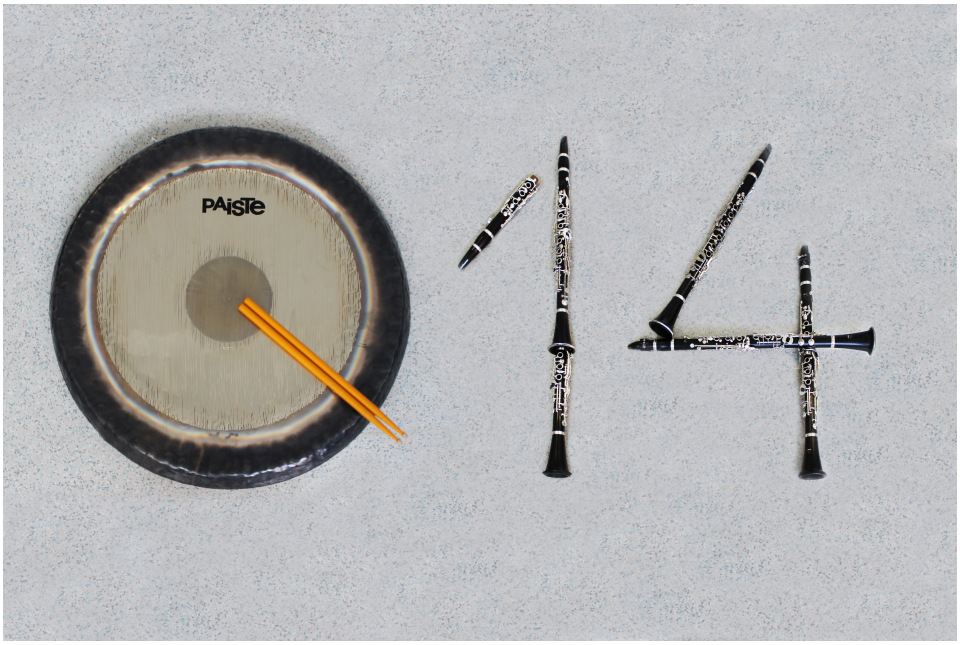
Austrian Journal of Statistics

AUSTRIAN STATISTICAL SOCIETY

Volume 44, Number 2, 2015

Special Issue

Q2014 , Vienna, Austria



Österreichische Zeitschrift für Statistik

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Austrian Journal of Statistics; Information and Instructions

GENERAL NOTES

The Austrian Journal of Statistics is an open-access journal with a long history and is published approximately quarterly by the Austrian Statistical Society. Its general objective is to promote and extend the use of statistical methods in all kind of theoretical and applied disciplines. Special emphasis is on methods and results in official statistics.

Original papers and review articles in English will be published in the Austrian Journal of Statistics if judged consistently with these general aims. All papers will be refereed. Special topics sections will appear from time to time. Each section will have as a theme a specialized area of statistical application, theory, or methodology. Technical notes or problems for considerations under Shorter Communications are also invited. A special section is reserved for book reviews.

All published manuscripts are available at

http://www.ajs.or.at

(old editions can be found at http://www.stat.tugraz.at/AJS/Editions.html)

Members of the Austrian Statistical Society receive a copy of the Journal free of charge. To apply for a membership, see the website of the Society. Articles will also be made available through the web.

PEER REVIEW PROCESS

All contributions will be anonymously refereed which is also for the authors in order to getting positive feedback and constructive suggestions from other qualified people. Editor and referees must trust that the contribution has not been submitted for publication at the same time at another place. It is fair that the submitting author notifies if an earlier version has already been submitted somewhere before. Manuscripts stay with the publisher and referees. The refereeing and publishing in the Austrian Journal of Statistics is free of charge. The publisher, the Austrian Statistical Society requires a grant of copyright from authors in order to effectively publish and distribute this journal worldwide.

OPEN ACCESS POLICY

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge.

ONLINE SUBMISSIONS

Already have a Username/Password for Austrian Journal of Statistics?
Go to http://www.ajs.or.at/index.php/ajs/login

Need a Username/Password?
Go to http://www.ajs.or.at/index.php/ajs/user/register

Registration and login are required to submit items and to check the status of current submissions.

AUTHOR GUIDELINES

The original LaTeX-file guidelinesAJS.zip (available online) should be used as a template for the setting up of a text to be submitted in computer readable form. Other formats are only accepted rarely.

SUBMISSION PREPARATION CHECKLIST

- The submission has not been previously published, nor is it before another journal for consideration (or an explanation has been provided in Comments to the Editor).
- The submission file is preferable in LaTeXfile format provided by the journal.
- All illustrations, figures, and tables are placed within the text at the appropriate points, rather than at the end.
- The text adheres to the stylistic and bibliographic requirements outlined in the Author Guidelines, which is found in About the Journal.

COPYRIGHT NOTICE

Manuscripts should be unpublished and not be under consideration for publication elsewhere. By submitting an article, the author(s) certify that the article is their original work, that they have the right to submit the article for publication, and that they can grant the above license.

Austrian Journal of Statistics

Volume 44, Number 2, 2015

Special Guest Editors: Gerhard NACHTMANN, Andreas QUATEMBER

Editor-in-chief: Matthias TEMPL

http://www.ajs.or.at

Published by the AUSTRIAN STATISTICAL SOCIETY

http://www.osg.or.at

Österreichische Zeitschrift für Statistik

Jahrgang 44, Heft 2, 2015

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Impressum

Editor:	Matthias Templ, Statistics Austria & Vienna University of Technology
Editorial Board:	Peter Filzmoser, Vienna University of Technology Herwig Friedl, TU Graz Bernd Genser, University of Konstanz Peter Hackl, Vienna University of Economics, Austria Wolfgang Huf, Medical University of Vienna, Center for Medical Physics and Biomedical Engineering Alexander Kowarik, Statistics Austria, Austria Johannes Ledolter, Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien & Management Sciences, University of Iowa Werner Mueller, Johannes Kepler University Linz, Austria Josef Richter, University of Innsbruck Milan Stehlik, Department of Applied Statistics, Johannes Kepler University, Linz, Austria Wolfgang Trutschnig, Department for Mathematics, University of Salzburg Regina Tüchler, Austrian Federal Economic Chamber, Austria Helga Wagner, Johannes Kepler University Walter Zwirner, University of Calgary, Canada
Book Reviews:	Ernst Stadlober, Graz University of Technology
Printed by	Statistics Austria, A-1110 Vienna

Published approximately quarterly by the Austrian Statistical Society, C/o Statistik Austria
Guglgasse 13, A–1110 Wien

Contents

Editorial

From the 2nd to the 5th of June 2014 more than 450 experts from national statistical institutes, universities and national or international organisations participated at the 7th European Conference on Quality in Official Statistics (Q2014) in Vienna. The conference, organized by EUROSTAT and STATISTICS AUSTRIA and held in the beautiful venue of Schönbrunn Castle, was partitioned into 39 parallel and three plenary sessions with a total of 170 given talks. The scientific conference programme covered a broad spectrum of topics. This is also reflected in the present special issue of the Austrian Journal of Statistics (AJS) on Q2014, for which contributors from different areas could be won as authors.

In the opening paper "The ESCB quality framework for European statistics", Aurel Schubert and Cathrine Ahsbahs from the European Central Bank (ECB) aim to provide a highly topical discussion on the European System of Central Banks (ESCB). This talk was part of the special session on "Quality Assurance Measures in the European System of Central Banks". In the first part, the article provides a comparison to other existing frameworks. In the second, it describes the main processes to guarantee high quality statistics by the ESCB.

This paper is followed by a paper from the Session on "Quality of Data Collection". The authors of "Measuring Nonresponse Bias in a Cross-Country Enterprise Survey", Katarzyna Bańkowska, Małgorzata Osiewicz, Sébastien Pérez-Duarte (again form the ECB), try to describe the effect of "missing values" on the European Commission and European Central Bank Survey on the Access to Finance and Enterprises (SAFE). Their argumentation uses representativity indicators, which were investigated in the "Representativity Indicators of Survey Quality" project (RISQ). The aim of the article is to measure the quality oft the SAFE at different fieldwork stages, across different survey waves as well as across different countries.

The next two papers presented in the special session "Big Data" make use of web-scraping technologies as an additional data source for official statistics.

Giulio Barcaroli, Alessandra Nurra, Sergio Salamone, Monica Scannapieco, Donato Summa from Istituto Nazionale di Statistica (Istat) describe together witch Marco Scarnò from Cineca how Internet as Data Source is used in the Istat Survey on Information and Communication Technologies (ICT) in Enterprises. They scraped websites of enterprises responding to the survey of year 2013 and processed the acquired texts (text mining) in order to try to reproduce the same information collected via questionnaire.

Maciej Beręsewicz uses Internet as Data Source as well to assess its representativeness for the real estate market in Poland. Therefore he developed a special R program for automated data collection (web spider), which is available at GitHub. He suggests that the

existing definitions and methodology, which are valid for existing statistical data sources, should be adopted or revised to deal with new data sources.

The next paper written by Jörg Drechsler, Hans Kiesl and Matthias Speidel comes from the Q-Session on "Non-sampling Errors". As the titel "MI Double Feature: Multiple Imputation to Address Nonresponse and Rounding Errors in Income Questions" indicates, this paper aims at two quality problems occuring together when asking for sensitive survey variables such as income. The Multiple Imputation approach is used to address both problems at the same time.

The concluding paper from Eoin MacCuirc, titled "You Don't Teach, Students Learn: A Report on a Project on Statistical Literacy in Ireland", is based on a talk that was part of the Q-Session on "Statistical Literacy". It is a report on a remarkably responsible Education Programme that was launched in 2007 by the Central Statistics Office of Ireland with the aim of improving statistical literacy and effective use of statistics. Here, the author gives an overview of the experiences from the project to date.

For both of us, it was an honour and a pleasure to co-operate as guest editors of this special issue "Q2014" of the Austrian Journal of Statistics. Thanks to all the people contributing to this issue – authors as well as reviewers and, in particular, to the editor-in-chief Matthias Templ.

Gerhard Nachtmann, Andreas Quatember
(Guest Editors)

BOKU – University of Natural Resources
and Life Sciences, Vienna &
STATISTICS AUSTRIA
Guglgasse 13
1110 Vienna

Johannes Kepler University Linz
Department of Applied Statistics
Altenberger Straße 69
4040 Linz

Vienna/Linz, April 2015

Austrian Journal of Statistics
April 2015, Volume 44, 3–11.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i2.56

The ESCB Quality Framework for European Statistics

Aurel Schubert
European Central Bank

Catherine Ahsbahs
European Central Bank

Abstract

The aim of the paper is to provide a general presentation of the ESCB statistics quality framework. The first part of the paper is dedicated to the "Public commitment on European statistics by the ESCB" and presents some of the similarities and differences of the public commitment relative to the main existing frameworks (e.g. the United Nations Fundamental Principles of Official Statistics and the IMF Data Quality Assessment Framework). It also provides some information on the process followed by the ESCB to converge with the European Statistics Code of Practice. In the second part, the paper presents the main quality assurance procedures put in place by the ESCB and applied to the whole statistical production chain (including the supporting IT infrastructure).

Keywords: ESCB statistics, quality principles, quality assurance procedures, monitoring.

1. Introduction

Credible statistics lie at the heart of the European Central Bank's monetary policy-making. This is the reason why the development, collection, compilation and dissemination of statistics designed to support the conduct of monetary policy and other tasks of the European System of Central Banks (ESCB)[1] is one of the core functions of the ESCB. For the ESCB, adhering to high quality standards is a key factor in maintaining public trust in its European statistics, upon which policy decisions are based. Therefore, since the start of Economic and Monetary Union the ESCB has emphasised key aspects of statistical quality, such as relevance, accuracy, reliability, timeliness, consistency, cost-effectiveness, non-excessive burden on reporting agents and statistical confidentiality.

To develop and implement a quality framework tailor-made for the objectives of the ESCB statistical function, the first step was to undertake a stock-taking exercise of the existing frameworks that have been developed by national central banks (NCBs), national statistical institutes (NSIs) and international organisations. However, due to differences in the institutional environments of these organisations, the models developed differ somewhat with regard to their stakeholders, definitions of quality and scope.

[1]The ESCB comprises the European Central Bank (ECB) and the national central banks of all EU Member States whether they have adopted the euro or not.

Nevertheless, a number of these frameworks have been assessed in order to determine whether they contain elements that could serve the ESCB's purposes. As a result, the ESCB has developed the "Public commitment on European statistics by the ESCB", which is based on selected components of the different frameworks that have been carved out in order to fit the ESCB's institutional environment and operational features.

The aim of the paper is to provide a general presentation of the ESCB statistics quality framework. The first section is dedicated to a brief overview of the governance structure guiding the provision of statistics in the European Union (EU). It is followed by a presentation of the ESCB statistics quality framework, its development and synergies with other quality models. The third part of the paper presents the main quality assurance procedures put in place by the ESCB and applied to the whole statistical production chain (including the supporting IT infrastructure).

2. Governance structure guiding the provision of statistics in the EU

European statistics are developed, produced and disseminated, within their respective spheres of competence, by the European Statistical System (ESS)[2] and the statistical function of the ESCB.[3] The ESS and the ESCB operate under separate legal frameworks reflecting their respective governance structures. However, they closely cooperate in order to minimise the reporting burden, to guarantee the coherence of European statistics, to eliminate inefficiencies and work duplication, to enhance transparency and accountability, and to ensure the statistical quality necessary for European statistics.

Such cooperation takes place at both strategic and operational levels.

Cooperation at the strategic level is organised via the European Statistical Forum (ESF),[4] which is composed of one representative per Member State from the ESS Committee (ESSC), one representative per Member State from the ESCB Statistics Committee, one representative from Eurostat and one representative from the ECB. The ESF ensures the exchange of appropriate information related to the ESS and the ESCB statistical activities, discusses priority-setting and advises the two statistical systems on:

(i) the content and consistency of the statistical work programmes of both systems, making proposals for a better coordination of the programmes;

(ii) possible future challenges for European statistics and medium-term strategic visions and actions allowing for such challenges to be addressed by the ESS and ESCB; and

(iii) priority issues for cooperation between the two systems and the time horizon over which they should be addressed. The ESF also adopts an annual operational work programme for cooperation, which is implemented by the Committee on Monetary, Financial and Balance of Payments Statistics (CMFB)[5] (i.e. the operational platform). For other topics, relevant ESS/ESCB bodies are assigned according to their specific subject-matter competences, avoiding duplications of work and ensuring cost-efficiency.

Overall, the ESCB has prime responsibility for money, banking and financial market statistics, quarterly financial accounts and financial stability statistics, while the ESS has prime responsibility for general economic statistics and non-economic statistics. The two systems

[2] See Article 4 of Council Regulation (EC) No 223/2009 on European statistics available on March 2009 at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:en:PDF

[3] See Article 5 of the Statute of the ESCB and of the ECB available on March 2001 at: http://www.ecb.europa.eu/ecb/legal/1341/1343/html/index.en.html

[4] See the Memorandum of Understanding on the cooperation between the Members of the European Statistical System and the Members of the European System of Central Banks of 24 April 2013 available at: http://www.ecb.europa.eu/ecb/legal/pdf/mou_between_the_ess_and_the_escb.pdf

[5] See Council Decision 2006/856/EC of 13 November 2006 available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32006D0856&from=EN

(see Figure 1) have shared responsibility for balance of payments statistics, European sector accounts and statistical infrastructure.

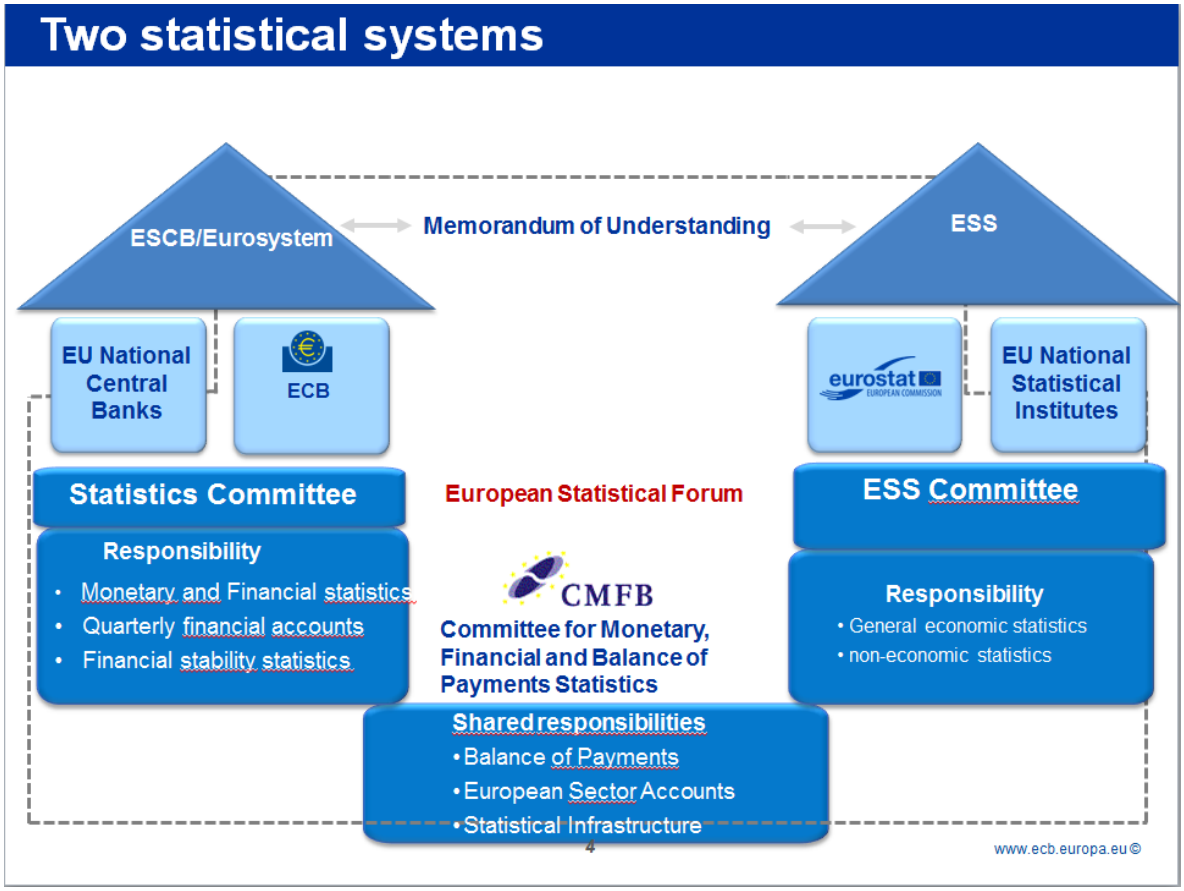


Figure 1: The two statistical systems.

3. Public commitment on European statistics by the ESCB: development and synergies with existing quality models

The "Public commitment on European statistics by the ESCB" (hereafter referred to as the "public commitment") was adopted by the ECB's Governing Council in May 2007. It was amended a first time in 2009, following the amendment of Council Regulation (EC) No 2533/98, with the definitions of the statistical principles governing the production of European statistics[6], and a second time in 2012 to enlarge the list of principles and provide practical indicators of good practice to characterise each principle in order to enhance the convergence of the ESCB and ESS quality frameworks.

3.1. Structure of the public commitment

The public commitment is structured around two main components:

(i) a concise, general, public statement; and

(ii) a list of quality principles and associated indicators of good practice grouped into three domains: the institutional environment, the statistical processes and high output quality, see Table 1.

[6]See http://www.ecb.europa.eu/stats/html/pcstats.en.html on April 2007.

Table 1: Public commitment on European statistics by the ESCB.

