) describes the volume as arising from a series of conferences. It is a valuable resource for researchers who were not able to attend these meetings, and for understanding some of the innovations occurring in methods in Germany and beyond.

## References

Stoop, I., J. Billiet, A. Koch, and R. Fitzgerald. 2010. *Improving Survey Response: Lessons Learned from the European Social Survey*. Chichester, UK: Wiley.

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.

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 1019–1024, http://dx.doi.org/10.1515/JOS-2016-0055

# 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, 2015–September 30, 2016. An asterisk indicates that the referee served more than once during the period.

Abbott, Owen, Office for National Statistics, Titchfield, UK
Abraham, Katharine, JPSM, College Park, MD, U.S.A.
Amaya, Ashley, RTI International, Washington, Washington DC, U.S.A.
Andersson, Per Gösta, Stockholm University, Stockholm, Sweden
Ashmead, Robert, US Census Bureau, Washington, Washington DC, U.S.A.
Bach*, Ruben, The Research Institute of the Federal Employment Agency, Nuremberg, Germany
Baffour, Bernard, The University of Queensland, Brisbane, Australia
Bautista, René, NORC at the University of Chicago, Chicago, Illinois, U.S.A.
Bavdaž* , Mojca, University of Ljubljana, Ljubljana, Slovenia
Benedetti*, Roberto, University of Chieti Pescara, Pescara, Italy
Beresovsky*, Vladislav, National Center for Health Statistics, Hyattsville, Maryland, U.S.A.
Bergsma, Wischer, London School of Economics and Political Science, London, UK
Berzofsky, Marcus, RTI International, Research Triangle Park, North Carolina, U.S.A.
Bianchi, Annamaria, Universita Degli Studi di Bergamo, Bergamo, Italy
Bijak*, Jakub, University of Southampton, Southampton, UK
Bivand, Roger, Norwegian School of Economics, Bergen, Norway
Bocci, Cynthia, Statistics Canada, Ottawa, Ontario, Canada
Bollineni-Balabay*, Oksana, Maastricht University, Maastricht, The Netherlands
Bonnéry, Daniel, University of Maryland at College Park, Maryland, U.S.A.
Bregar, Lea, University of Ljubljana, Ljubljana, Slovenia
Brick, Michael, Westat, Rockville, Maryland, U.S.A.
Brion, Philippe, INSEE, Paris, France
Bryant*, John, Statistics New Zealand, Christchurch, New Zealand
Buelens*, Bart, Statistics Netherlands, Heerlen, The Netherlands
Bueno, Edgar, Stockholm University, Stockholm, Sweden
Buono*, Dario, Eurostat, Mamer, Luxembourg
Burger, Joep, Statistics Netherlands, Heerlen, The Netherlands
Burton, Jonathan, University of Essex, Colchester, UK
Callegaro, Mario, Google, London, UK
Caldwell, Carol, US Census Bureau, Washington, Washington DC, U.S.A.
Campbell, Paul, Australian Bureau of Statistics, Sidney, Australia
Cardot*, Hervé, University de Bourgogne, Dijon, France
Caspar, Rachel, RTI International, Research Triangle Park, North Carolina, U.S.A.