Institutional environment	**Statistical processes**	**Statistical output**
P1. Professional independence	P7. Sound methodology	P11. Relevance
P2. Mandate for data collection	P8. Appropriate statistical procedures	P12. Accuracy and reliability (incl. stability)
P3. Adequacy of resources	P9. Minimisation of reporting burden	P13. Timeliness (incl. punctuality)
P4. Commitment to quality	P10. Cost-effectiveness	P14. Consistency and comparability
P5. Statistical confidentiality		P15. Accessibility and clarity
P6. Impartiality and objectivity		

With regard to the first component, the purpose of the public statement is to first spell out the legal mandate of the ESCB statistical function which enables the collection of all necessary and relevant data[7] in the areas under the ESCB's responsibility and to recall that the ESCB statistical function produces European statistics in accordance with European and internationally agreed standards, guidelines and good practices. It also aims to highlight the following important elements:

(i) the independence of the ESCB statistical function in the compilation and dissemination of statistical information (in line with Article 130 of the Treaty[8] on the Functioning of the European Union);

(ii) the collaboration with the ESS, the NSIs and other national statistical authorities, taking into account the principles laid down in the European Statistics Code of Practice;[9] and

(iii) the efficient use of resources when collecting, compiling and disseminating statistics, keeping the reporting burden on respondents to a minimum, guaranteeing their privacy and protecting the confidentiality of the non-public information that they provide.

Concerning the second component, the principles and indicators selected in the public commitment are based on the principles of existing quality frameworks which were adapted to the institutional environment and operational features of the ESCB. From all of the frameworks analysed, five different models have been selected owing to:

(i) a similar scope and similar objectives;

(ii) interesting features in their implementation and monitoring; and

(iii) the existence of clear similarities between the institutional environment of the institution of application and the ESCB statistical function.

Table 2 below presents the elements that were of interest when drawing up the list of principles of the ESCB public commitment.

3.2. Synergies in terms of scope and objectives

The public commitment is a comprehensive quality framework addressed to users, data suppliers and producers of statistics alike. It aims to establish core principles for the compilation and dissemination of ESCB statistics.

The main foundation in terms of scope and objectives is the UN Fundamental Principles of Official Statistics. Both models aim to maintain public trust in official statistics, enhance their quality and establish standards and concepts allowing cross-country comparisons.

[7] i.e. monetary and financial statistics, payment and payment systems statistics, balance of payments and international investment position statistics and financial stability statistics.

[8] See http://www.ecb.europa.eu/ecb/legal/1341/1342/html/index.en.html available on October 2012.

[9] See http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF available on September 2011.

Table 2: Elements of interest when drawing up the list of principles of the ESCB public commitment.

Model	**Reason for interest**
United Nations (UN) Fundamental Principles of Official Statistics	- Provides the foundations for all quality frameworks
IMF Data Quality Assessment Framework (DQAF)	- Rooted in the UN Fundamental Principles - Provides an organised structure to assess existing practices against best practices, including international methodologies - Easily applicable to the ESCB statistical function
OECD Quality Framework	- Adaptation of the UN Fundamental Principles and the IMF DQAF to the OECD institutional environment, including a proximity to users in the same institution - Interesting monitoring of the implementation via the use of quality assurance procedures
Bank of England Code of Practice	- Based on the Code of Practice of the UK Office for National Statistics - Similar institutional environment to the ESCB - Concepts of cost-efficiency and non-excessive burdens on respondents - The Code is addressed to all stakeholders
European Statistics Code of Practice	- Coherent with the UN Fundamental Principles and the IMF DQAF - Of high relevance for the ESCB in view of the close cooperation with the ESS

Therefore, the scope and objectives of the public commitment are consistent with those of all models inspired by the UN Fundamental Principles of Official Statistics (e.g. the IMF DQAF, the Bank of England Code of Practice, the OECD Quality Framework and the European Statistics Code of Practice).

Regarding the set of quality principles, the public commitment shares the same holistic structure as the IMF DQAF, the Bank of England Code of Practice and the European Statistics Code of Practice. This structure supports the comprehensive assessment of data quality by covering the institutional environment, the whole statistical production chain (including the supporting IT infrastructure) and the characteristics of the statistical products.

3.3. Synergies in terms of implementation and monitoring features

With regard to the structure of the public commitment, the quality principles selected are directly inspired by the IMF DQAF and the European Statistics Code of Practice. However, in terms of presentation, the structure of the public commitment is identical to that of the European Statistics Code of Practice. In both models, the principles are organised into three main blocks: the institutional environment (ensuring the integrity and credibility of the production and dissemination of statistics); the statistical processes (dealing with the application of good practices regarding the processes used for the development, collection, processing and dissemination of statistics); and the statistical output (ensuring the fulfilment of users' needs). Each principle is then accompanied by a set of practical indicators of good practice.

The main difference between the public commitment and the European Statistics Code of Practice relates to the monitoring procedure for the implementation of the frameworks. While the ESCB uses a number of well-defined quality assurance procedures (see Section 4 below) to enable the monitoring and reporting of the implementation of the framework by the ESCB statistical function, the implementation of the European Statistics Code of Practice is monitored via the launch of regular peer review exercises. The implementation of the set of practical guidelines which accompanies the ESS framework is not mandatory. In this respect, the public commitment follows the system introduced by the OECD and to a lesser extent that by the Bank of England.

With regard to the selected quality principles, contrary to the IMF DQAF, both the public commitment and the European Statistics Code of Practice include quality principles dealing with the notions of cost-efficiency and non-excessive burden on respondents, which are deemed very significant for the ESCB statistical function. Moreover, due to the importance attached to it by the data providers, confidentiality constitutes an individual principle in the public commitment (and in the European Statistics Code of Practice), whereas in the IMF DQAF it is only an indicator of the quality prerequisites. Both issues are also part of the Bank of England Code of Practice.

3.4. Similarities between the institutional environments

The ESCB statistical function presents a number of similarities with the institutional environments of all five of the other institutions mentioned above. It is widely acknowledged that the institutional environment significantly affects the integrity and credibility of the production and dissemination of statistics. An important feature that characterises both the public commitment and the Bank of England Code of Practice is the statutory independence of the institutions concerned which also applies to their statistical activities. This is why this particular aspect is included in both quality models. The public commitment also integrates the notion of professional independence of the statistical function, which is present in all models selected.

3.5. Alignment with the European Statistics Code of Practice

As mentioned previously, the public commitment was last amended in 2012. As a result, the list of principles was enlarged and each principle was characterised by a set of practical indicators of good practice. This last amendment was mainly triggered by the need to enhance the convergence of the ESCB and ESS quality frameworks.

Concretely, this enhancement resulted in the ESCB amending the public commitment with: (i) practical indicators of good practice based on the mapping of the indicators of the ECB Statistics Quality Framework with those of the European Statistics Code of Practice; and (ii) five new principles that were missing compared with the principles of the ECB Statistics Quality Framework and of the European Statistics Code of Practice.

The updated ESCB public commitment is now more complete and better structured as it is fully aligned with the ECB Statistics Quality Framework and the European Statistics Code of Practice.

4. Quality assurance procedures associated with the public commitment

The ESCB statistical function has at its disposal a precise set of quality assurance procedures, associated with each principle, to ensure the adherence of the ESCB to the quality principles. But these procedures are not yet formalised via official guidelines made available on the ECB's website.

4.1. General presentation

The aim of the quality assurance procedures is to provide information on the activities, methods and tools used by the ESCB to implement the public commitment. They address all key stakeholders and cover the whole statistical production chain, namely the development, collection, compilation and dissemination of statistics, as well as the supporting IT infrastructure. They can be grouped into the following domains:

Governance issues: legal framework; cooperation within the ESCB, cooperation with other international organisations and bodies.

Strategy, work programme and procedures to identify new user requirements and develop new statistics: the ESCB statistical function's medium-term work programme and annual work programmes; merits and costs procedure.

Protection of statistical confidentiality: legal requirements; IT infrastructure; rules and monitoring ("Annual Confidentiality Report").

Quality assurance procedures related to the collection of data: data transmission calendars; data compliance monitoring; metadata management; data collection standards, methods and tools.

Quality assurance procedures related to compilation and statistical analysis: completeness checks; revision studies; plausibility checks; internal consistency and consistency across frequencies; regular quality assessment notes.

Quality assurance procedures related to data accessibility and dissemination policy: adherence to the IMF Special Data Dissemination Standard; data release calendars; press releases; regular ECB publications; online access via the ECB's website; joint tables of euro area statistics and national breakdowns on the websites of the ECB and euro area NCBs; Statistical Data Warehouse; statistics hotline; Real Time Database.

Monitoring and reporting: process management; audit reviews; ECB Annual Report; annual quality reports and output quality indicators.

Special attention should be given to the specific procedures in place concerning the review of the efficiency and effectiveness of the ESCB statistical processes and the monitoring of the quality of the ESCB outputs.

4.2. Monitoring of statistical processes

Currently, the monitoring of the ESCB statistical processes is performed by the ESCB Internal Audit Committee (IAC). The IAC regularly reviews the different statistical domains and the ESCB statistical framework in general. During these reviews, ESCB auditors generally assess the compliance of the ECB and all EU NCBs with a number of principles and indicators of good practice of the ESCB public commitment. Among other things, the IAC has reviewed past statistical audits of whether new statistical reporting requirements were

backed by sound user requirements and were evaluated using a cost/benefit analysis, whether statistical reporting requirements were consistently implemented into national collection and compilation procedures, and whether data gaps and shortcomings in the data collection and/or compilation procedures were properly identified, reported and followed up by the ESCB statistical function. When reviewing the whole process management, auditors also assess the overall efficiency of the statistical processes within the ESCB statistical function, whether all statistical processes are documented in a consistent manner and whether the appropriate risk management and change management procedures are implemented. For the ESCB statistical function, the main advantages of the ESCB audit reviews are:

(i) the same audit reviews are performed simultaneously in the whole ESCB; and

(ii) they are carried out by a team of independent auditors, not (peer) statisticians, specialised in statistical issues.

4.3. High output quality monitoring and reporting

The monitoring and reporting of the availability and quality of all European statistics produced by the ESCB is based on the annual assessments conducted by the ECB, the results of which are set out in four separate reports, one for each statistical field, namely: quarterly financial accounts; balance of payments and international investment position statistics; annual government finance statistics; and monetary and financial statistics. The main objectives of these quality reports are to:

(i) provide a good overview of the availability of the European statistics compiled and disseminated by the ESCB;

(ii) assess their quality in terms of punctuality, timeliness, methodological soundness, reliability and stability;

(iii) report on the state of affairs concerning the implementation of enhanced or new statistical requirements in the Member States; and

(iv) provide information on possible further improvements to data collection and compilation methods and the level of detail and/or frequency of input data. Quality aspects are assessed by using a set of quantitative indicators covering, for example, the frequency, number, direction, magnitude and patterns of revisions, and assessing internal and external consistency of data (e.g. the size of net errors and omissions; comparisons across statistics).

5. Conclusions and outlook

The public commitment as it stands now is a very useful instrument for the ESCB statistical function in two ways. First, it is a good communication tool to help in maintaining the confidence of the general public regarding the quality and independence of the European statistics produced by the ESCB. Second, it provides staff of the ESCB statistical function with a useful complementary benchmark in their day-to-day compilation work and release of statistics and in the development of new statistical products.

However, striving for the best possible quality in terms of statistical output and statistical processes is a continuous task for statistical authorities. While already well-developed, the quality management and assurance procedures continue to be scrutinised by the ESCB and will be further fine-tuned in the future. In this context, the following initiatives in particular are under way.

The case for a comprehensive ESCB quality framework is currently being further investigated. Making explicit in one document the quality principles applied and the quality assurance procedures followed would further increase the transparency of the ESCB's statistical procedures.

Another ongoing activity relates to a further enhancement of the already fruitful collaboration with the ESCB auditors. The ESCB statistical function is currently envisaging to further

formalise the monitoring of the implementation of the principles associated with the statistical processes by the ESCB auditors. Work is under way to establish a comprehensive list of all aspects that need to be integrated in the scope of all statistical audits when auditors are planning their reviews.

Last but not least, user-friendly access to statistical data and metadata remains a continuous challenge that is receiving much attention from ESCB statisticians.

In all of these areas, and in the work on statistical quality in general, the value of sharing good practices with other international and national statistical authorities can hardly be overestimated. The biennial quality conferences provide an excellent opportunity in this respect.

Affiliation:

Aurel Schubert
European Central Bank
Kaiserstrasse 29
60113 Frankfurt, Germany
Phone: +49-69-1344-7555
Fax: +49-69-1344-7693
E-mail: aurel.schubert@ecb.europa.eu

Catherine Ahsbahs
Directorate General Statistics
European Central Bank
Kaiserstrasse 29
60113 Frankfurt, Germany
E-mail: catherine.ahsbahs@ecb.int

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 44 *Submitted:* 2014-08-29
April 2015 *Accepted:* 2014-10-17

Austrian Journal of Statistics
April 2015, Volume 44, 13–30.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i2.60

Measuring Nonresponse Bias in a Cross-Country Enterprise Survey

Katarzyna Bańkowska, Małgorzata Osiewicz*, Sébastien Pérez-Duarte
European Central Bank, Frankfurt am Main

Abstract

Nonresponse is a common issue affecting the vast majority of surveys. Efforts to convince those unwilling to participate in a survey might not necessary result in a better picture of the target population and can lead to higher, not lower, nonresponse bias.

We investigate the impact of nonresponse in the European Commission & European Central Bank Survey on the Access to Finance of Enterprises (SAFE), which collects evidence on the financing conditions faced by European SMEs compared with those of large firms. This survey, conducted by telephone bi-annually since 2009 by the ECB and the European Commission, provides a valuable means to search for this kind of bias, given the high heterogeneity of response propensities across countries.

The study relies on so-called "Representativity Indicators" developed within the Representativity Indicators of Survey Quality (RISQ) project, which measure the distance to a fully representative response. On this basis, we examine the quality of the SAFE at different stages of the fieldwork as well as across different survey waves and countries. The RISQ methodology relies on rich sampling frame information, which is however partly limited in the case of the SAFE. We also assess the representativeness of the SAFE particular subsample created by linking the survey responses with the companies' financial information from a business register; this sub-sampling is another potential source of bias which we also attempt to quantify. Finally, we suggest possible ways how to improve monitoring of the possible nonresponse bias in the future rounds of the survey.

Keywords: business survey, representativeness, bias, nonresponse, R-indicators.

1. Nonresponse bias and its measurement

Nonresponse bias occurs when the survey estimates for the respondents are different from those who did not answer to the survey. While initially the nonresponse was treated as a fixed characteristic of a respondent, the more currently popular stochastic approach assumes that people have a certain probability ρ_i of participating, which varies depending on circumstances. In this sense, the bias of the respondents' mean $\bar{y}_r$ is approximated by $\frac{\sigma_{y\rho}}{\bar{\rho}}$, where $\sigma_{y\rho}$ is the population covariance between the survey variable, y, and the response propensity, ρ, and $\bar{\rho}$ is the mean propensity in the target population over sample realisations (Groves 2006).

*Corresponding author: malgorzata.osiewicz@ecb.europa.eu

However, the relation between the response propensities and the nonresponse biases is not straightforward and higher response rates do not necessarily lead to lower bias, if higher efforts to convert the nonrespondents are effective only for particular groups, e.g. in a business survey, larger companies or enterprises encountering financial difficulties. Groves (2006) presents the absolute relative bias together with corresponding response rate for over 200 estimates from 30 different methodological studies and shows weak correlation between the two. Interestingly, most of the variation comes from the estimates within the same survey.

Dependent on the available information, various approaches are applied to analyse the nonresponse (Montaquila & Olson 2012). First, the survey estimates can be compared to the external sources, like administrative records. In this case, highly accurate benchmark and consistent measurement of analysed indicators between both datasets are prerequisite to the meaningful evaluation.

A second set of methods compares the survey estimates under alternative weighting schemes using additional characteristics associated with the key survey estimates or response propensities. Sensitivity of the results to different weighting would indicate the presence of nonresponse bias. On the other hand, no or insignificant differences might stem rather from lack of good predictors than absence of bias.

A third approach relies on the information from the sampling frame and observations collected during the fieldwork for the whole sample. Such data are the basis for the calculation of different statistics (e.g. sample means, proportions) separately for respondents and nonrespondents or various reasons for nonparticipation (noncontact, refusal). Additionally for longitudinal studies, past information on the initial respondents, who turned nonrespondents in the subsequent rounds, help to detect response patterns and possible causes of attrition (National Research Council 2013). Furthermore, the auxiliary sample information allows computing response rates by characteristics. Within the respondent set, the survey estimates can be presented for cooperative and more reluctant respondents, measured by variables like number of call attempts, early versus late respondents, provided incentives and techniques used for refusal conversion. Large variation between specific subgroups would point to the potential bias and its source. R-indicators, which are the focus of this paper, fall also into this set of methods for nonresponse analysis.

Fourth, follow-up surveys, aimed at collecting information on the initial nonrespondents, are another possibility to investigate how distinct they are from the respondents. Such studies usually apply enhanced recruitment techniques, different survey modes, and shorter questionnaires targeted on the main variables. Apart from the drawbacks of the extra cost and the extended fieldwork, achieving high response rate in the follow-up survey is essential, which might prove a difficult objective[1].

In this paper, we apply the third approach based on the sample information to the Survey on Access to Finance of Enterprises (SAFE), with the main focus on the R-indicators developed within the Representativity Indicators of Survey Quality (RISQ) project[2].

SAFE is a qualitative telephone survey conducted with the purpose of providing regular information on the financing conditions of micro, small and medium-sized enterprises (SMEs). A sample of large firms (250 employees or more) is also included in order to be able to compare developments for SMEs with those for large firms. A subset of the survey is run by the ECB every six months to assess the latest developments of the financing conditions of firms in the euro area countries. A more comprehensive version of the survey with an extended questionnaire is run every two years, in cooperation with the European Commission. The survey is conducted by an external survey company. The sample is a probability sample based on quo-

[1] Additional data collection can also take the form of randomised nonresponse experiments, where different design features (e.g. "warm-up" questions, mode) are assigned to different random subsamples. The results and the response rates of the treatment groups are then compared and effective design identified, although it might be challenging to find one treatment which performs well in terms of reducing nonresponse bias, not only for a particular group, but for the full sample (Kruskal & Mosteller 1979).

[2] http://www.risq-project.eu/

tas by country and size. The SAFE has also a rotating opt-in panel component – at the end of the interview the respondents are asked whether they would like to participate in the future survey rounds. Around 80% of firms agree, however, afterwards only a part is successfully re-contacted. As a result, panel constitutes currently around 50% of the respondents.

Given the restricted length of phone interview and respondent's difficulties in answering questions related to quantitative accounting elements, to obtain balance sheet information of the interviewed companies, the survey data are matched with the quantitative financial information from the Bureau van Dijk's Amadeus database.

The objective of this study is to examine the representativity of the SAFE sample, as well as the subsample containing the matched financial information. This paper gives first the rationale for applying the R-indicators to the probability sample based on quotas. Secondly, we present an overview of the nonresponse in SAFE. In the following sections, we describe briefly the methodology of various types of R-indicators and present the implementation of the indicators in SAFE and the matched dataset of SAFE and Amadeus. In final section, we conclude and give the recommendation for fieldwork monitoring.

2. Probability sampling based on quotas in the SAFE

A word is warranted on the nature of the sample in the SAFE, as "quota sample" carries a negative connotation among survey statisticians, and indeed, when improperly done, data collected through such a sample offer no guarantee of representativity and do not allow any sort of probabilistic analysis. However, the SAFE sample is very far from the quota samples of the 1950s where interviewers had to choose a convenience sample respecting quotas. The SAFE sample follows the work of Sudman (1966) in order to confer probabilistic properties to quota sampling.

We describe the selection of the sample of first-time participants in the survey; panel firms are not considered here[3]. The sample is drawn from the Dun & Bradstreet company database, which has the benefit of adequately, if not perfectly, covering the universe of enterprises in the euro area. From Dun & Bradstreet, a stratified random sample is drawn, with strata composed of country (11 in the euro area surveys) and size class (4 such classes). In line with other cold-call business surveys, response rates are very low. Consequently, the initial sample is 10 to 15 times larger than the desired sample, to account for nonresponse.

As in other surveys working with firm data in a multinational setting, we assume that the Dun & Bradstreet population is a good image of the population of firms. The total number of firms in the target population is known from Eurostat's Structural Business Statistics, by country, sector, and size class. If, conditional on country, sector and size class, firms have the same probability of being included in Dun & Bradstreet, then firms not in that register can be considered to be missing at random (MAR, in Donald Rubin's terminology). Hence, the initial sampling probability can be estimated for all firms in the population and thus in the initial sample.

The interviews are based on this initial sample, with targets or quotas for the number of interviews conducted by country and size class (the same as above). The initial sample is randomly sorted, and the firms are dialled from this sample. Up to ten calling attempts are made to each address, at different times or even outside normal office hours. Call-back appointments are not subject to the limit of ten attempts. From this interviewing strategy, a certain number of firms will not have been called at all ("fresh" sample), some firms will have been called and not contacted ("non-contact"), others contacted but they refuse to participate ("refusal") and others successfully interviewed ("respondents"). At the end of the fieldwork, some firms will still be "fresh" and will be so at random (conditional on the quota cell).

[3]For the description of the panel selection, see section 1.

In order to analyse response behaviour and response rates across countries, those "fresh" firms are dropped from the initial sample. The initial number of records drawn from the register is a decision of the survey company based on the past response rates in the SAFE and similar studies. Usually, a sample ten times larger than the targeted number of interviews is sufficient. However, in some countries with lower quality of the contact information (e.g. incorrect telephone numbers, out-of-date records) or lower than expected cooperation rates, there is a need of topping-up the initial sample with additional fresh records. Thus, the amount of unused sample would not be comparable across countries as the ratio of the initial sample to targeted interviews varies. However, even if this ratio was the same in each quota, the amount of the fresh sample is an arbitrary decision and should not be taken into account in the analysis of the response indicators[4]. Consequently, the unused records are removed and only the records, where at least one contact attempt was made, are taken into account in the analysis.

During fieldwork, however, the way the fresh sample is integrated into the calling roster is crucial for the probabilistic nature of the sample. A firm in the fresh sample should not be called only because it is more probable to conclude the interview than trying to contact again the firm, for which the previous contacts were unsuccessful. If this is the case, then the quota sampling is not less probabilistic than a probability sample where nonresponse causes randomness in the firms that are interviewed. Of course, since the survey has a tight deadline and priority is given to the timeliness of the results, towards the end of the fieldwork it is more likely that not enough contact attempts for the firms in the calling roster are carried out. We study this phenomenon in section 5.2 below, when we consider the representativity of the sample through the length of the fieldwork.

The final estimation weight is then obtained by calibrating on official counts by country, size class, and sector (4 main sector groupings), correcting in this manner differential response rates as long as the nonresponse can indeed be considered conditionally random by country, sector and size class.

One interesting theoretical aspect that would need to be further explored in connection with the R-indicators is the randomness of the effective initial sample (excluding the fresh firms) and the fixed number of firms in the final, respondent sample, which is the converse of the standard probabilistic setup of fixed initial sample but random final one. We consider this issue to be of a secondary nature in the measure of the representativity of the final sample, and will hence take the effective initial sample as the true initial sample and the final sample as the result of the interviewing process of all the firms in the initial sample.

3. Nonresponse in the Survey on Access to Finance of Enterprises (SAFE)

A common problem across nearly all types of surveys is low response rates, which in fact have dropped substantially over the last decades (see e.g. National Research Council 2013, p. 12-30). A low response rate is also a concern for the SAFE. The overall response rate reached around 14% in the last survey rounds[5], below those of other business surveys run by central banks. While these other surveys are not directly comparable, given the differences in how they are conducted, in absolute terms the response rates for the SAFE can nevertheless be objectively deemed low. As this may be a source of uncertainty about the quality of the results, in this paper we apply R-indicators to analyse from several angles possible nonresponse bias and its origin.

[4] To illustrate it, we can consider two initial samples: one ten times larger and another one hundred times larger than the number of targeted interviews. Computed responses rate would be very different for those two scenarios, although the response behaviour is the same.

[5] Response rate 3, following the definition of outcome rates advocated by AAPOR (see American Association for Public Opinion Research 2011). Since the original AAPOR definitions refer to household surveys, they were adapted to the features of a business survey.

In the first step, we present the outcome rates for the SAFE by main characteristics of enterprises: country of residence, size and sector. In addition, we split firms into those which participate for the first time in the survey (non-panel firms) and those which took part at least in one of the earlier survey rounds (panel).. Those results will be later cross-checked with the findings coming from the R-indicators. We focus on the three latest survey rounds (8^{th} to 10^{th}) as detailed information on the full sample including nonrespondents, was not available in the earlier rounds. When computing response and cooperation rates, break-off interviews are treated as nonresponse. In case of unknown eligibility, the proportion of cases of unknown eligibility that are eligible is estimated[6] and increased from 0.6 in 8^{th} survey round to 0.8 in the 10^{th} round, which is rather conservative, since the higher this proportion, the lower the response rate. While contact, cooperation and response rates vary considerably across countries, neither companies' sector nor size class have a large impact on the response rates (small firms have a slightly higher propensity to participate, while construction firms have a lower one; see Figure 1). The largest divergence shows between panel and non-panel enterprises with relatively high response rate of 40% for panellist in 8^{th} survey round, either through a positive image of the survey acquired through previous participation or a higher propensity to participate (see Figure 2).

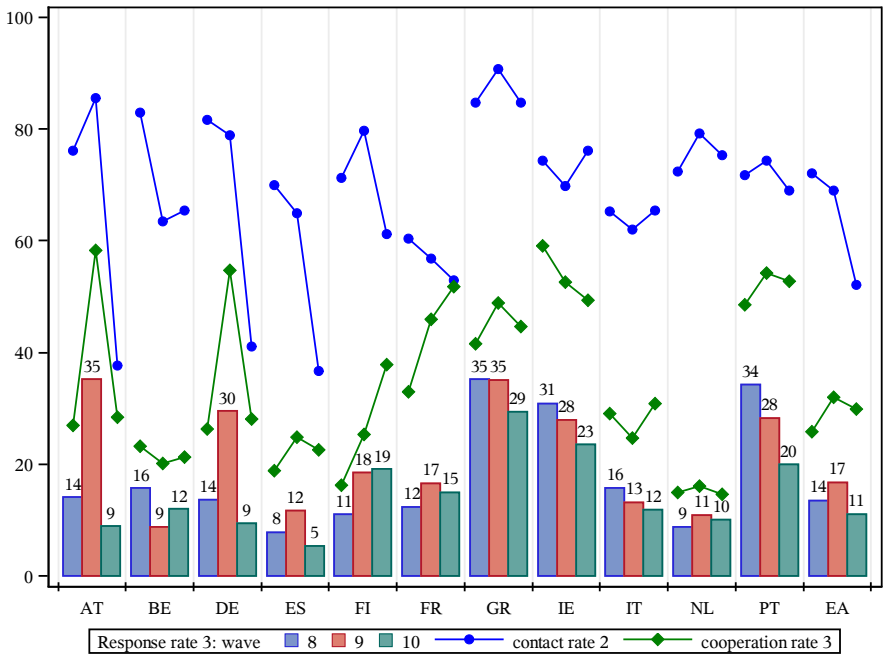


Figure 1: Outcome rates for SAFE from 8^{th} to 10^{th} survey round by country.
Note: Residency of a firm is indicated by country ISO-3166 code (AT – Austria, BE – Belgium, DE – Germany, ES – Spain, FI – Finland, FR – France, GR – Greece, IE – Ireland, IT – Italy, NL – the Netherlands, PT – Portugal). EA stands for aggregated figure for all presented euro area countries combined.

Country variation can stem from many factors. First, cultural differences play a role. In some countries, the respondents strongly refuse to participate, asking to be excluded from any future

[6]Following the definitions:

- response rate 3: I/((I+P) + (R+NC+O)+ e*U),
- cooperation rate 3: I/(I+P+R),
- refusal rate 2: R/((I+P)+(R+NC+O) + e*U),
- contact rate 2: ((I+P)+R+O) / ((I+P)+R+O+NC+ e*U),
- e: (I+P+R+NC+O)/(I+P+R+NC+O+NE),

where I – Interview, P – Partial interview, R – Refusal, NC – Non-contact, O – Other contact (non-refusals), U – Unknown if firm, NE – Non-eligible, e – the estimated proportion of cases of unknown eligibility that are eligible.

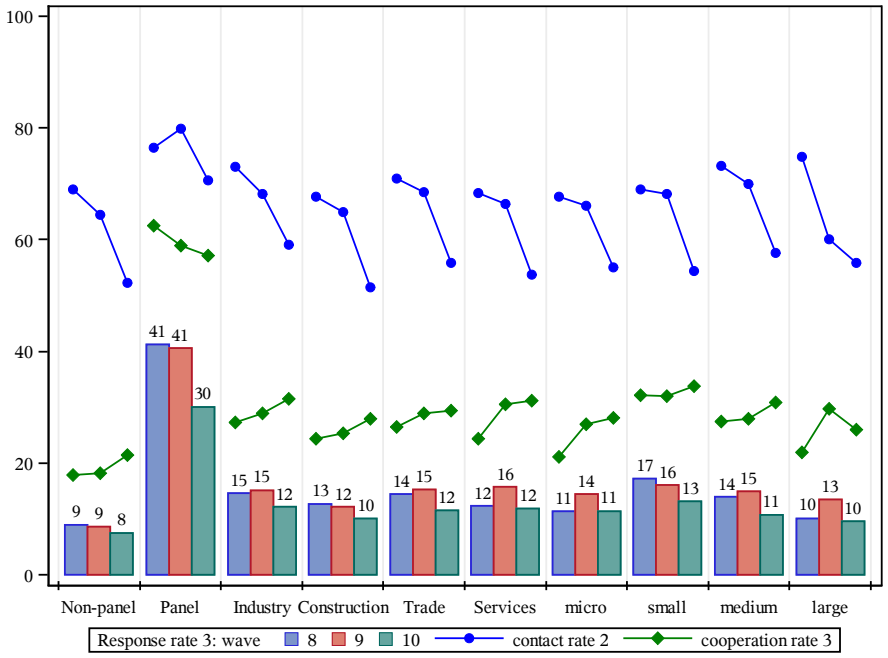


Figure 2: Outcome rates for SAFE from 8th to 10th survey round by panel dummy, sector and size (excluding Austria and Germany).

surveys conducted by the survey company, while in other countries, where the refusals are softer, good interviewers can more easily convince initial nonrespondents to eventually take part in the study. Second, the quality of the sampling frame differs across countries. The low quality of the enterprises' contact information, number of employees or sector will result in unsuccessful phone calls (in case of wrong company's number) or necessity to exclude a respondent after the screener questions (in case of SAFE, if the firm is non-profit, has no employees other than the owner or belongs to a sector which is out of the scope of the SAFE). Third, the situation in the local offices of the survey company, such as the experience and training of the interviewers, work load at the time of conducting the survey can also have an impact on the response rate. In case of SAFE, additional factor which can explain the divergences is different CATI system used by the survey company in Germany and Austria and it is apparent that the outcome codes are not fully harmonised with offices in other locations. For that reason, we excluded those two countries from the subsequent analysis.

4. R-indicators as a measure of representativity

The concept of "representativeness" does not have single clear interpretation. Kruskal & Mosteller (1979) review the statistical and other scientific literature and divides the meaning of term "representative" into no less that nine different groups, varying from "general acclaim for data", through "miniature of the population" to "representative sampling as permitting good estimation".

Representativity indicators (R-indicators) are based on definition linked to the mechanism of Missing Completely at Random (MCAR) and individual response propensities. Following Schouten, Bethlehem, et al. (2012, p. 384), "response is called representative with respect to [the vector of auxiliary variables] X when the response propensities of all subpopulations formed by the auxiliary variables are constant and equal to the overall response rate", in other words, "when the respondents form a random subsample of the survey sample". In this sense, the R-indicators attempt to capture the overall impact of the nonresponse for the whole survey, and not only at the level of a particular estimate.

Although it is not the point of this paper to describe in details the theoretical properties of the R-indicators, which is much better done in Shlomo & Schouten (2013) or in Schouten, Cobben & Bethlehem (2009), we present their definition and main features.

The R-indicator is based on the standard deviation of the response propensities transformed to lie between 0 and 1, where 1 is representative response: $R = 1 - 2S(\rho)$. The response propensities and then the variance of the response propensities are estimated, leading to the following estimator of R:

$$\hat{R} = 1 - 2\hat{S}\left(\hat{\rho}\right) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{n} d_i(\hat{\rho}_i - \hat{\bar{\rho}})^2}$$

where d_i are the design weights, $\hat{\bar{\rho}} = \frac{1}{N}\sum_{i=1}^{n} d_i\hat{\rho}_i$ is the weighted sample mean of the estimated response propensities and N is the size of the population (see Shlomo & Schouten 2013, p. 4).

It can be shown that the lower bound of the R-indicator (see Schouten, Cobben & Bethlehem 2009, p. 104) depends on the response rate: $R \geq 1 - 2\sqrt{\bar{\rho}(1-\bar{\rho})}$. Notably, it reaches its minimum of 0 for response rate of 0.5, i.e. when the individual response propensities can have largest variation, while it increases when the response rate decreases from 0.5 to 0.

The decomposition of the variance $S^2(\rho)$ into between- and within components of the response propensities for the sample subgroups is the foundation of the partial R-indicators at variable level. The unconditional partial R-indicator corresponds to the between subgroup variance, while the within variances are the basis for the conditional partial indicators (Schouten, Bethlehem, et al. 2012). Those indicators can be further decomposed into the category level R-indicators showing the contributions to the variation of the respective categories (de Heij, Schouten, Shlomo 2010).

	Unconditional	**Conditional**
$S^2\left(\rho\right) =$	$S^2_{between}\left(\rho\right)$	$S^2_{within}\left(\rho\right)$
Variable level	$P_U(X_k) = \sqrt{\frac{1}{N}\sum_{h=1}^{H} n_h(\bar{\rho}_h - \bar{\rho})^2}$	$P_C(X_k) = \sqrt{\frac{1}{N}\sum_{l=1}^{L}\sum_{i\in U_l} d_i(\rho_i - \bar{\rho})^2}$
Category level	$P_U(X_k, h) = \sqrt{\frac{n_h}{N}}(\bar{\rho}_h - \bar{\rho})$	$P_C(X_k, h) = \sqrt{\frac{1}{N}\sum_{l=1}^{L}\sum_{i\in U_l} d_i\Delta_{h,i}(\rho_i - \bar{\rho}_l)^2}$
Notation	X_k is a categorical variable with H categories and it is a component of the vector $\underline{X}$. $n_h = \sum_{i=1}^{n} d_i\Delta_{h,i}$ is the weighted sample size in the category h, where $\Delta_{h,i}$ is a 0-1 dummy variable for sample unit i being a member of stratum h. U_l is a cell in the cross-classification of all model variables except X_k.	

Standardised maximal absolute bias (in short "maximal bias"), in the worst case scenario, if the nonresponse correlates maximally with the variable of interest is $B_m\left(X\right) = \frac{1-R(\rho)}{2\bar{\rho}} \leq 1 - \bar{\rho}$ and it can be shown that it cannot be larger than the nonresponse rate (see Schouten, Morren, et al. 2009).

5. R-indicators for SAFE survey

For the computation of R-indicators and associated statistics, we used the SAS code available at the website of the RISQ project[7] (see also de Heij, Schouten, Shlomo 2010) for the methods of bias adjustment and computation of confidence intervals of the R-indicators).

[7] http://www.risq-project.eu/tools.html; We would like to thank Natalie Shlomo for providing additional SAS code for stratified simple random samples and useful suggestions.

The main requirement for the computation of the R-indicators is the availability of the auxiliary information from the sampling frame. The microdata for the whole sample of SAFE were provided only from 7th survey round, although not fully harmonised yet, and contain detailed outcome codes of a phone call (interview, refusal, answering machine, etc.), size class and sector from business register Dun & Bradstreet (D&B) and a dummy for panel firms (only from 8th survey round onwards). We also have the date of the last attempt or contact, which in case of respondent is the time of the interview.

Although the methods to estimate representativity were not designed for quota samples, and consistent with the description of the probability sample based on quotas used in the SAFE, we will neglect this issue in this paper and assume that the respondents were obtained through a simple random sample. We will consider that every firm for which a contact was attempted (the "non-fresh" sample) is to be included in the sample as a nonrespondent. Since the objective of the paper is to assess the influence of the firm characteristics on the response behaviour, we do not use the R-indicators for stratified samples, as this would mask the impact of stratification variables (country and size). However, for comparison we computed the R-indicators for stratified samples[8]. As expected, the overall R-indicator improves; however, the effect of the remaining variables (sector and panel) is similar to the presented results without stratification.

All R-indicators were computed using four above mentioned variables, i.e. country (9 euro area countries), size class (micro, small, medium and large), sector (industry, construction, trade and services) and panel dummy. The response propensities were estimated by a logistic regression with all mentioned variables as predictors, without interactions.

5.1. R-indicators across survey rounds (8 to 10)

We start the examination from the R-indicators for each survey round looking at the overall response and contact rates. It would be possible to split the response process into successive sub-processes of contact, cooperation and final response, as it was done in Schouten, Bethlehem, et al. (2012). However, being unsure to which extent the outcome codes are harmonised among countries, we limit this initial analysis to two processes mentioned.

Interestingly, the R-indicator for overall response is the lowest for the 9th round, although the highest response rate was achieved in that round (see Table 1). Notably, it was the time when longer questionnaire was used. We cannot draw conclusion from this one observation, but it would be recommended to monitor in the future the development of the nonresponse bias in the rounds with the extended questionnaire.

Table 1: R-indicators and other associated information for the survey rounds 8 to 10.

	Response			**Contact**		
Round	**8**	**9**	**10**	**8**	**9**	**10**
Total sample	70,432	58,689	62,090	70,432	58,689	62,090
Response rate 3 / contact 2	13.4%	15.0%	11.6%	70.0%	67.3%	55.5%
R-indicator	0.853	0.822	0.859	0.725	0.686	0.666
Standard error	0.003	0.004	0.003	0.003	0.003	0.003
Ave propensity	0.085	0.102	0.097	0.622	0.651	0.556
Maximal bias	0.863	0.868	0.729	0.221	0.241	0.300
Lower bound for R	0.441	0.394	0.408	0.030	0.047	0.006

A higher response rate does not guarantee better representativeness. For instance, the R-indicator for the response is the highest and maximal bias is the lowest for round 10, although

[8] Available upon request.

the response rate was higher for the previous round (see Table 1 and Figure 3). It is also useful in the analysis of the overall representativeness to look at the maximal bias, especially since it is not sensitive to the level of the response rates. Figure 3 illustrates that R-indicators are higher for the overall response process than for the contact, but the maximal bias is much lower for the contact. It seems that other sub-processes of the overall response behaviour (such as cooperation of the respondents) may play a bigger role and contribute to the potential loss of representativeness.