Cernat, Alex, University of Manchester, Manchester, UK
Charest, Anne-Sophie, Université Laval, Quebec, Canada
Chauvet*, Guillaume, ENSAI, Bruz Cedex, France
Chesnut, John, Owing, Maryland, U.S.A.
Chessa, Antonio, Centraal Bureau voor de Statistiek, Den Haag, The Netherlands
Chevarley, Frances, Agency for Healthcare Research and Quality, Rockville, MD, U.S.A.
Citro*, Constance, Committee on National Statistics, Washington, Washington DC, U.S.A.
Cornelis*, Èric, Université de Namur, Namur, Belgium
Couper, Mick, University of Michigan, Ann Arbor, Michigan, U.S.A.
Crequer, John, Statistics New Zealand, Christchurch, New Zealand
Cruyff, Maarten, University Utrecht, Utrecht, The Netherlands
Czajka, John, Mathematica Policy Research Inc., Washington, Washington DC, U.S.A.
Dalla Valle, Luciana, Plymouth University, Plymouth, UK
Dasylva, Abel, Statistics Canada, Ottawa, Ontario, Canada
Davern, Michael, NORC at the University of Chicago, Chicago, IL, U.S.A.
Davis, Karen, Agency for Healthcare Research, Rockville, Maryland, U.S.A.
Defays, Daniel, University of Liége, Liége, Belgium
Del Barrio Castro*, Tomas, University of Balearic Islands, Palma, Mallorca, Spain
Demnati, Abdellatif, Ottawa, Ontario, Canada
Dever, Jill, RTI International, Washington, Washington DC, U.S.A.
Diewert, Erwin, University of British Colombia, Vancouver, Canada
Dipko, Sarah, Westat, Rockville, MD, U.S.A.
Di Consiglio, Loredana, ISTAT, Rome, Italy
Di Zio, Marco, ISTAT, Rome, Italy
Djerf*, Kari, Statistics Finland, Helsinki, Finland
Durrant*, Gabrielle, University of Southampton, Southampton, UK
Dykema, Jennifer, University of Wisconsin-Madison, Madison, Wisconsin, U.S.A.
Earp, Morgan, Bureau of Labor Statistics, Washington, Washington DC, U.S.A.
Edwards*, Brad, Westat, Rockville, MD, U.S.A.
Elezovic, Suad, Statistics Sweden, Stockholm, Sweden
Elliott, Duncan, Office for National Statistics, Newport, UK
English*, Ned, NORC at the University of Chicago, Chicago, Illinois, U.S.A.
Felderer*, Barbara, University of Mannheim, Mannheim, Germany
Fowler, Floyd Jackson Jr., University of Massachusetts, Boston, MA, U.S.A.
Franco, Carolina, US Census Bureau, Washington, Washington DC, U.S.A.
Gargiulo, Floriana, Universite de Namur, Namur, Belgium
Gerland, Patrick, United Nations, New York, New York, U.S.A.
Giusti, Caterina, University of Pisa, Pisa, Italy
Goebel, Jan, Deutsches Institut fur Wirtschaftsforschung eV SOEP, Berlin, Germany
Gonzalez, Jeffrey, U.S. Bureau of Labor Statistics, Washington, Washington DC, U.S.A.
Graham, Patrick, University of Otago, Christchurch, New Zealand
Greenway-McGrevy*, Ryan, University of Auckland, Auckland, New Zealand
Guarnera, Ugo, ISTAT, Rome, Italy
Gweon, Hyukjun, University of Waterloo, Waterloo, Ontario, Canada

Lenis, David, Johns Hopkins University, Baltimore, Maryland, U.S.A.
Lesage, Éric, INSEE, Paris, France
Lewis*, Daniel, Office for National Statistics, Newport, UK
Lewis, Taylor, U.S. Office of Personnel Management, Washington DC, U.S
Liberts, Mārtinš, Central Statistical Bureau of Latvia, Riga, Latvia
Lin, Yan-Xia, University of Wollongong, Wollongong, Australia
Lipps*, Oliver, FORS, Lausanne, Switzerland
Liu, Benmei, National Cancer Institute, Rockville, Maryland, U.S.A.
Liu, Mingnan, Survey Monkey, Palo Alto, California, U.S.A.
Loft, John, RTI International, Chicago, Illinois, U.S.A.
Loong, Bronwyn, Australian National University, Canberra, Australia
Looseveldt, Geert, KU Leuven, Leuven, Belgium
Lozar Manfreda, Katja, University of Ljubljana, Ljubljana, Slovenia
Lugtig, Peter, Utrecht University, Utrecht, The Netherlands
Luiten, Annamieke, Statistics Netherlands, Heerlen, The Netherlands
Lumley*, Thomas, University of Auckland, Auckland, New Zealand
Lundquist, Peter, Statistics Sweden, Stockholm, Sweden
Lynn*, Peter, ISER, University of Essex, Essex, UK
Ma, Yinyi, Rotterdam, The Netherlands
Magnussen, Steen, Canadian Forest Service, Victoria BC, Canada
Maples, Jerry, U.S. Census Bureau, Washington, Washington DC, U.S.A.
Maravall*, Agustin, Madrid, Spain
Marti Sempere, Monica, University of Alicante, Alicante, Spain
Mathew, Thomas, University of Maryland, Baltimore, Maryland, U.S.A.
McElroy, Tucker, U.S. Census Bureau, Washington, Washington DC, U.S.A.
Meeden, Glen, University of Minnesota, Minneapolis, MN, U.S.A.
Meinfelder, Florian, University of Bamberg, Bamberg, Germany
Meng, Xiao-Li, Harvard University, Cambridge, MA, U.S.A.
Mercer, Andrew, Pew Research Center, Washington, Washington DC, U.S.A.
Miller, Peter, US Census Bureau, Washington, Washington DC, U.S.A.
Miranda, Javier, U.S. Census Bureau, Suitland, Maryland, U.S.A.
Mirel, Lisa, Centers for Medicare and Medicaid Services OEDA, Baltimore, Maryland, U.S.A.
Mneimneh, Zeina, University of Michigan, Ann Arbor, Michigan, U.S.A.
Mohler, Peter, University of Mannheim, Leinsweiler, Germany
Moors, Guy, Tilburg University, Tilburg, The Netherlands
Morren*, Meike, VU University Amsterdam, Amsterdam, The Netherlands
Moura da Silva, Fernando, UFRJ, Rio de Janeiro, Brazil
Muñoz, Juan, University of Granada, Granada, Spain
Muralidhar, Krishnamurty, University of Oklahoma, Norman, Oklahoma, U.S.A.
Namazi-Rad, Mohammad-Reza, University of Wollongong, Wollongong, Australia
Nayak, Tapan, George Washington University, Washington, Washington DC, U.S.A.
Nolte, Sandra, Lancaster University, Bailrigg, Lancaster, UK
Norberg, Anders, Statistics Sweden, Stockholm, Sweden
Oberski, Daniel, Tilburg University, Tilburg, The Netherlands