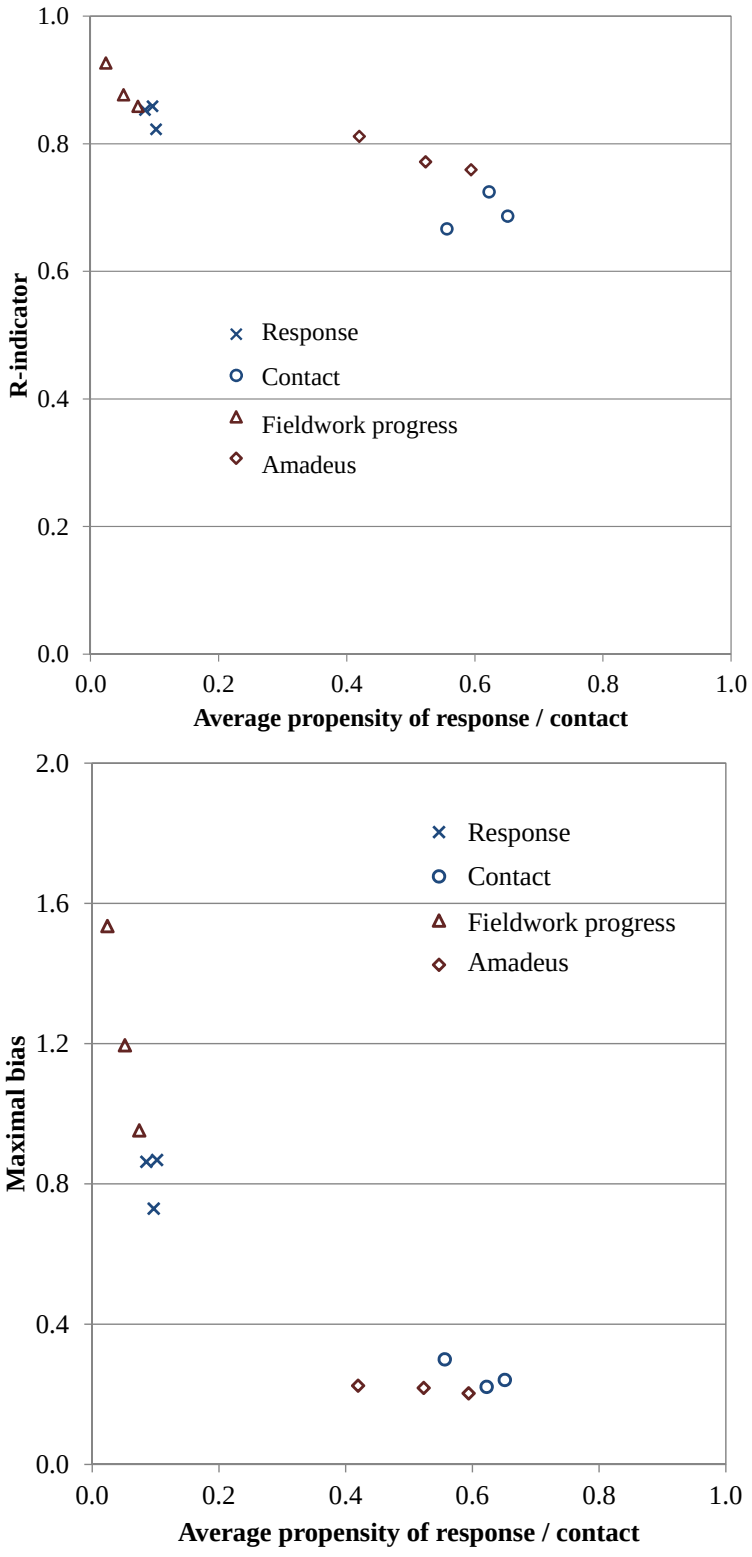
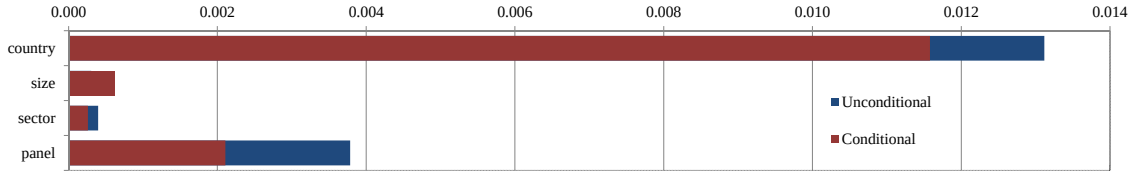
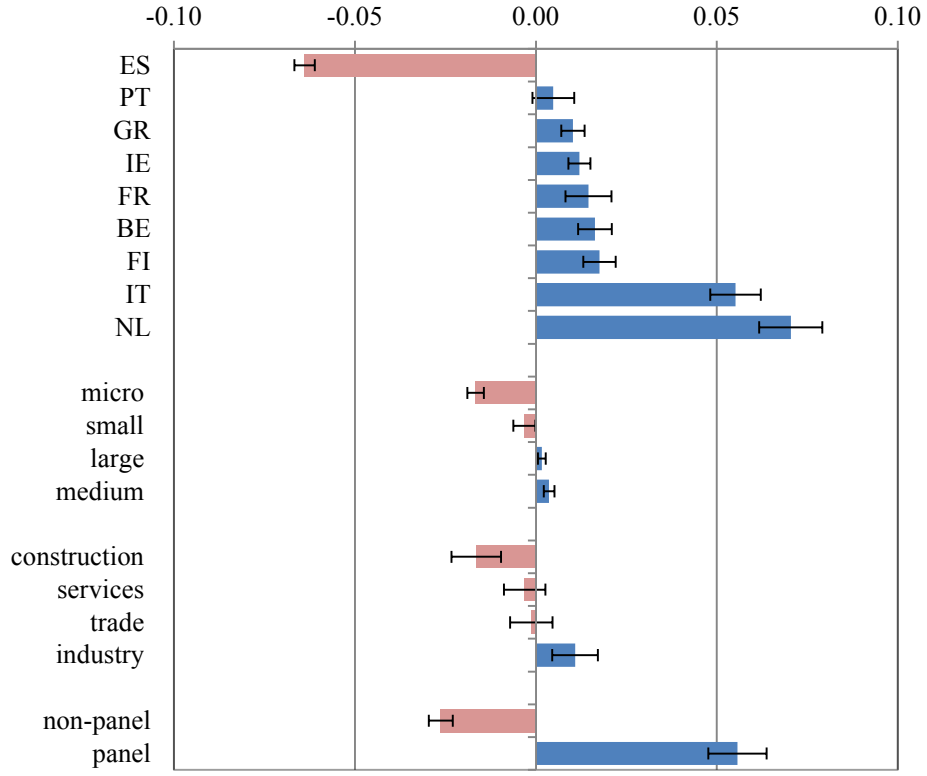


Figure 3: R-indicators and maximal bias as a function of the average propensity of response/contact for the survey rounds 8 to 10.

Looking at the R-indicator corresponding to contact propensities, the 10th survey round scores the worst. It was already visible from the investigations of outcome rates, where the contact rate dropped dramatically from round 9 to 10, particularly in three countries: Austria, Germany (both excluded from the analysis) and Spain[9]. In this case, low contact rate is also associated with higher bias – the large negative unconditional values for R-indicator point to the underrepresentation of Spanish businesses in the pool of contacted enterprises, while the Netherlands and Italy with high positive unconditional values are in comparison overrepresented (see Table 4 and Figure 4).



(a) Variable level: Conditional and unconditional partial indicators



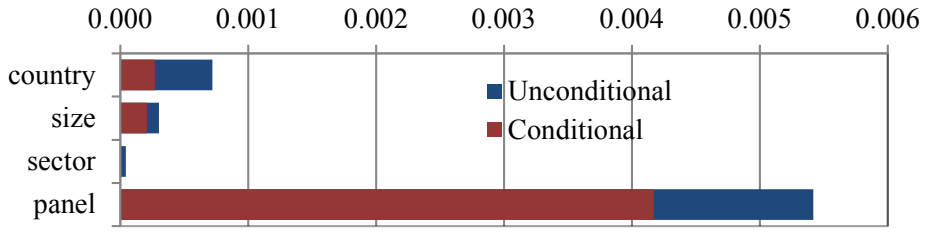
(b) Category level: Unconditional partial indicators with 95% confidence bands

Figure 4: Partial indicators for contact in 10th survey round.

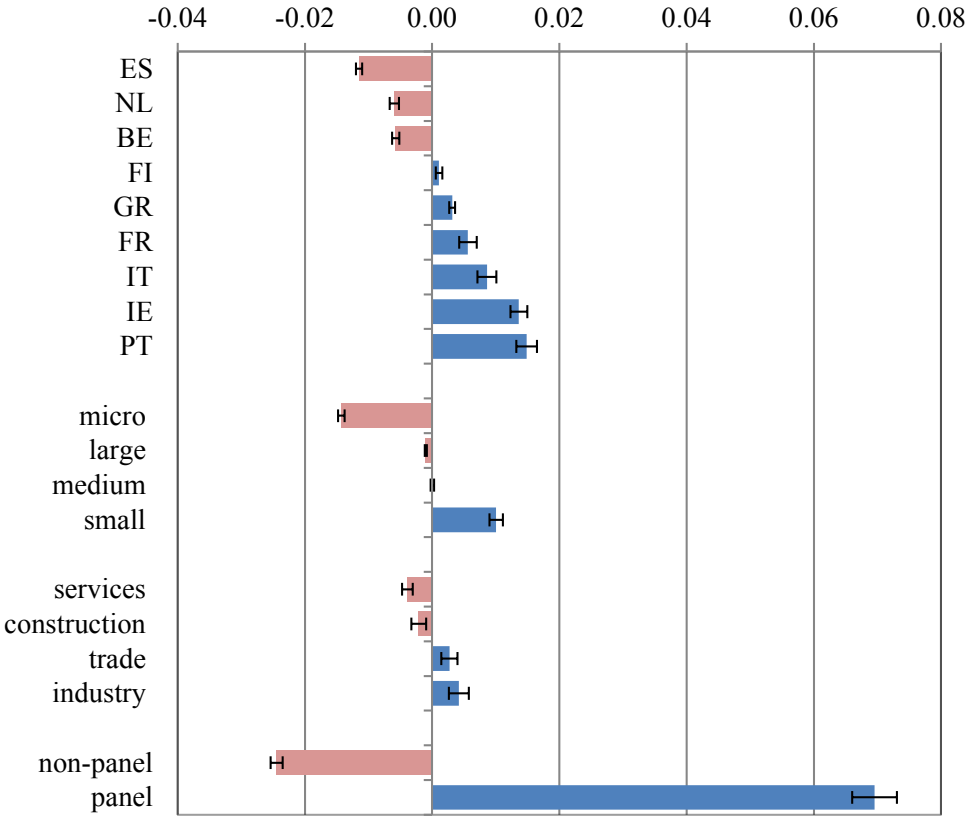
More generally, with respect to contact the unconditional and conditional partial R-indicators are the highest for the variable country and the country variation contributes the most to the loss of representativeness in all examined survey rounds. It seems that enterprises in some countries are more difficult to contact than in other regions, which points out also to the issues with the quality of the sampling frame. For SAFE, enterprises are all sampled from Dun & Bradstreet; however, the availability and accuracy of the contact information is not homogenous, given that the underlying sources of information differ by country. Consequently, it would be recommended to increase the efforts in the improvement of the sampling frame.

[9]The disproportionately high non-contact rate in drop in the 10th wave was a result of approaching relatively many enterprises at the beginning of the fieldwork. Enterprises, which were not contacted successfully, were not re-approached since the quotas were already filled. In other countries, such companies would be re-contacted and possibly converted into the respondents.

If we turn to the overall response, unsurprisingly, the fact whether the enterprise belongs to the panel or not plays the biggest role while the company's characteristics, such as country, size and sector are not statistically significant at the variable level (see Table 4 and Figure 5). This is consistent with the earlier finding about much higher response propensities of the panel firms. It is also comforting that the firm's characteristics available in the registers do not play a role in the response patterns. This is confirmed when the R-indicators were calculated separately for the firms which in a given round participated for the first time in the survey – also in this case the unconditional and conditional indicators at the variable level were not statistically different from zero[10].



(a) Variable level: Conditional and unconditional partial indicators



(b) Category level: Unconditional partial indicators with 95% confidence bands

Figure 5: Partial indicators for response in 8th survey round.

5.2. R-indicators during the SAFE fieldwork

The R-indicators can be implemented as a tool for monitoring the representativeness during the data collection. They can be computed for different amount of efforts, e.g. number of attempts, level of interviewer's experience. In SAFE such fieldwork information is limited and we analyse the development of the R-indicators during fieldwork progress.

[10]These results are not presented in the paper but are available upon request.

The SAFE is conducted usually within one month, however, the start and end of the fieldwork can slightly vary by country. To account for these differences, we divide fieldwork into four periods based on the quartiles of the total number of fieldwork days, calculated separately for each country. The results for the 8th round are presented in Table 2.

Table 2: R-indicators for the response and other associated information for each quartile on the fieldwork (8th survey round)

	1st quartile	2nd quartile	3rd quartile	Full fieldwork
Total sample	70,432	70,432	70,432	70,432
R-indicator	0.926	0.877	0.859	0.853
Standard error	0.003	0.004	0.003	0.003
Ave response propensity	0.024	0.052	0.074	0.085
Maximal bias	1.535	1.195	0.953	0.863
Lower bound for R	0.694	0.558	0.476	0.441

For the first fieldwork quartile, which corresponds to approximately the first week of the data collection, the representativity is the highest with R-indicator reaching 0.93. It drops slightly in the second quartile to 0.88 and remains broadly stable till the end of the fieldwork. In this case, the split of the sample into the enterprises which are part of the panel and those participating for the first time plays the major role as indicated by increasing partial R-indicator as the fieldwork progresses (see Table 5). However, a positive impact of each additional week of the fieldwork is visible when looking at maximal bias – it decreases steadily from maximum of 1.54 standard deviation of a survey estimate of interest in the first part of the fieldwork to 0.86 at the end of the fieldwork (see also Figure 3).

6. R-indicators for SAFE data matched with Amadeus database

In this section, first we describe briefly the matching methodology of the SAFE dataset with the Bureau van Dijk's Amadeus database and comment on the quality of the matching. Second, with the dataset, containing both qualitative and quantitative firm-level information, we analyse the R-indicators looking at the availability of the financial information among respondents.

To link the companies from SAFE and Amadeus the information on tax identification number, company name, street, postcode, city and country are used. In the 8th round, 86% of SAFE respondent[11] were successfully matched with Amadeus business register. The quality of matching varies substantially between countries, with success rates over 90% in Belgium, Spain, France and the Netherlands and the lowest in Greece of 67%. There is also a significant difference between the size classes, with the large companies being successfully matched in 98% of cases, whereas the micro firms only in 72%. The difference on the sector level is much less pronounced (see also Bańkowska, Osiewicz and Pérez-Duarte 2014 for more information on matching results).

Being in Amadeus is not enough; a record may have missing financial information. For that reason, we examine separately the representativeness of the SAFE subsamples containing the respondents with the available information on loans, value added and turnover in 8th survey round (in short, "Amadeus sample").

The R-indicators were computed using the same auxiliary variables as in earlier analysis (i.e. country, size, sector and panel dummy), and amount to 0.81 for the value added and are a bit lower for loans and turnover (0.76 and 0.77 respectively; see Table 3). In all three cases, the

[11] As in the previous section, Austria and Germany were excluded from the analysis.

lack of representativity, measured by both partial conditional and unconditional R-indicators, comes from the country variable, similarly to results for the contact (see section 5.1). However, given the smaller sample size, the unconditional partial indicator is statistically significant at 0.1 level only for value added (for turnover p-value equals to 0.12 for country and 0.11 for size variable; see Table 6). Estimated negative values for the category level partial indicators, suggest that the enterprises in the Netherlands and to a lesser extent in Greece are underrepresented in the set of companies with available financial information. Looking at value added and turnover this applies also to Belgium and Ireland. On the other hand, France and Spain are strongly overrepresented with respect to all the three variables considered.

Table 3: R-indicators and other associated information for the availability of information on loans, value added and turnover (8^{th} survey round, respondents)

	Loans	Value added	Turnover
Total sample	6,008	6,008	6,008
R-indicator*	0.759	0.812	0.772
Standard error	0.003	0.002	0.003
Ave propensity	0.594	0.420	0.523
Maximal bias	0.203	0.225	0.218
Lower bound for R	0.018	0.013	0.001

*Due to smaller sample size R-indicator adjusted for bias is used as in de Heij, Schouten & Shlomo (2010).

Similarly to the analysis of the whole SAFE sample with respect to the contact, the size class breakdown also contributes to the loss of representativity in the dataset matched with quantitative financial variables[12]. As expected, micro companies, for which financial information are scarce, are strongly underrepresented also in the matched SAFE subsample (see Table 6). The findings are also reflected in the overall matching rates at the enterprise level, as mentioned above.

It is also worth noting that in the 8^{th} round the maximal bias for the SAFE respondents among the whole sample is higher than for the subsample of the respondents with financial information (0.86 for the SAFE sample in comparison to 0.23 for value added in the Amadeus subsample). However, it should be borne in mind that this is an additional potential bias since the matched SAFE-Amadeus dataset is already a subsample of the SAFE respondents.

7. Conclusions and outlook

In this paper we present R-indicators for SAFE and show that the level of representativity is comparable to other surveys (e.g. see Schouten, Bethlehem, et al. 2012). We found that for the SAFE sample, the country variation contributes mostly to the loss in representativity, while for the Amadeus subsample also size class plays some role with the evident underrepresentation of micro firms.

Based on these findings, we make the following recommendations: i) increase efforts to enhance the quality of the sample contact information, ii) fully harmonise the use of the outcome codes across countries and interviewers, and iii) collect more detailed information from the fieldwork useful for the monitoring of the data collection, i.e. outcome codes for each attempt and possibly interviewers' performance and experience.

Since September 2014 (corresponding to 11^{th} survey round), a new survey company has been in charge of the SAFE fieldwork. Given that this new supplier conducts interviews from one central call centre, as opposed to having local agencies in each region, we will have the

[12]The level of the partial indicators for the size variable are comparable to the partial R-indicators for the contact. However, given the smaller sample size they turn to be statistically not significant at 0.1 level (p-value for value added is 0.15 and for turnover 0.11).

opportunity to disentangle the country variation from the differences in the organisation of local offices. Since the introduction of the online questionnaire in September 2014, it will be important to investigate and monitor the representativity of different survey modes.

This paper could be extended in three directions. First, the representativity of the sample frame can be assessed with respect to the official statistics on the enterprises' population. Second, the sensitivity of the survey results can be tested using different weighting schemes. Finally, as mentioned before, the analysis presented in this paper can be extended using newly available information from the fieldwork and splitting response process into several sub-processes (like contact, cooperation and response) to identify the main causes of potential nonresponse bias.

References

The American Association for Public Opinion Research (2011). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition.* AAPOR.

Bańkowska K, Osiewicz M, Pérez-Duarte S (2014). Linking Qualitative Survey Responses With Quantitative Data. Methodology, Quality and Data Analysis from the Matching of the ECB/EC Survey on Access to Finance of Enterprises and the Amadeus database, URL http://www.bis.org/ifc/events/7ifcconf_bankowska.pdf.

Groves R M (2006). *Nonresponse Rates and Nonresponse Bias in Household Surveys.* Public Opinion Quarterly, Vol. 70, 646-675.

Heij V de, Schouten B, Shlomo N (2010). *RISQ manual, Tools in SAS and R for the Computation of R-Indicators and Partial R-Indicators*, URL http://www.risq-project.eu/publications.html.

Kruskal W, F. Mosteller (1979). *Representative Sampling, III: The Current Statistical Literature.* International Statistical Review 47, 245-265.

Montaquila J M, Olson K M (2012). *Practical Tools for Nonresponse Bias Studies.* URL http://www.amstat.org/sections/srms/webinarfiles/NRBiasWebinarApril2012.pdf.

National Research Council (2013). *Nonresponse in Social Science Surveys: A Research Agenda*, Washington DC: The National Academies Press.

Schouten B, Bethlehem J G, Beullens K, Kleven O, Loosveldt G, Luiten A, Rutar K, Shlomo N & Skinner C (2012). *Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators.* International Statistical Review, Vol. 80, 382-399.

Schouten B, Cobben F & Bethlehem J (2009). *Indicators of Representativeness of Survey Response.* Survey Methodology Vol. 35, 101-113.

Schouten B, Morren M, Bethlehem J, Shlomo N & Skinner C (2009). *How to Use R-Indicators?.* URL http://www.risq-project.eu/publications.html.

Shlomo N & Schouten B (2013). *Theoretical Properties of Partial Indicators for Representative Response.* Technical Report, University of Southampton.

Sudman S (1966). *Probability Sampling with Quotas.* Journal of the American Statistical Association, Vol. 61, No. 315 (Sep 1966), 749-771.

Table 4: Unconditional and conditional partial R-indicators for contact and response in 8 to 10 survey round.

	Unconditional						***Conditional***					
	response			**contact**			**response**			**contact**		
Round	**8**	**9**	**10**	**8**	**9**	**10**	**8**	**9**	**10**	**8**	**9**	**10**
Variable level												
country	0.001	0.001	0.001	0.003***	0.005***	0.013***	0.000	0.000	0.001	0.005***	0.005***	0.012***
size	0.000	0.000	0.000	0.002***	0.002**	0.000	0.000	0.000	0.000	0.003***	0.001	0.001
sector	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
panel	0.005***	0.007***	0.004***	0.001	0.002***	0.004***	0.004***	0.006***	0.004***	0.001	0.002	0.002*
Category level												
BE	-0.006***	-0.010***	-0.002***	0.031***	-0.023***	0.016***	0.005***	0.008***	0.003***	0.037***	0.018***	0.017***
ES	-0.011***	-0.009***	-0.016***	-0.037***	-0.032***	-0.064***	0.007***	0.007***	0.014***	0.040***	0.033***	0.067***
FI	0.001***	0.006***	0.010***	0.011***	0.023***	0.018***	0.002***	0.004***	0.011***	0.013***	0.025***	0.018***
FR	0.006***	0.001	0.007***	-0.012***	0.026***	0.015***	0.007***	0.004***	0.007***	0.023***	0.029***	0.008**
GR	0.003***	0.009***	0.012***	0.021***	0.014***	0.010***	0.001**	0.007***	0.012***	0.021***	0.013***	0.010***
IE	0.014***	0.004***	0.008***	-0.004***	0.000	0.012***	0.006***	0.001	0.003***	0.007***	0.002***	0.008***
IT	0.009***	-0.002***	0.002***	0.006*	-0.026***	0.055***	0.004***	0.003***	0.008***	0.014***	0.023***	0.041***
NL	-0.006***	-0.007***	-0.003***	-0.002	0.040***	0.070***	0.006***	0.008***	0.004***	0.004***	0.043***	0.069***
PT	0.015***	0.022***	0.011***	0.014***	0.012***	0.005	0.008***	0.011***	0.008***	0.012***	0.008***	0.003**
micro	-0.014***	-0.006***	-0.007***	-0.044***	-0.039***	-0.017***	0.012***	0.002***	0.004***	0.050***	0.029***	0.024***
small	0.010***	0.004***	0.004***	0.000	0.000	-0.003**	0.007***	0.003	0.003	0.010***	0.005**	0.005
medium	0.000	-0.001***	0.000	0.009***	0.008***	0.004***	0.000	0.000	0.000	0.010**	0.008	0.005
large	-0.001***	0.000	-0.001***	0.001***	-0.002***	0.002***	0.001	0.001	0.001	0.003	0.002	0.002
industry	0.004***	0.002***	0.003***	0.012***	0.008**	0.011***	0.002	0.003*	0.003	0.004	0.007***	0.007***
construction	-0.002***	-0.008***	-0.006***	-0.014***	-0.018***	-0.016***	0.001	0.006***	0.004***	0.009***	0.014***	0.014***
trade	0.003***	0.001	0.000	0.008***	0.001	-0.001	0.002***	0.001**	0.001	0.008***	0.004***	0.004***
services	-0.004***	0.001	0.000	-0.008***	0.000	-0.003	0.002	0.001	0.001	0.003**	0.001	0.001
non-panel	-0.024***	-0.037***	-0.028***	-0.010***	-0.021***	-0.026***	0.027***	0.035***	0.026***	0.010***	0.018***	0.019***
panel	0.070***	0.077***	0.060***	0.028***	0.043***	0.056***	0.059***	0.073***	0.056***	0.024***	0.041***	0.042***

Note: *** indicates significance at 0.01 level, ** indicates significance at 0.05 level and * indicates significance at 0.1 level.

Table 5: Unconditional and conditional partial R-indicators for response during fieldwork progress in round 8.

	Unconditional				***Conditional***			
	1st quartile	**2nd quartile**	**3rd quartile**	**Full fieldwork**	**1st quartile**	**2nd quartile**	**3rd quartile**	**Full fieldwork**
Variable level								
country	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.000
size	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
sector	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
panel	0.001	0.004**	0.005***	0.005***	0.001	0.003***	0.004***	0.004***
Category level								
BE	-0.002***	-0.003***	-0.005***	-0.006***	0.002***	0.003***	0.005***	0.005***
ES	-0.003***	-0.006***	-0.010***	-0.011***	0.001	0.003***	0.006***	0.007***
FI	-0.001***	0.000**	0.001***	0.001***	0.001*	0.000	0.002***	0.002***
FR	0.001***	0.000	0.003***	0.006***	0.003***	0.001*	0.004***	0.007***
GR	0.001***	0.002***	0.003***	0.003***	0.001**	0.001	0.001*	0.001**
IE	0.004***	0.011***	0.014***	0.014***	0.000	0.006***	0.007***	0.006***
IT	0.000	0.006***	0.007***	0.009***	0.005***	0.002	0.003***	0.004***
NL	-0.002***	-0.005***	-0.006***	-0.006***	0.002**	0.006***	0.007***	0.006***
PT	0.010***	0.012***	0.014***	0.015***	0.009***	0.007***	0.007***	0.008***
micro	-0.002***	-0.007***	-0.012***	-0.014***	0.001	0.005***	0.010***	0.012***
small	0.003***	0.007***	0.010***	0.010***	0.003	0.005**	0.008***	0.007***
medium	0.000***	-0.001***	-0.001***	0.000	0.001	0.001	0.001	0.000
large	0.000***	-0.001***	-0.001***	-0.001***	0.001	0.000	0.001	0.001
industry	0.001**	0.003***	0.004***	0.004***	0.001	0.002	0.002	0.002
construction	0.000	-0.001**	-0.002***	-0.002***	0.001**	0.000	0.001	0.001
trade	0.001**	0.002***	0.003***	0.003***	0.001**	0.001	0.001*	0.002***
services	-0.001***	-0.003***	-0.003***	-0.004***	0.000	0.001	0.001	0.002
non-panel	-0.011***	-0.020***	-0.024***	-0.024***	0.014***	0.023***	0.026***	0.027***
panel	0.032***	0.058***	0.067***	0.070***	0.031***	0.051***	0.057***	0.059***

Note: *** indicates significance at 0.01 level, ** indicates significance at 0.05 level and * indicates significance at 0.1 level.

Table 6: Unconditional and conditional partial R-indicators for SAFE respondents matched with Amadeus database (8^{th} survey round).

	Unconditional			***Conditional***		
	Loans	**Value added**	**Turnover**	**Loans**	**Value added**	**Turnover**
Variable level						
country	0.002	0.004*	0.003	0.001	0.002	0.002*
size	0.001	0.002	0.002	0.001	0.002	0.002
sector	0.000	0.001	0.001	0.000	0.000	0.000
panel	0.000	0.000	0.000	0.000	0.000	0.000
Category level						
BE	0.012***	-0.015***	-0.017***	0.015***	0.010***	0.014***
ES	0.007***	0.018***	0.011***	0.006**	0.016***	0.009***
FI	0.004***	0.000	0.007***	0.004**	0.001	0.007***
FR	0.006**	0.016***	0.029***	0.004***	0.011***	0.028***
GR	-0.007***	-0.013***	-0.005***	0.006	0.009**	0.004
IE	-0.002	-0.017***	-0.022***	0.002***	0.011***	0.016***
IT	0.008**	0.031***	0.020***	0.004***	0.021***	0.012***
NL	-0.038***	-0.031***	-0.033***	0.031***	0.020***	0.025***
PT	0.004*	0.019***	0.011***	0.004***	0.019***	0.012***
micro	-0.036***	-0.044***	-0.045***	0.034***	0.046***	0.042***
small	0.005***	0.002	0.003	0.007***	0.008***	0.007***
medium	0.003***	0.006***	0.006***	0.003	0.004*	0.004
large	0.002***	0.003***	0.003***	0.002	0.002	0.002
industry	0.010***	0.015***	0.015***	0.004***	0.004***	0.006***
construction	0.002	0.002	0.001	0.002**	0.003***	0.001
trade	-0.015***	-0.020***	-0.018***	0.007***	0.006***	0.007***
services	-0.001	-0.002	-0.002	0.002**	0.001*	0.002**
non-panel	-0.002	-0.001	0.002	0.001	0.000	0.001**
panel	0.003	0.001	-0.003	0.001*	0.000	0.001***

Note: *** indicates significance at 0.01 level, ** indicates significance at 0.05 level and * indicates significance at 0.1 level.

Affiliation:

Katarzyna Bańkowska
European Central Bank – Statistics Development/Coordination Division
60640 Frankfurt am Main, Germany
E-mail: katarzyna.bankowska@ecb.europa.eu

Małgorzata Osiewicz
European Central Bank – Statistics Development/Coordination Division
60640 Frankfurt am Main, Germany
E-mail: malgorzata.osiewicz@ecb.europa.eu

Sébastien Pérez-Duarte
European Central Bank – Statistics Development/Coordination Division
60640 Frankfurt am Main, Germany
E-mail: sebastien.perez_duarte@ecb.europa.eu

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 44
April 2015

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2014-09-03
Accepted: 2015-01-08

Austrian Journal of Statistics
April 2015, Volume 44, 31–43.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i2.53

Internet as Data Source in the Istat Survey on ICT in Enterprises

Giulio Barcaroli
Istat
Alessandra Nurra
Istat
Sergio Salamone
Istat
Monica Scannapieco
Istat
Marco Scarnò
Cineca
Donato Summa
Istat

Abstract

The Istat sampling survey on *Information and Communication Technologies (ICT) in enterprises* aims at producing information in particular on the use of Internet and other networks by Italian enterprises for various purposes (e-commerce, e-skills, e-business, social media, e-government, etc.). To such a scope, data are collected by means of the traditional instrument of the questionnaire. Istat began to explore the possibility to use web scraping techniques, associated, in the estimation phase, to text and data mining algorithms, with the aim to replace traditional instruments of data collection and estimation, or to combine them in an integrated approach. The 8,687 websites, indicated by the 19,114 enterprises responding to the survey of year 2013, have been *scraped* and the acquired texts have been processed in order to try to reproduce the same information collected via questionnaire. Preliminary results are encouraging, showing in some cases a satisfactory predictive capability of fitted models (mainly those obtained by using the *Naïve Bayes* algorithm). Also the method known as *Content Analysis* has been applied, and its results compared to those obtained with classical learners. In order to improve the overall performance, different systems for web scraping and mining have been experimented and evaluated. On the basis of the final results of this test, an integrated system harnessing both survey data and data collected from Internet to produce the required estimates will be taken into consideration, based on systematic scraping of the near 100,000 websites related to the whole population of Italian enterprises with 10 persons employed and more, operating in industry and services. This new approach, based on *Internet as Data source (IaD)*, is characterized by advantages and drawbacks that need to be carefully analysed.

Keywords: web scraping, web mining, data mining, text mining, Internet as Data source, Big Data, R.

1. Introduction

Internet can be considered as a data source (belonging to the vast category of Big Data), that may be harnessed in substitution of, or in combination with, data collected by means of the traditional instruments of a statistical survey. In case of substitution, the aim is to reduce

respondent burden; in case of integration the increase in accuracy of the estimates is the main goal. The *Community survey on ICT usage and e-commerce in enterprises* (in short, *ICT in enterprises*) carried out by Istat (together with all EU Statistical Institutes) is a natural candidate to experiment this approach, as the questionnaire contains a number of questions, related to the characteristics of the websites owned or used by the enterprises, whose answers can be deduced directly by the content of these websites. An experiment has been conducted, whose aim is twofold:

- from a technological point of view, to verify the capability to access the websites indicated by the enterprises participating to the sampling survey, and collect all the relevant information;
- from a methodological point of view, to use the information collected from Internet in order to predict the characteristics of the websites not only for surveyed enterprises, but for the whole population of reference, in order to produce estimates with a higher level of accuracy.

Previous work on the usage of Internet as a data source for Official Statistics was carried out by Statistics Netherlands in recent years (ten Bosch and Windmeijer 2014). In particular, a first domain of experimentation was related to *air tickets*: the prices of air tickets were collected daily by Internet robots, developed by Statistics Netherlands supported by two external companies, and the results were stored for several months. The experiment showed that there was a common trend between the ticket prices collected by robots and existing manual collection (Hoekstra, ten Bosch, and Harteveld 2012). Two additional domains of experimentation were *Dutch property market* and *clothes prices*, the first exhibiting more regularity in the sites structure, the latter more challenging with respect to automatic classification due to lack of a standard naming of the items, and variability in the sites organization. The scraping task described in this paper goes a step further with respect to such experiences, by collecting data without any assumption on the structure of the websites and by providing the ability to scale up to a huge number of them.
This paper is organized as follows. In section 2, a general description of the survey is given with a focus on the section of the questionnaire interested to the experiment. In section 3, some different solutions for the web scraping system are described. In section 4, different inference approaches are outlined, together with their results. In the conclusions, pros and cons of the *Internet as Data source* approach are evaluated, and indications about future work are outlined.

2. Description of the survey

The *ICT in enterprises* survey is carried out annually by the Italian National Statistical Institute (Istat)[1], according to a common questionnaire and a harmonised methodology set out by Eurostat, shared in all the EU member states and in cooperation with OECD. The survey collects information on ICT usage by enterprises with 10 and more persons employed working in industry and services[2] and, in particular, involves a sample of small and medium firms and all the large enterprises (with at least 250 persons employed). The survey, on the basis also of a benchmarking framework adopted for the Information Society policy, is annually adapted to the needs of users and policy makers. Moreover, technological evolution

[1]For a complete description, see http://siqual.istat.it/SIQual/visualizza.do?id=5000078

[2]The enterprises are classified in the following economic activity (NACE Rev. 2): Manufacturing; Electricity, gas and steam, water supply, sewerage and waste management; Construction; Wholesale and retail trade repair of motor vehicles and motorcycles; Transportation and storage; Accommodation and food service activities; Information and communication; Real estate activities; Professional, scientific and technical activities; Administrative and support activities; Repair of computers. In 2013 the sample was of 32,328 enterprises and the frame population of 193,130 enterprises. The survey frame is represented by the Italian Business Register of active enterprises (BR). The sampling design is stratified with one-stage selection of units with equal probability. Strata are defined by the combination of economic activities, size classes and administrative regions of the administrative office of enterprises.

requires flexible statistical measurements of the phenomena observed and this survey responds to the need to better tailor some indicators from year to year while keeping the others fixed and more comparable in accordance with the general criteria of reduction or maintenance of response burden on enterprises within a given limit. For ICT survey this limit was fixed to 66 variables per questionnaire and was one of the main reason to begin discussing about the use of Internet as source of data and the possibility to substitute or to complete the information asked through more traditional statistical instruments like self-administered survey. The survey aims at measuring the adoption of ICT, broadband Internet connection, website functionalities, the impact of new technologies on the relationships with customers and suppliers (sharing information electronically on Supply Chain Management, exchanging automatically business documents), on organizational and marketing aspects (sharing electronically information on sales and/or purchases with any internal function, using applications to analyse information collected on clients), e-commerce, e-government. Figures considered in this paper derive from raw survey data of year 2013. In 2013 respondents to the ICT survey were 19,114, equal to 59% of the total initial sample and 9.9% of the universe of Italian active enterprises with 10 and more persons employed. The ICT questionnaire includes a section on access and use of the Internet with a subsection on use of website that is the subject of this paper. We used information coming from questions about facilities supplied through website and those given by respondents in a final section dedicated to enterprises, indicating the website URL. The observed variable (*Does your enterprise have a Website or Home Page, or one or more Internet pages?*) does not refer specifically to the ownership of the website, but to the use of a website by the enterprise to present its activities. The enterprises answering 'yes' to this filter question can include not only the existence of a website which is located on servers belonging to the enterprise, but also third party websites (e.g. one of the group of enterprises to which it belongs, other third party websites[3]). This definition represents a first possible limit of this experimental study, as we will discuss later. For the enterprises having a website, this question focuses on the measurement of its specific uses. In particular we concentrated our study on the possibility to sell product or services via web (*Online ordering or reservation or booking, shopping cart facility*, from now on *Web sales functionality*)[4]. This choice was due to particular potential importance of this facility for positive impact on enterprise's performance, for giving a measure of level of e-business readiness and intensity of firms and sectors and also because, in terms of web contents, should be easier to recognize keywords that could detect the same phenomenon through automated tools.

3. The web scraping system

Web scraping is the process of automatically collecting information from the World Wide Web, based on tools (called scapers, internet robots, bots etc.) that navigate and extract the content of a website, and store scraped data in local data bases. Web scraping may be against the terms of use of some websites: courts are prepared to protect proprietary content of commercial sites from undesirable uses, even though the degree of protection for such content is not clearly settled. The amount of information accessed and copied depends on the degree to which the access is perceived as adversely affecting the site owner's system, and the types and manner of restrictions to such access.

[3]An enterprise may offer web sales functionality and still not have a website as the sales are through e-marketplaces that are not included in questions considered in this paper.

[4]From the Methodological Manual (2013): 'This item refers to a facility which allows the user to order products or services with no additional contact offline or via e-mail necessary (for the ordering). It includes also websites which allow the reservation of hotel rooms or the booking of flights. It does not include a link in the website which directs the user to an e-mail application which requires the user to send the order via e-mail. Payment may or may not be included in the ordering facility, e.g. payment may be made on reception of the product also by other means than electronic payment'. Inside the benchmarking framework 2011-2015 it is included the indicator D7 asking for percentage of enterprises having a website with web sales facilities (Website or a Home Page with online ordering or reservation or booking).

In the following, different solutions for the web scraping are described: the first one is already available and has been used specifically for this experiment, while the others are still under investigation. Indeed, we are carrying out a dedicated activity with the purpose of testing and comparing different technological solutions for scraping in order to figure out the most suitable solution for a specific purpose.

3.1. The web scraping application based on JSOUP and ADaMSoft

A first choice was to develop the scraping application by referring to an open source library called JSOUP (available at `http://jsoup.org/`) and by integrating it in the ADaMSoft system (available at `http://adamsoft.sourceforge.net/`). JSOUP is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM (Document Object Model), CSS (Cascading Style Sheets), and jQuery-like methods. More in detail this tool allows:

- scraping and parsing HTML from a URL, file, or string;
- findind and extracting data, using DOM traversal or CSS selectors;
- manipulating the HTML elements, attributes, and text.

On the other side, ADaMSoft was selected because it included facilities in managing huge data sets and because it already contains procedures to deal with textual information. We then developed an appropriate step that can be considered a mix between a web spider (i.e. a tool to extract the structure of a website) and a retriever for both the content and the tags of identified pages. More specifically, the content is intended as the text that can be viewed when browsing a page, while the tags are all the hidden HTML keywords that guide the way in which the page is displayed, the actions associated to a button, etc.

Inside such step we implemented methods that permit to:

- keep into account the limitations eventually defined in the `robots.txt` file;
- extract the structure of the website for a given level of depth (i.e. the sub-links from the main URL);
- filter the resultant URLs (in order to avoid, for example, those links that redirect to other websites);
- emulate a specified user agent;
- pass to the website a series of cookies;
- specify the method (GET or POST);
- use a given time limit to explore the website;
- identify and access different content types (HTML, obviously, but also PDF, DOC, etc.).

The input to the procedure is a dataset containing the identifiers of the enterprises, the indication of related URLs and also the indication of the level of depth that is considered as acceptable in the exploration of the websites (i.e. to what extent sub-links will be taken into consideration: in our tests we set this parameter to '2' and '3'). To increase the efficiency of the task, the procedure permits to examine at the same time more than one website.

We considered really crucial to retrieve also the elementary information of a HTML tag (i.e. its type, name and content), because it could contain discriminant terms that can help us in identifying the nature of the website; for example a button associated to an image called "paypal.jpg" could be a clear sign of web sales functionality.

Running the Web Scraper procedure for the original 8,687 websites took less than one day on a Windows PC platform. Actually we observed some difficulties in accessing some websites; these derive from a not correct specification of the main URL, and/or making use of technologies not entirely based on standard HTML text (like, for example, the websites realized with Flash technology).

By considering only those websites for which at least a page was accessed, we found an average value of 235,108 characters retrieved.