Oganian*, Anna, National Center for Health Statistics, Hyattsville, Maryland, U.S.A.
Oguz-Alper, Melike, Statistics Norway, Oslo, Norway
O'Muircheartaigh, Colm A., NORC at the University of Chicago, Chicago, IL, U.S.A.
Piva, Thais, Universidade Federal de Minas, Pampulha, Belo Horizonte, Brazil
Pascale, Joanne, U.S Census Bureau, Washington, Washington DC, U.S.A.
Pedlow, Steven, NORC at the University of Chicago, Chicago, IL, U.S.A.
Persson*, Andreas, Statistics Sweden, Örebro, Sweden
Peytchev*, Andy, RTI International, Research Triangle Park, North Carolina, U.S.A.
Peytcheva, Emilia, RTI International, Research Triangle Park, North Carolina, U.S.A.
Pinheiro, Eliane, University of São Paulo, São Paulo, Brazil
Plewis, Ian, University of Manchester, Manchester, UK
Puts, Marco, Statistics Netherlands, Heerlen, The Netherlands
Raftery*, Adrian, University of Washington, Seattle, WA, U.S.A.
Ranalli*, Giovanna, University of Perugia, Perugia, Italy
Raymer, James, Australian National University, Canberra, Australia
Redline, Cleo, NCES, Washington, Washington DC, U.S.A.
Reedman, Laurie, Statistics Canada, Ottawa, Ontario, Canada
Reiter, Jerome, Duke University, Durham, North Carolina, U.S.A.
Robinson, Edwin, Bureau of Labor Statistics, Washington, Washington DC, U.S.A.
Rohe*, Karl, University of Wisconsin-Madison, Madison, WI, U.S.A.
Ronning, Gerd, University of Tübingen, Tübingen, Germany
Russ*, Daniell, National Institutes of Health, Bethesda, Maryland, U.S.A.
Rust, Keith, Westat, Rockville, MD, U.S.A.
Saei*, Ayoub, University of Southampton, Southampton, UK
Sarathy*, Rathindra, Oklahoma State University, Stillwater, Oklahoma, U.S.A.
Scheffer, Fredrik, Statistics Sweden, Stockholm, Sweden
Schierholz*, Malte, The Institute for Employment Research, Nuremberg, Germany
Schmid*, Timo, Freie Universität Berlin, Berlin, Germany
Scholtus, Sander, Statistics Netherlands, The Hague, The Netherlands
Schonlau*, Matthias, University of Waterloo, Waterloo, Ontario, Canada
Schouten*, Berry, Statistics Netherlands, The Hague, The Netherlands
Sempere Marti, Monica, Universidad de Alicante, Alicante, Spain
Shook-Sa*, Bonnie, RTI International, Research Triangle Park, North Carolina, U.S.A.
Skalland, Benjamin, NORC at the University of Chicago, Chicago, Illinois, U.S.A.
Smith, Tom, NORC at the University of Chicago, Chicago, Illinois, U.S.A.
Smith, Peter W.F., University of Southampton, Southampton, UK
Spencer, Bruce, Northwestern University, Evanston, Illinois, U.S.A.
Steel*, Philip, US Census Bureau, Washington, Washington DC, U.S.A.
Steinhauer, Hans Walter, Leibniz-Institutet für Bildungsverläufe, Bamberg, Germany
Stoop, Ineke, SCP Sociaal en Cultureel Plansbureay, The Hague, The Netherlands
Suzer-Gurtekin*, Tuba, University of Michigan, Ann Arbor, Michigan, U.S.A.
Tancredi, Andrea, University of Rome, Rome, Italy
Thalji, Lisa, RTI International, Research Triangle Park, North Carolina, U.S.A.
Thibaudeau, Yves, U.S. Census Bureau, CSRM, Washington, Washington DC, U.S.A.
Thomas, Troy, University of Guyana, Georgetown, Guyana