For what concerns the HTML tags (that we restricted to one of the following types: *address, button, fb:like, form, label, menu, input, meta, option, rss, select, textarea*), we collected more than 17.5 million of elementary information contained in tags (which corresponds to an average of 2,649 tags collected for each one of the 6,632 websites for which at least one tag was retrieved).

Due to the huge amount of terms, we proceeded by tokenizing each of these by transforming all non valid ASCII code characters in spaces (i.e. `paypal.jpg` is transformed in two terms: `paypal` and `jpg`) and we considered their main lemma, by deleting all the determiners, the articles, the prepositions, etc. (i.e. all the terms that can be considered generic). To this purpose, we used a package named TreeTagger, directly executed from inside Adamsoft.

TreeTagger is a tool for annotating text with part-of-speech and lemma information, developed at the Institute for Computational Linguistics of the University of Stuttgart (`http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`). It has been used by referring to the Italian and to the English lexicon in order to permit the selection of less than 60,000 terms to be considered as the basis for further processing steps.

3.2. Other solutions for web scraping: Nutch/Solr suite and HTTrack

The Apache suite used for crawling, content extraction, indexing and searching results is composed by Nutch and Solr. Nutch (available at `https://nutch.apache.org/`) is a highly extensible and scalable open source web crawler, it facilitates parsing, indexing, creating a search engine, customizing search according to needs, scalability, robustness, and scoring filter for custom implementations. Built on top of Apache Lucene and based on Apache Hadoop, Nutch can be deployed on a single machine as well as on a cluster, if large scale web crawling is required. Apache Solr (available at `https://lucene.apache.org/solr/`) is an open source enterprise search platform that is built on top of Apache Lucene. It can be used for searching any type of data; in this context, however, it is specifically used to search web pages. Its major features include full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document handling. Providing distributed search and index replication, Solr is highly scalable. Both Nutch and Solr have an extensive plugin architecture useful when advanced customization is required.

Starting from a list of URLs (root pages), Nutch fetches, parses and indexes for each of them all the linked resources according to a series of constraints, the most important are (i) the link depth from the root page that should be crawled, (ii) the maximum number of pages that will be retrieved at each level up to the depth. Nutch offers a series of fine configurable NLP (Natural Language Processing) functions applicable on fetched web resources, such as tokenization, stop-words removal and stemming. Finally Nutch delegates searching to Solr.

All problems encountered are relative to environment configuration and tools integration, for example it was necessary to manage case sensitivity, site load balancing, page redirections, plugins and OS configuration, etc..

Although this web scraping approach requires an initial effort in terms of technological expertise, in the long run it can lead to a substantial return on investment as it can be used on many other contexts to access Big Data sources. As an example, it can be used as a platform to access and analyse web resources like blogs or social media to perform semantic analyses or sentiment analyses tasks.

HTTrack (available at `http://www.httrack.com/`) is a software tool that permits to "mirror" locally a web site, by downloading each page that compose its structure. For the specific case study described in this paper, we access URLs of enterprises websites and download related HTML resources. Once such resources are locally available, we are able to access specific the content of HTML elements (e.g. title, HTML links, body, etc.) that are used for subsequent analysis steps.

The main differences between HTTrack and Nutch/Solr are:

- generic parsing and indexing tasks are only performed by Nutch/Solr;
- direct processing of HTML resources is easily available with HTTrack, while it requires dedicated effort with Nutch/Solr/Lucene.

According to our first tests, HTTrack approach results to be more suitable for the scraping task of the case study here described, as it requires access to specific HTML pieces.

4. The inference system

Once completed the web scraping activities, before proceeding with the inference phase a pre-processing step was applied, consisting in treating the text terms (reduction to lower case, elimination of punctuation and stop-words, stemming) and in selecting only the words that showed a significant influence on the target variables. This influence was determined by a two-step procedure:

- a first selection was made by applying a *correspondence analysis* bewteen a given target variable and the words contained in the scraped texts;
- a second selection was obtained by evaluating the chi-square associated to the cross-classification of a given target variable with respect to the presence/absence of a given word.

By applying the first step a subset of words can be selected, still too numerous to be managed in the modelling phase. This is why we applied the second step, in which four different subsets of words have been defined: having set as thresholds the percentiles 99.5, 99.0, 97.5, 95.0 related to their chi-square distributions, only words with a chi-square exceeding those thresholds have been considered.

The final result of this pre-processing consists in a document/term matrix, where each row represents a website, each column is referred to an influent word, and the intersection indicates the frequency (or just the presence or the absence) of the word in the website. In order to choose the best instruments useful to build the inference system, in this exploratory phase we tested several of them, distinguished in:

- data mining learners, applicable to this text mining problem: *Classification Trees*, ensemble learners (*Random Forest, Adaptive Boosting, Bootstrap Aggregating*), *Neural Networks, Maximum Entropy, Support Vector Machines, Latent Dirichlet Allocation* (James, Witten, Hastie, and Tibshirani 2013);
- the approach followed in the *Content Analysis* (Hopkins and King 2010);
- the learner most suitable for text mining: *Naïve Bayes* (Lantz 2013).

As usual, available data have been partitioned in a training set and in a test set: each model, fitted using the training set, has been applied to the test set in order to evaluate its performance, by comparing observed and predicted values for the target variables, both at individual and aggregate level. In general, the proportion between the two sets was determined in 75/25, but a sensitivity analysis has been performed for Naïve Bayes and Content Analysis defining 9 different rates for the training set (from 0.1 to 0.9). Experiments have been carried out considering the four different subsets of words defined accordingly to their chi-square, and the most favorable in terms of performance has been retained.

Performance has been measured by considering the following indicators: (i) *precision rate* (number of correctly classified cases on the total number of cases), (ii) *sensitivity* (rate of correctly classified positive cases), (iii) *specificity* (rate of correctly classified negative cases). In addition, we also introduced (iv) the *proportion of predicted positive cases*, as it corresponds to the final estimate that we want to produce, and whose accuracy we want to maximize. These four indicators can be easily computed from the *confusion matrix*.

4.1. Data mining learners

In Table 1 we report the results of the application of the different learners in order to predict web sales functionality (we made use of R packages **RTextTools** (Jurka, Collingwood, Boydstun, Grossman, and Atteveldt 2014) and **rattle** (Williams 2011)) (R Core Team 2014). It is possible to notice that the precision level is in general acceptable: it ranges from a minimum of 79% to a maximum of 85%. Specificity is always very high. The real problem is given by sensitivity, that is the capability to correctly classify positive cases, i.e. the websites that offer web sales functionality: in many cases its value is too low to be considered as acceptable. As for the proportion of web sales functionality, in general data mining learners fail in reproducing the correct aggregates.

Table 1: Performance indicators for data mining learners (variable *Web sales functionality*).

METHOD	Precision	Sensitivity	Specificity	Proportion Web sales = Yes (observed)	Proportion Web sales = Yes (predicted)
Classification tree	0.83	0.28	0.98	0.21	0.08
Random forest	0.85	0.34	0.99	0.22	0.08
Bootstrap aggregation	0.82	0.48	0.91	0.21	0.10
Adaptive boosting	0.80	0.39	0.91	0.22	0.17
Maximum entropy	0.80	0.46	0.90	0.22	0.18
Support Vector Machines	0.79	0.02	0.99	0.22	0.01
Neural networks	0.82	0.21	0.98	0.20	0.06
Latent Dirichlet Allocation	0.81	0.18	0.98	0.21	0.05

4.2. Text mining specific approaches

Content analysis

Hopkins and King (2010) proposed a method quite different from all the others so far considered, as it does not require statistical or machine learning modeling of data and consequent individual predictions. It does not even require a training set to be a representative sample of the whole population: the only requirement is that the training set must contain a sufficient number of cases for each combination of terms.

In order to verify the robustness of this method, different training sets have been obtained by drawing samples from the available websites, varying the sampling rate from 0.1 to 0.9 (100 samples for each sampling rate), and related estimates of *Web sales functionality* rate have been produced for each sample by using Content Analysis. The software used for its application is described in Hopkins, King, Knowles, and Melendez (2012), and is available at `http://gking.harvard.edu/readme`.

It can be seen (Figure 1) that, especially in cases from 0.1 to 0.3 of training set rate, the method seems to be unbiased, as the mean of the estimates tends to coincide with the proportion calculated in the total number of cases. But the range of the estimates is considerably large: for example, in the case of 0.1, interval of estimates goes from 0.08 to 0.31, and we can observe even worse situations for the other training set rates.

Naïve Bayes algorithm

The Naïve Bayes algorithm is the most used in the field of the text classification, where it can be considered as a standard choice. It is called "naïve" because of its (simplistic) assumptions concerning data, as it assumes that all the features in a dataset are independent and equally important, a condition that is seldom verified in real situations. Actually, words in a text are not equally important in order to predict a given category to be associated to the text, and words are not independent each other. But Naïve Bayes works well despite the fact that its basic assumptions are very seldom fulfilled. We made use of the implementation available in the R package **e1071** (Meyer, Dimitriadou, Hornik, Weingessel, and Leisch 2014). In Table 2 the results obtained by the application of Naïve Bayes are reported.

Table 2: Confusion matrix for Naïve Bayes application (variable *Web sales functionality*).

	Predicted Values			
Observed Values	**1 (YES)**	**2 (NO)**	**Total**	**Relative Frequencies**
1 (YES)	120	119	239	0.22
2 (NO)	121	748	869	0.78
Total	241	867	1,108	1.00
Relative Frequencies	0.22	0.78	1.00	

From this confusion matrix it is possible to calculate the usual performance indicators (Table 3). It can be seen that Naïve Bayes is slightly inferior to some data mining learners in terms of precision, but performs better in terms of sensitivity, and reaches a practically perfect coincidence between the predicted proportion and the observed one.

Table 3: Values of performance indicators for Naïve Bayes application (variable *Web sales functionality*).

Indicator Name	**Indicator Value**
Precision	0.78
Sensitivity	0.50
Specificity	0.86

As in the case of Content Analysis, also for evaluating the robustness of Naïve Bayes solutions a simulation has been carried out, under the same setting.

The graph in Figure 2 shows that the method is slightly biased[5], as it systematically overestimates the true value (in the order of one or two percentage points). But the variability of the estimates is much lower than in the case of the Content Analysis: considering the case related to the training set rate equal to 0.1, the range goes from 0.19 to 0.24.

[5]In presence of bias, a method to correct the aggregations resulting from individual predictions obtained by a given learner has been proposed by Hopkins and King (2010). Given a variable D with two possible values (1 and 2), we know that

$$P(\hat{D} = 1|D = 1) : sensitivity \tag{1}$$

$$P(\hat{D} = 2|D = 2) : specificity \tag{2}$$

Then, by the law of total probability:

$$P(\hat{D} = 1) = (sensitivity)P(D = 1) + (1 - specificity)P(D = 2) \tag{3}$$

we can obtain:

$$P(D = 1) = \frac{P(\hat{D} = 1) - (1 - specificity)}{sensitivity - (1 - specificity)} \tag{4}$$

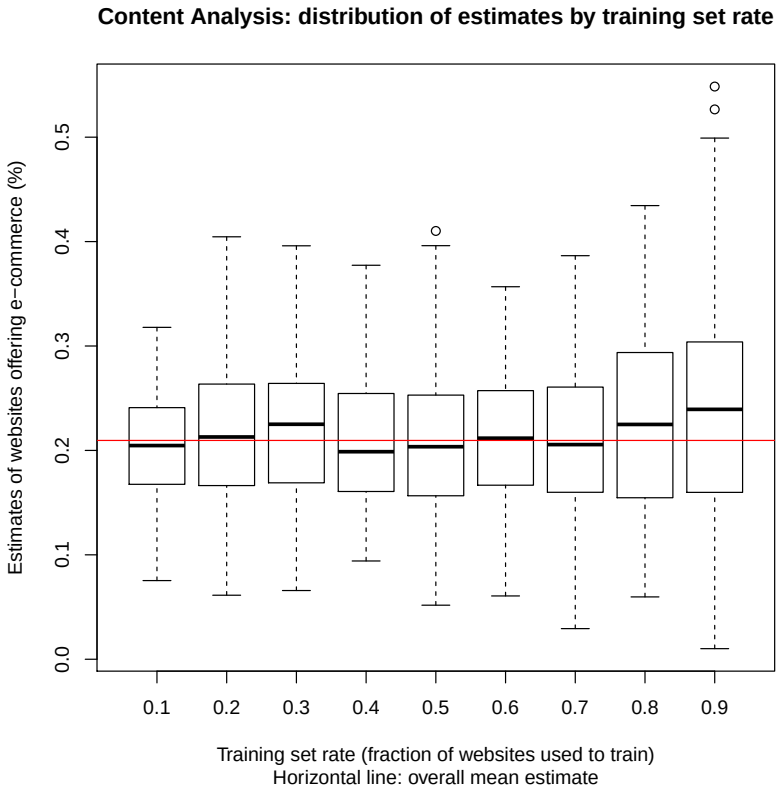


Figure 1: Content Analysis: distributions of estimates calculated on test sets varying the training set rate (variable *Web sales functionality*)

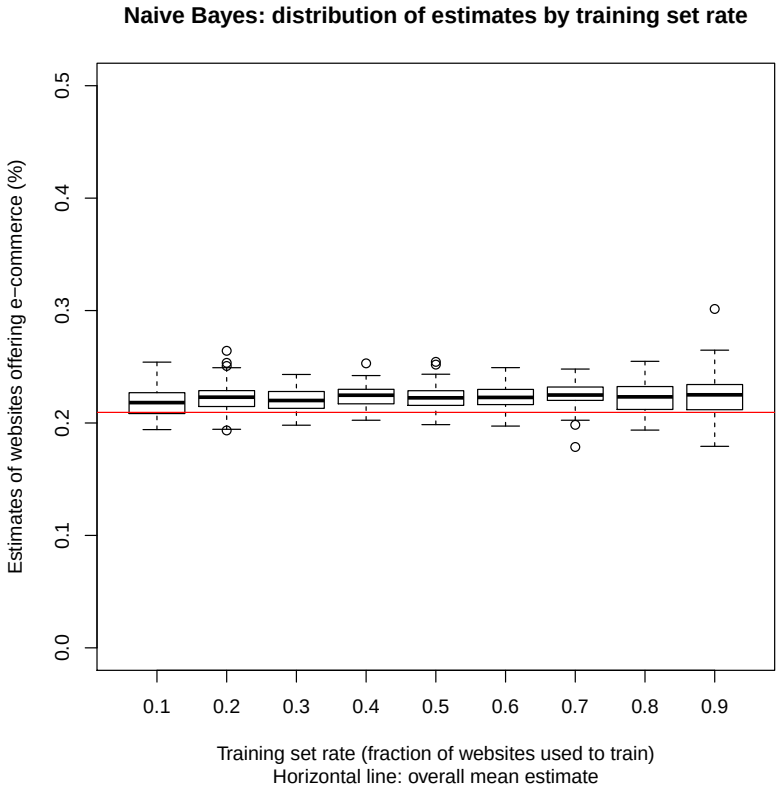


Figure 2: Naïve Bayes: distributions of estimates calculated on test sets varying the training set rate (variable *Web sales functionality*)

As it resulted to be the best method among those considered, Naïve Bayes has been applied to other suitable variables in the questionnaire, obtaining the results reported in Table 4.

As for the software that has been used to produce the results reported above, the ADaMSoft modules for web scraping and texts handling, together with the R scripts for the application of Naïve Bayes algorithm, are available at http://adamsoft.sourceforge.net/appscripts.html.

Table 4: Results of the application of Naïve Bayes to the complete set of B8 question.

QUESTION	Precision	Sensitivity	Specificity	Proportion Web sales = Yes (observed)	Proportion Web sales = Yes (predicted)
Web sales functionality	0.78	0.50	0.86	0.21	0.21
Orders tracking	0.82	0.49	0.85	0.18	0.11
Description and price list of goods	0.62	0.44	0.79	0.48	0.32
Personalised content for regular visitors	0.74	0.41	0.781	0.09	0.23
Possibility to customise online goods	0.86	0.53	0.87	0.05	0.14
Privacy policy statement	0.59	0.57	0.64	0.68	0.51
Online job application	0.69	0.521	0.78	0.35	0.33

5. Conclusions

The best method resulting from the experiment seems to be the Naïve Bayes. The values of the first three indicators of performance (precision, sensitivity and specificity) are all good, and the forth (alignment between observed and predicted aggregate) is the best with respect to the other learners. It is slightly biased with respect to the Content Analysis, but is much better in terms of variability of the estimates.

With regard to the relatively low levels of sensitivity, due to the high number of false negatives (represented by enterprises declaring in the survey to have web ordering facility but resulting as *not having* this possibility on the site-centric scraping basis), it is important to underline that the use of *website centric measurements* allows only a partial measurement of the phenomenon detected by the survey. In fact, in the questionnaire the wording of the questions permits to the respondent to answer "yes" with reference not only to the owned website but also to those sites of the linked companies (subsidiaries or owning the brand, or other third parties). Moreover, the positive answers in the survey consider also e-sales between enterprises: commercial transactions between the responding enterprise and other enterprises, named business-to-business (B2B, e.g. manufacturer and a wholesaler, a wholesaler and a retailer). With respect to business-to-consumer (B2C) e-sales or reservation systems, B2B is often based on a protected access requiring a login and a password, making difficult to identify automatically e-sales functionalities of investigated websites. These two factors can explain why using different instruments (survey vs scraping) we measure the same phenomenon but delimited by different boundaries.

The extension of the *IaD* methods to further technical indicators (i.e. number of pages, downloading speed, technical or language accessibility, etc.) requires to consider also other issues. In the following, we report the main trade-offs, strengths and weaknesses of web scraping and mining methods presented above, compared with the traditional statistical survey.

Benefits and opportunities

- in terms of accuracy: it is possible to extend the analysis to the whole population and not only a subsample (avoiding sampling errors), therefore producing more detailed figures (e.g. for enterprises with less than 10 persons employed not observed in ICT survey); degree of closeness of estimates to the true values could be improved thanks to technology and programming new code (reducing measurement errors);
- in terms of relevance of information: it is possible to discover new services, new information; to investigate other web functionalities as e.g. advertisement of open job positions or online job application, usage of website safety certificate, possibility for customer to submit electronic complaints (via e-mail, web form, etc.), links or references to the enterprise's social media profiles, etc.;
- in terms of comparability among countries: it could be improved if same automatic website centric tools are used;
- in terms of transparency of process: it is avoided human misunderstanding among concept/definition and scope of the question of survey;
- in terms of statistical burden: the respondent burden can be reduced (but we discussed about only one variable out of 66);
- in terms of timeliness: it is improved;
- in terms of reiteration of process: it is possible to repeat the entire automatic data collection during the same period of traditional survey.

Costs and disadvantages

- in terms of accuracy: it is necessary to manage and maintain a list of URLs for the entire population; there is a non-negligible risk to introduce bias into the estimates;
- in terms of coherence of measured concepts: web mining applications described may not catch the same phenomenon of ICT survey;
- in terms of comparability among countries: using different tools (survey vs IaD) or a different list of words could produce less comparability;
- in terms of technology used: there are technical limits to solve as the long run time necessary for the crawler to get the entire content; security barriers inside the website preventing automatic access (restrictions); website not in HTML (i.e. in Flash), redirect problems;
- semantic limits of automatic tools: not all services offered on websites can be well semantically delimited;
- time spent in analysis and programming: to discover new information requires to analyse data collection in different ways and then to update program code;
- in terms of development and maintenance efforts of the web mining applications: persons with high level skill are required.

The web mining (or Internet as Data source) approach experimented in the *ICT in enterprises* survey revealed to be promising and can be continued and extended in different directions:

- with reference to the population of interest: we can consider the URLs of all the units belonging to the Business Register, and perform a mass scraping of related websites (in this case also experimenting more properly the high volume problems (scaling) related to Big Data), considering the whole survey sample as a training set, so to obtain a model that can be applied the whole population. The aim is twofold: (i) to produce estimates under a full predictive approach, reducing the sampling errors at the cost of introducing additional bias (both components of Mean Squared Error should be evaluated); (ii) to identify the subpopulation of enterprises active in web sales transactions with individuals as the end consumer (B2C), that can be considered as a new sampling frame to consider in the ICT survey or useful to carry out new *ad hoc* surveys;

- with reference to the content of the questionnaire: the approach used with the set of variables contained in the 'B8' section of the questionnaire will be evaluated also with regard to other suitable sets of variables in the questionnaire (e-recruitment, use of social networks, etc.).

Anyway, it is necessary to improve the results of the web mining applications by investigating specific situations. While conceptual reasons justify, as discussed above, the high percentage of false negatives, it is more difficult to understand cases in which web scraping finds web sales functionality signals in websites contrary to answers of the survey (false positives). In the future it is necessary to better explore these false positives because, for example, they could be a signal that respondents do not understand correctly the question. Different explanations could be found in time lag between survey and web scraping, or in flaws in the methods and tools used for web scraping and text mining.

References

Hoekstra R, ten Bosch O, Harteveld F (2012). "Automated Data Collection from Web Sources for Official Statistics: First Experiences." *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, **28**(3-4), 99–111.

Hopkins D, King G (2010). "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science*, **54**(1), 229–247.

Hopkins D, King G, Knowles M, Melendez S (2012). *ReadMe: Software for Automated Content Analysis*. Version 0.99835, URL http://gking.harvard.edu/files/gking/files/readme.pdf.

James G, Witten D, Hastie T, Tibshirani R (2013). *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics.

Jurka T, Collingwood L, Boydstun A, Grossman E, Atteveldt vM (2014). *RTextTools: AutomaticText Classification via Supervised Learning*. R package version 1.4.2., URL http://CRAN.R-project.org/package=RTextTools.

Lantz B (2013). *Machine Learning with R*. Packt Publishing Ltd.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2014). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-3, URL http://CRAN.R-project.org/package=e1071.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

ten Bosch O, Windmeijer D (2014). "On the Use of Internet Robots for Official Statistics." In *MSIS-2014*. URL http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2014/Topic_3_NL.pdf.

Williams G (2011). *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!)*. Springer.

Affiliation:

Giulio Barcaroli
Istituto Nazionale di Statistica (Istat)
Via Cesare Balbo 16
00184 Roma, Italy
E-mail: giulio.barcaroli@istat.it

Alessandra Nurra
Istituto Nazionale di Statistica (Istat)
Via Tuscolana 1788
00173 Roma, Italy
E-mail: alessandra.nurra@istat.it

Sergio Salamone
Istituto Nazionale di Statistica (Istat)
Via Tuscolana 1788
00173 Roma, Italy
E-mail: sergio.salamone@istat.it

Monica Scannapieco
Istituto Nazionale di Statistica (Istat)
Via Cesare Balbo 16
00184 Roma, Italy
E-mail: monica.scannapieco@istat.it

Marco Scarnò
Cineca
Via dei Tizi 6/B
00185 Roma, Italy
E-mail: m.scarno@cineca.it

Donato Summa
Istituto Nazionale di Statistica (Istat)
Via Cesare Balbo 16
00184 Roma, Italy
E-mail: donato.summa@istat.it

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 44
April 2015

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2014-07-07
Accepted: 2014-10-31

Austrian Journal of Statistics
April 2015, Volume 44, 45–57.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i2.79

On the Representativeness of Internet Data Sources for the Real Estate Market in Poland

Maciej Beręsewicz
Poznan University of Economics

Abstract

Shifting paradigms in Official Statistics lead to the widespread use of administrative records in an effort to support or create an alternative for censuses and surveys. At the same time the demand for diversified detailed information is increasing. In order to meet this demand Official Statistics needs to seek new data sources. Internet data sources (IDS), or more generally, big data could be one of them. The potential usefulness of these new sources of statistical information should not be neglected.

The aim of the paper is to report on a study intended to assess the representativeness of IDS for the real estate market in Poland. These sources could be used for describing the demand and supply on the secondary real estate market in a more detailed way than is possible with the existing methodology. The degree of representativeness is assessed on the basis of information from official surveys and other data sources. Due to the shortage of relevant literature on the subject, the article provides a definition of IDS and draws on insights from a study conducted by the author to enhance information from Official Statistics. The study involved using information on street names from the National Official Register of the Territorial Division of the Country (TERYT) to harmonize street names obtained from IDS. A special program for automated data collection (*web spider*) was developed. All the calculations were made with R (R Core Team 2014) statistical software and additional R packages (**XML**, **RCurl**, **httr** and **ggplot2**).

Keywords: Big data, Internet data sources, secondary real estate market, web scraping, R.

1. Introduction

Increasing information needs at a low level of aggregation not only encourage the development of small area estimation but also stimulate the search for new data sources that could support or enhance existing sources (reporting, censuses or surveys). This process has been continuing since 1970s when statisticians and National Statistical Institutes (NSIs) started using and adopting administrative records into their statistical systems (Wallgren and Wallgren 2014). However, the statistical theory underlying the use of administrative registers is currently the subject of research and development (Zhang 2011, 2012). Nonetheless, the process has brought about a change in thinking about statistical data sources. In the literature and during statistical conferences this process is often described as a change of paradigm in Official Statistics, which involves the adoption of existing data sources instead of creating new ones.

Although administrative records provide unit-level data, their scope is usually limited to a specific field that was crucial to the register's administrator. Initially, registers were not created for statistical purposes, which means that these sources need to be transformed to become a statistical data source. In addition, it is assumed that registers cover the whole target population, which is not always the case (see Gołata 2014; Zhang 2015). However, in the environment of electronic economy, characterized by the increasing use of the Internet (both by households and companies) and the Internet of Things (e.g. mobile technologies), administrative registers as well as surveys tend to lag behind the changing setting. Therefore, information gaps in certain fields are growing and new data sources should be examined to improve information coverage.

In this context the term *big data* has gained wide recognition as a potential source of statistical information, although it does not have a clear definition. In an information system, it refers to data that cannot easily be handled within the existing infrastructure. From the statistical point of view, it is considered as a potential source for describing ongoing changes in society. The following sources are discussed in the context of Official Statistics: mobile networks (e.g. to track movement, travel routes), social networking sites (e.g. Facebook, Twitter, Linkedin), e-commerce (e.g. eBay, Amazon, price comparison services) or Google search trends. However, they are not being investigated widely as a statistical data source or from the point of view of estimation theory. The purpose of this paper is to bridge this gap by discussing the representativeness issue in the context of new data sources, specifically concentrating on IDS.

The paper has the following structure. The second section defines and presents IDS in the context of survey methodology. The key concept – representativeness – is defined and discussed in the context of new data sources in the Internet. Relation to the characteristics of big data is underlined in the light of statistical data sources. The third section is devoted to data sources for the real estate market in Poland and possibilities of using IDS to obtain statistical information. The penultimate section contains an empirical evaluation of representativeness using data from the Polish real estate web portal – `http://www.nieruchomosci-online.pl`. The data were collected from the web portal automatically by means of a special R (R Core Team 2014) program. The article ends with the discussion of results and final remarks.

2. Internet data sources

While Internet access in households is increasing (Mohorko, Leeuw, and Hox 2013), the way people or companies communicate is changing as well (e.g. Customer to Customer C2C, Business to Customer B2C, Business to Business B2B). This process opens new opportunities for statisticians to track and measure economy and society. For example, it is possible to use web services to assess auctions (e.g. e-Bay), compare prices on the Internet with off-line prices using e-commerce services (e.g. Amazon) or access hard-to-reach populations. In the literature we can find evidence of using data scraped from web pages to measure inflation, predict unemployment or flu risk. For instance, *the Billion Price Project*[1] conducted by the Massachusetts Institute of Technology (MIT) web-scrape data from over 60 countries and calculates price indexes and measures of macroeconomic phenomena (Cavallo 2012, 2013). However, the exact methodology and web scraping technique is protected by PriceStats[2], a start-up based in Cambridge MA. Another well known project that has highlighted the potential usefulness of the Internet is Google Flu Trends, which was widely discussed a few years ago (Ginsberg, Mohebbi, Patel, Brammer, Smolinski, and Brilliant 2008).

Nonetheless, new data sources have not been discussed widely in the statistical literature. The first reference to the statistical aspect known to the author is mentioned in Shmueli, Jank, and Bapna (2005) and is devoted to on-line auction research. Bapna, Goes, Gopal, and Marsden (2006) discusses the problem of data-driven research in e-commerce studies. However, in

[1] `http://bpp.mit.edu`
[2] `http://www.pricestats.com`

recent years new research addressing big data and IDS in the context of statistical data source has been growing. Below are the main topics and selected literature:

- Predicting unemployment - Fondeur and Karamé (2013), Xu, Li, Cheng, and Zheng (2012);
- Source of information for small area estimation - Pratesi, Pedreschi, Giannotti, Marchetti, Salvati, and Maggino (2013), Pratesi, Giannotti, Giusti, Marchetti, Pedreschi, and Salvati (2014), Porter, Holan, Wikle, and Cressie (2013)
- Opinions / Sentiment analysis - Daas, Roos, van de Ven, and Neroni (2012); Daas and Puts (2014b), Miller (2011)
- Indexes - Vosen and Schmidt (2011)
- Representativeness and quality - Buelens, Daas, Burger, Puts, and van den Brakel (2014), Daas and Puts (2014b)
- General on new data sources - Choi and Varian (2012), Daas, Roos, de Blois, Hoekstra, ten Bosch, and Ma (2011); Daas and Puts (2014a), Hoekstra, ten Bosch, and Harteveld (2012)

However, in order to assess the issue of representativeness of this new data source it is important to define precisely what kind of data sources are involved and compare them with the existing ones. First of all, it should be emphasized that IDS are not defined in a statistical system or in the literature. For example, according to the The United Nations Economic Commission for Europe (UNECE) IDS can be a part of administrative sources, which are defined as *data collected by sources external to statistical offices*. On the other hand, the recent project *Big Data for Official Statistics*[3], led by UNECE, classified IDS as one type of big data sources. The project divides big data into three groups: Social Networks (human-sourced information), Traditional Business systems (process-mediated data) and the Internet of Things (machine-generated data). According to this division, IDS could be classified into the first two classes as Social Networks, Internet searches or E-commerce.

Big data is not a statistical term, but more of a general description used to capture certain characteristics of data sources. There is no specific time when the term was introduced but references can be found in Bayer (2011) and Bayer and Laney (2012). The definition consists of three aspects - high volume, high velocity and high variety (often described as 3V). The first V refers to the amount of data counted in tera- and petabytes, which are hard to analyse within the existing infrastructure. The second V denotes how these data are generated and change in time (e.g. web-logs, photo uploads). The last V indicates that big data occur in different formats, such as photos, texts, logs, videos etc. In comparison to classical data sources, like censuses or surveys, big data processing requires more effort in order to extract meaningful information. Some types of big data can occur in administrative sources, i.e. traffic sensor data, patients registers, land photos or car registers. However, in most cases such data are generated by users of specific types of portals (e.g. social networks) or services (e.g. mobile apps). That is why it is important to consider big data as a potential source of information about people or business activity.

For the purpose of this study, IDS are defined as *data collected and maintained by units external to statistical offices and administrative regulations available (mainly) on the Internet (through web-based databases)*. The definition contains two main aspects – first it explicitly states that data are collected by units other than official institutions and the purpose is not defined by official regulations. This element is crucial since the majority of data sources on the Internet are created by private companies. The definition also excludes official web pages that contain reports or statistics resulting from surveys or the use of registers (e.g. Eurostat Database, STATcube). The second part states that these data sources are available on the Internet through queries (e.g. via web-forms). Such portals could be devoted to price comparison, e-commerce, portals that include unit-data (e.g. offers on real estate market) or reports and aggregated data (e.g. Google Trends).

[3] http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics

IDS and big data have recently been under evaluation by NSIs for the production of statistics and enhancement or replacement of the existing data sources or data collection techniques. NSIs working papers address different aspects connected with the use of such data, for instance privacy or legality. These issues are outside the scope of this study and therefore will not be discussed in this paper. However, before new data sources can be used for statistics, they should meet the criteria that are applied to classical data sources. The following aspects should be discussed in the future: *conceptualisation, representativeness, selectivity, nonsampling errors, measurement of uncertainty, sampling, estimation (e.g. model-based estimation, Bayesian approach) or the place in statistical information system.*

The literature provides many definitions of representativeness, but none is given explicitly. Kruskal and Mosteller (1979a,b,c) provide a comprehensive literature review and list nine definitions of representativeness that refer to the following aspects: a general opinion about data, the lack of selective forces, the scaled-down version of the population, typical/ideal cases, whether it reflects variability of the population, how it refers to specific sampling methods (equality of probability of inclusion), whether it provides good estimation, whether it fits specific purposes. Most of the definitions in the statistical literature refer to respondents (people or companies) (see Schouten, Cobben, and Bethlehem 2009) and the suggestion of using propensity weighting. Bethlehem (2009) defines representativeness with respect to the sample when relative distributions are the same in the sample and in the population. It means that the sample is representative when characteristics of the sample and the population are the same. Following Kruskal and Mosteller (1979a,b,c) this statement can be understood to mean that a representative sample is the same as a scaled-down population. The measurement of representativeness of Internet research mainly refers to online surveys, online panels and pop-ups (Bethlehem 2008; Bethlehem and Biffignandi 2011). Buelens *et al.* (2014) recently proposed a diagram flow to measure selectivity of big data. In the first phase unit-level data are checked if they contain units and then their representativeness is assessed by linking them to existing sources or aggregating them for comparisons with other sources. Daas and Puts (2014b) proposed using co-integration tests to measure representativeness of trends.

3. Data sources on real estate market in Poland

Poland's real estate market is partially covered by official data sources. Data about this market come from three surveys on the management of housing resources, property sales and residential and commercial property prices supported by administrative registers and non-official data bases. The survey is conducted by the National Bank of Poland (NBP) in co-operation with the Central Statistical Office in Poland (CSO); it concerns both the primary and secondary market. Since NBP is mainly responsible for the analysis, the report mostly covers aspects connected with the macroeconomic analysis at the country and city level. It is a survey of brokers who deliver information on the primary and secondary market in the biggest cities in Poland. In addition, data from various administrative sources are collected, for instance, the number of brokers and other market participants are obtained from the National Official Business Register (REGON[4]) register and the transaction data are taken from the Register of Prices and Market Value of Property (pol. *Rejestr Cen i Wartości Nieruchomości*, PVP) that is administered by local government at Local Administrative Unit 1 level (LAU1 level, counties) and contains information on transactions on both markets. In addition, non-official databases (created in collaboration with brokers' associations) are used as well as databases created and supplied with information by NBP employees. However, from the statistical point of view, the methodology of this research is not clear. For instance, there is no information on the quality and response rate of survey data, nor is it clear how NBP databases are created or what the quality of the PVP register is.

[4] http://bip.stat.gov.pl/en/regon/

Nonetheless, the PVP register is an important data source of statistical data for researching the real estate market. The legal basis is described in the Act of Geodetic and Cartographic Law with amendments (1989) and the Regulation on the Land and Buildings by the Minister of Regional Development and Construction in Poland (2001). Under these two Acts notaries are obliged to inform local authorities about transactions involving land and property. Each transaction is described with detailed characteristics (e.g. floor area, location, building characteristics) and includes the transaction price. As stated by the law, the PVP register should cover all transactions in the biggest Polish cities at the LAU1 level. There are no reports on the quality of data that PVP contains nor about how PVP is used for statistics. In addition, access to the register is limited and granted only for the purpose of evaluating new properties or to NBP/CSO employees (as at the end of 2014).

Results of the research are published in two reports. The first one is devoted to information on prices (offers and transactions) on the primary and the secondary market on a quarterly basis (National Bank Of Poland 2014a). It contains point estimates and hedonic indexes for 17 biggest Polish cities aggregated at the LAU1 level and is based on a survey of brokers, non-official data and the PVP register. The second report is delivered on a yearly basis and provides a detailed description of macroeconomic indicators and characteristics of the real estate market for 17 biggest cities in Poland excluding their agglomerations (National Bank Of Poland 2014b). However, the second report is produced and published with a delay, for instance information for 2013 was available at the end of the 2014, which indicates that information is outdated and does not reflect the current state of the real estate market. In consequence, one can observe a growing interest in reports and surveys created in by other institutions in non-official settings. On the other hand, there is a lack of research devoted to the assessment of quality and uncertainty regarding this non-official information, given that the main data source is the Internet and web portals.

For the sake of clarity it should be noted how the Polish real estate market is organized. Market participants (excluding buyers and tenants) are brokers and owners and properties can be put up for sale directly by the owner or brokers. Properties are offered by agents under two types of agreements – exclusive and open. An exclusive agreement states that only one broker can offer a given property on the market. This type of agreement is not popular owing to the limited number of possible ways of reaching potential buyers. The second type of agreement is more popular and allows brokers to co-operate and exchange information on properties for sale. The organization of the market affects research – relations between properties for sale and owner/broker could be of the "many to many" type, making identification of units difficult. In particular, when agents are using web-portals devoted to the real estate market, offers may appear more than once. Nonetheless, in order to sell, brokers and owners need to inform potential buyers about properties for sale and the Internet is becoming the main channel.

IDS have been used for real estate market research in the past. Examples of such studies can be found in working papers of Statistics Netherlands (CBS). CBS uses Funda.nl (Hoekstra *et al.* 2012), maintained by the association of Dutch brokers (nl. *Nederlandse Vereniging van Makelaars*), which is responsible for the majority of transactions on the Dutch market, to obtain data on the secondary market and to link it with registers. To achieve this, CBS has adopted a web-scraping technique, whereby all necessary information is downloaded automatically (Hoekstra *et al.* 2012). IDS concerning the real estate market can be classified into four groups – brokers' portals, brokers' association portals, portals offering brokering assistance (both for agents and owners) and services that aggregate other web-pages. The proposed classification is important in terms of quality and coverage. For instance, one broker's official website contains nearly 4,100 offers of flats for sale on the secondary market in Warsaw, Poland, while portals offering brokering assistance feature 4,500 to 5,000 offers posted by the same agent. The differences can be seen not only between brokers' activities but also between web portals. For example, four biggest web portals in Poland (measured in terms of the number of visitors) `www.otodom.pl`, `www.dom.gratka.pl`, `www.domiporta.pl`

and www.szybko.pl offer respectively 304,000, 380,000, 321,000 and 167,000 flats on the secondary market in Poland[5]. Certainly, these numbers are biased for different factors – selectivity connected with preferences in the selection of portals, duplicate adverts within and between portals, outdated, erroneous or false sale offers.

Another issue reflecting the quality of research of the real estate market is the extent of Internet coverage. The CSO conducts *Information and Communications Technologies* (ICT, Central Statistical Office 2014) survey, which is part of The Digital Agenda for Europe programme run by the European Commission. According to this survey in 2012 98.6% of companies in section L (described below) had an Internet connection (97% in 2011), 74.5% have their own website (63,3% in 2011) and 37.7% used it to present their products and prices. In Poland companies are classified into different sectors and sections. Section L refers to the real estate market and consists of four groups of companies – *purchase and sale of property on one's own account*, *leasing and management of one's own or leasehold property*, *property brokerage* and *freelance property management*. Given the level of aggregation within this section, it is difficult to directly estimate the coverage of the Internet in the group of agencies and brokers that operate on the secondary real estate market. However, it could be assumed that this level is high in the 13 biggest Polish cities. In addition, the ICT survey does not measure the use of external portals. For instance, Polish web portals devoted to the real estate market enable brokers to have private websites within their domains. Another issue is that brokers can specialize in different aspects of the real estate market - houses, flats, commercial property, sale or renting, which could affect the use of the Internet. Moreover, the Polish property market is not regulated: there is no legal control over what agent is offering a property and where the original offer has been placed. However, taking into account that most buyers are young people, the IDS should not be neglected as a source of the statistical information.