Thorsdottir, Fanney, University of Iceland, Reykjavik, Iceland
Tijdens*, Kea, University of Amsterdam, Amsterdam, The Netherlands
Tille, Yves, University of Neuchâtel, Neuchâtel, Switzerland
Toth, Daniell, U.S. Bureau of Labor Statistics, Washington, Washington DC, U.S.A.
Tourangeau, Roger, Westat, Rockville, Maryland, U.S.A.
Trappman, Mark, Institute for Employment Research, Nuremberg, Germany
Trimbur, Thomas, US Census Bureau, Washington, Washington DC, U.S.A.
Tucker*, Clyde, American Institute of Research, Washington, DC, U.S.A.
Tuolemon, Laurent, Institute National D'Études, Paris, France
Tuoto*, Tiziana, ISTAT, Rome, Italy
Tzavidis, Nikos, University of Southampton, Southampton, UK
Van Delden*, Arnout, Statistics Netherlands, The Hague, The Netherlands
Van den Brakel, Jan, Statistics Netherlands, Heerlen, The Netherlands
Vandenplas*, Caroline, UNIL-FORS, Lausanne, Switzerland
Van de Pol, Frank, Central Bureau of Statistics, Heerlen, The Netherlands
Van Kints, Marcel, Australian Bureau of Statistics, Belconnen, Australia
Vermount, Jeroen, Tilburg University, Tilburg, The Netherlands
Vicente, Paula, ISCTE- Lisbon University, Lisbon, Portugal
Vilhuber, Lars, Cornell University, Ithaca, New York, U.S.A.
Villund*, Ole, Statistics Norway, Oslo, Norway
Virgile, Matthew, US Census Bureau, Suitland, Maryland, U.S.A.
Wagner, James, University of Michigan, Ann Arbor, Michigan, U.S.A.
Walejko*, Gina, US Census Bureau, Suitland, Maryland, U.S.A.
Watson, Nicole, University of Melbourne, Melbourne, Australia
Welsh, Alan, Australian National University, Canberra, Australia
West*, Brady, University of Michigan, Ann Arbor, Michigan, U.S.A.
Winkler, William, U.S Census Bureau, Washington, Washington DC, U.S.A.
Wit, Ernst, Rijksuniversiteit Groningen, Groningen, The Netherlands
Xie*, Yingfu, Statistics Sweden, Stockholm, Sweden
Yan, Ting, Westat, Rockville, Maryland, U.S.A.
Yansaneh, Ibrahim, United Nations, New York, New York, U.S.A.
Yao, Jing, University of Glasgow, United Kingdom of Great Britain and Northern Ireland
You, Yong, Statistics Canada, Ottawa, Ontario, Canada
Yuskavage*, Robert, U.S. Bureau of Economic Analysis, Springfield, Virginia, U.S.A.
Zabala, Felipa, Statistics New Zealand, Wellington, New Zealand
Zangeneh, Sahar, University of Michigan, Ann Arbor, Michigan, U.S.A.
Zwick, Markus, European Commission, Luxembourg, Germany

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 1025–1028, http://dx.doi.org/10.1515/JOS-2015-0056

# Index to Volume 32, 2016

Contents of Volume 32, Numbers 1–4



## Author Index

## Book Reviews



Printed in December 2016