4. Empirical evaluation of representativeness

For the purpose of the study the secondary real estate market was limited to flats (units of interest) that were offered only in Poznań, Poland between 2nd quarter of 2012 to 2nd quarter of 2014. It was motivated by the availability of official data and the limited scope of this paper. The http://www.nieruchomosci-online.pl portal (NOPL) was chosen, which, unlike other portals mentioned in section 3, offers free-of-charge access to historical unit-level data. However, it should be noted that the proposed approach can be extended not only to other cities or websites but also to different fields, where IDS could be used for statistics. For this study special R code was developed to scrape information from the portal.[6] **XML** (Lang 2013), **RCurl** (Lang 2014) and **httr** (Wickham 2015) packages were used for this purpose. Algorithm 1 presents the pseudo-code for the web-scraping.

Data: Web pages, N - number of search result pages (i), n - number of results on search page (j)
Result: Text file with scraped data
Send query through form on webpage and save link to results;
Set cookies for session ;
for $i \leftarrow 1$ **to** N **do**
  Enter i result page;
  Set n ;
  for $j \leftarrow 1$ **to** n **do**
    Scrape data from j result from search result page and write it into text file;
    Enter j page from the search result page;
    Scrape all text data from j page and write it into text file;
  end
end

Algorithm 1: Pseudo-code for the algorithm for web-scraping

[5] Information on 2014-10-07

[6] Available at GitHub https://github.com/BERENZ/Papers-supplements/blob/master/AJS/Codes/NOPL_scraper.R

The algorithm produces a text file containing all scraped information on prices, floor area, number of rooms and other characteristics that could be used to identify units. In the process of data cleaning the Register on Street Names and Addresses (TERYT) was used to harmonize street names. In addition, long text descriptions included in the ads were compared with information in the remaining ads. Offers that had erroneous price per square meter (eg. lower than 1,000 PLN/m^2 or higher than 100,000 PLN/m^2) were excluded from the analysis. In the next step data were cleaned and de-duplicated using probabilistic record linkage (Fellegi and Sunter 1969) implemented in **RecordLinkage** (Borg and Sariyar 2015). Probabilistic record linkage takes into account numeric, character and missing values to link records. A 80% threshold was set to determine the probability of two records referring to the same unit. To measure the degree of representativeness data were aggregated by quarter and the number of observations for each quarter can be found in Table 1. The number of observations varies over time and is connected with the availability of data for the beginning of 2012.

Table 1: Number of Poznań real estate offers from the secondary market obtained from http://www.nieruchomosci-online.pl

Quarter	2012Q2	2012Q3	2012Q4	2013Q1	2013Q2	2013Q3	2013Q4	2014Q1	2014Q2
Nobs	2,896	3,904	6,095	6,447	6,569	9,483	13,079	11,159	4,477

The main goal of the paper is to assess the representativeness of IDS for the real estate market. For this purpose, the definition proposed by Bethlehem (2009) was adopted and the distribution of three characteristics – price per square meter, number of rooms and floor area – were compared with official reports produced by NBP/CSO. The variables describing the number of rooms and floor area were harmonized to reflect the values in the official statistics data. As a result, the number of rooms and floor area had four levels: 1 room, 2 rooms, 3 rooms, 4+ rooms and under 40 m^2, [40 m^2, 60 m^2), [60 m^2, 80 m^2), over 80 m^2 respectively. For the sake of comparison, both the primary and the secondary market data reported by NBP/CSO were used. On the following plots each red, green and blue color indicates NBP/CSO offer estimates, NBP/CSO transaction estimates and NOPL estimates respectively. Due to the variability of the estimates, only trends for the three variables are compared. Trend estimation was conducted using loess regression (with default value for `span = 0.75`) implemented in `stat_smooth` function from **ggplot2** (Wickham 2009). `stat_smooth` is a wrapper for the `loess` function from **stats** package (R Core Team 2014). The loess regression was used for three reasons: the time series for comparison is short (9 quarters) and the application of time series models (e.g. AR, MA or ARMA) is difficult. Second, estimates obtained from the NBP/CSO and NOPL vary over time and are nonlinear, which makes direct comparison of data points impossible. Finally, estimates obtained from NOPL website contain nonsampling errors, which may introduce bias in the point estimates. That is why a comparison of the trend that is estimated from the data may indicate whether the changes in the trend are in the same direction as in NBP/CSO. In addition, the loess regression approach is often used for the aggregation of polls (see Bergman and Holmquist 2014).

Figure 1 presents the price per square meter according to NOPL (blue) and NBP/CSO (green and red). At the beginning of the period of interest we can observe an increase in the price; however, this change is probably due to the quality of the data in the first years of the research. From the 2nd quarter of 2008 the trend slightly decreases until the end of the 2012. The NBP/CSO price keeps increasing from the beginning of 2013 to the end of the period analysed. Between 2nd quarter of 2012 to 3rd quarter of 2013 the NOPL offer price is closer to the transaction price reported by NBP/CSO than to the offer price. However, the trend is comparable to the NBP/CSO price due to stability in time. At the end of 2014 the NOPL price rapidly rises to catch up with the NBP/CSO offer price, with little difference between the two trends of estimates. A comparison of the direction and level of price per m^2 indicates that with respect to this variable NOPL is not representative in the light of the Bethlehem (2009) definition. However, the variable is considered without detailed information on flats

and the limitations of published data do not allow comparisons in subgroups to detect which groups are under- or over-represented. In addition, due to the fact that price per m^2 is considered as an output statistics, the relative distribution of flat characteristics need to be investigated.

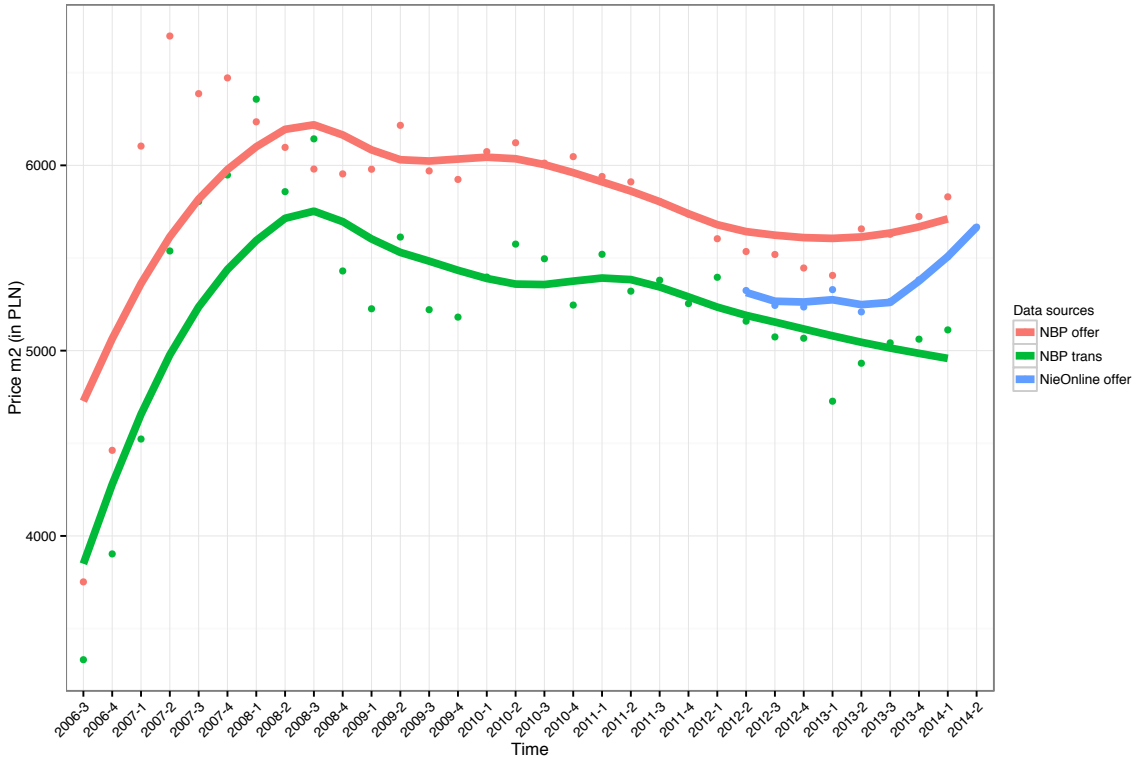


Figure 1: Comparison of offer and transaction price per square meter in Poznań, Poland reported by NBP/CSO and obtained from NOPL

In order to assess the representativeness of NOPL data, a relative distribution of floor area and the number of rooms of flats was compared with NBP/CSO reports. Figure 2 presents a comparison of harmonized floor area of flats in four categories reported by NBP/CSO. The smallest flats (under 40 m^2) and medium-sized (60-80 m^2) are represented at the same level as in official statistics data, and the trends are consistent for these groups. The blue line representing NOPL is close to the red line representing the NBP/CSO offers and is different from the green line denoting transactions. Nonetheless, the biggest (over 80 m^2) and smaller (40-60 m^2) flats are under-represented compared to official statistics data. The difference between the fraction of 40-60 m^2 flats is constant in time and the trend is consistent with that reported by NBP/CSO. In the case of the biggest flats the fraction of flats offered for sale is not at the same level as in official statistics data - the trend has a smaller slope while the direction is the same.

Figure 3 presents a comparison of number of rooms aggregated to the four categories defined in the NBP/CSO reports – 1 room, 2 rooms, 3 rooms and 4 and more rooms. In the case of flats with 1 room the NOPL trend is slightly shifted in time and reaches the same percentage of flats but with one quarter delay. However, the trend is consistent in the sense of shape and direction with the one plotted on the basis of official statistics data. The trend for flats with 3 rooms reaches the same level as in NBP/CSO reports and the differences between the trends are minor. The main differences are visible in the group of biggest flats that are under-represented in comparison with official statistics data. Nonetheless, the trend is comparable to the NBP/CSO flat offers denoted by the red line. In contrast, the trend for 2-room flats for most of the period is consistent, although in the 2nd quarter of 2012 a change in the slope can be observed. which is more comparable with transaction data, or perhaps, it indicates a change in the trend which will be visible in the upcoming report.

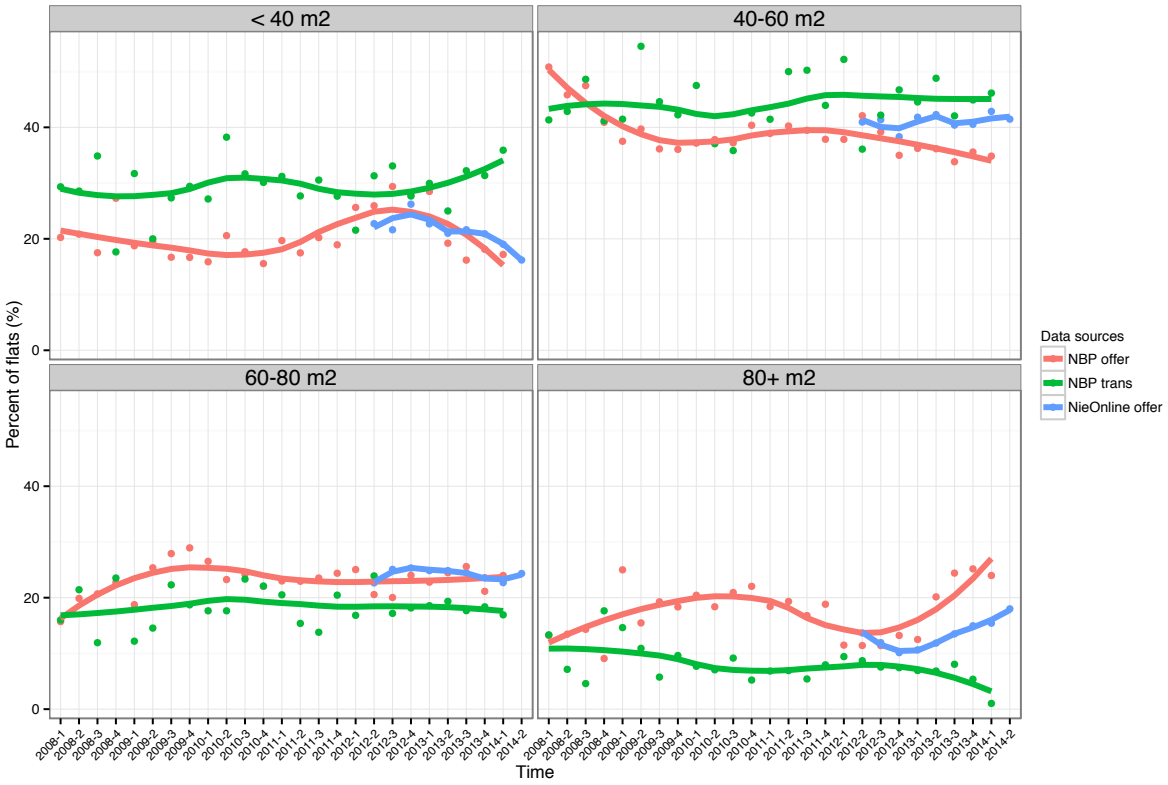


Figure 2: A comparison of characteristics of flats offered and sold on the secondary market as reported by NBP/CSO and obtained from NOPL by floor area

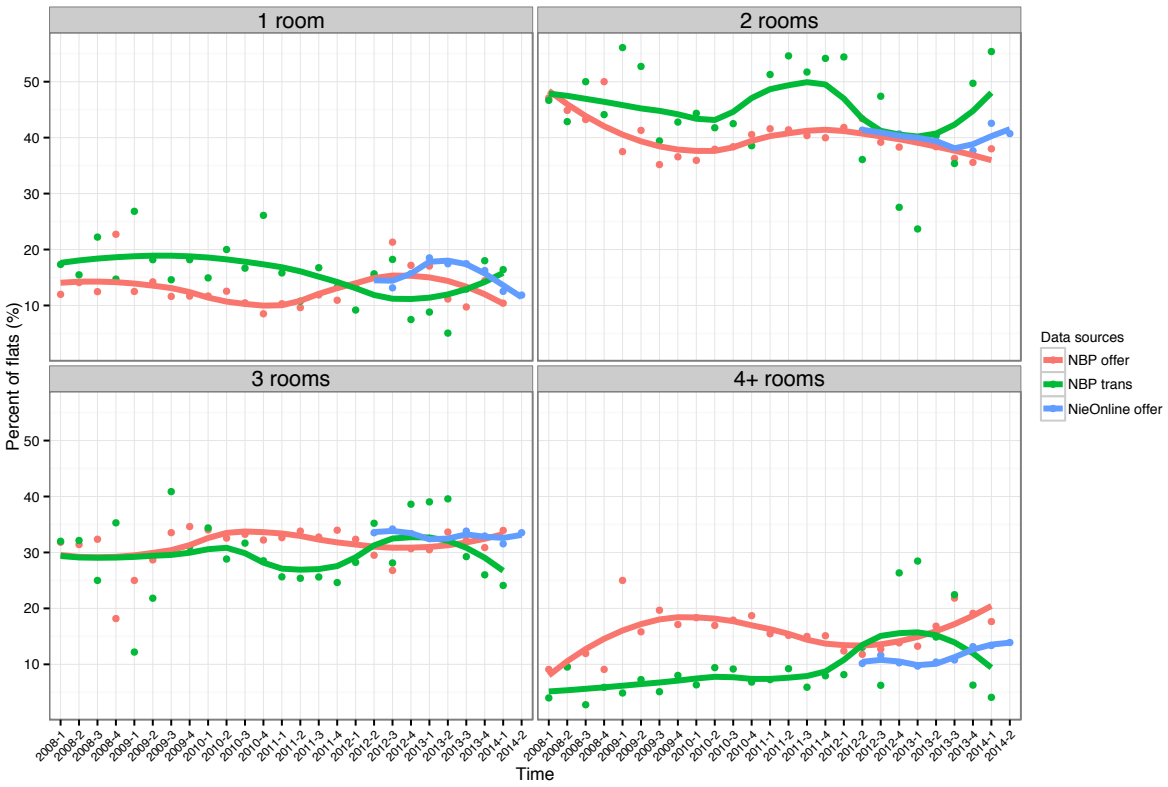


Figure 3: Comparison of characteristics of flats offered and sold on the secondary market as reported by NBP/CSO and obtained from NOPL by the number of rooms

The comparison of trends indicates that data obtained from NOPL are representative for the smallest (below 40 m^2, 1 room) and medium-sized (60-80 m^2, 3 rooms) flats. On the other hand, the group of the biggest (80+ m^2, 4+ rooms) are under-represented in comparison with official statistics data presented in NBP/CSO reports. However, the decrease in the category of under 40 m^2 flats appears first in the NOPL source and is then reflected in the NBP/CSO source, because reports from 2013 appeared at the end of the 2014. This indicates that the NOPL source could be regarded as an indicator for this category of flats. A similar relationship for big flats with floor area over 80 m^2 and 4+ rooms can be observed where an increase in the trend is first indicated in the NOPL source. As a result, the differences in the categories of flats may influence the discrepancies in the price per m^2 presented in Figure 1.

In addition, due to the non-sampling character of data obtained from the Internet, it is challenging to estimate standard errors for the estimated characteristics. In addition, NBP/CSO reports do not contain any information on standard errors of estimates, which again limits the scope of comparison of distributions. Therefore, visual analysis can be useful to detect trends and their relations with official data sources could be the first indicator of representativeness for new data sources.

5. Summary and discussion

IDS and big data have recently become the subject of evaluation by statisticians as potential statistical data sources. Despite the increasing interest in these new data sources there are several aspects that need to be considered in order to meet the criteria of statistical data sources. In order to discuss the representativeness of the IDS the definition of IDS was presented in the paper. The comparison of trends estimated by loess regression was proposed as a indication of the representativeness of the NOPL in comparison with official statistics data. Results presented in the paper indicate that for two categories (under 40 m^2 and 60-80 m^2) of flats the NOPL source could be considered representative, while the biggest flats are under-represented. However, it should be noted that the results reflect the secondary real estate market only in one city and such analysis should be extended to include other cities as well as other web-portals. Furthermore, estimation can be affected by the selection of web portals and by the data cleaning process, which in turn can influence the measuring of representativeness.

The results of the study suggest that the existing definitions and methodology, which are valid for existing statistical data sources, should be adopted or revised to deal with new data sources. Another problem is the lack of reference data from official statistics or its limited scope, which makes it difficult to measure representativeness or, more importantly, uncertainty of estimates. On the other hand, IDS cannot often be compared with existing research due to the lack of consistency and harmonized definitions. Therefore, information obtained from IDS could be treated as a proxy measure of sociological or economical phenomena (e.g. Google Trends). Moreover, there is a lack of statistical literature directly connected with estimation problems related to new data sources. Furthermore, IDS and big data should be treated as non-probability samples, whose representativeness is hard to measure. Recently Wanga, Rothschildb, Goelb, and Gelman (2014) proposed Bayesian model-based estimation and post-stratification that could be one of the possible approaches to the problem. Nonetheless, new data sources open up possibilities of extending the set of statistical sources, which should not be neglected.

Acknowledgement

The article and the research has been financed by National Science Centre Poland, Preludium 7 grant no. 2014/13/N/HS4/02999. In addition, I would like to thank two anonymous reviewers for helpful comments.

References

Bapna R, Goes P, Gopal R, Marsden JR (2006). "Moving from Data-Constrained to Data-Enabled Research: Experiences and Challenges in Collecting, Validating and Analyzing Large-Scale e-Commerce Data." *Statistical Science*, **21**(2), 116–130. ISSN 0883-4237. doi:10.1214/088342306000000231. 0609136v1, URL http://projecteuclid.org/Dienst/getRecord?id=euclid.ss/1154979815/.

Bayer M (2011). "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data." URL http://www.gartner.com/newsroom/id/1731916.

Bayer M, Laney D (2012). "The Importance of 'Big Data': A Definition." URL https://www.gartner.com/doc/2057415/importance-big-data-definition.

Bergman J, Holmquist B (2014). "Poll of Polls : A Compositional Loess Model." *Scandinavian Journal of Statistics*, **41**(2), 301–310. doi:10.1111/sjos.12023.

Bethlehem J (2008). "Representativity of Web Surveys–An Illusion?" *Access panels and online research, panacea or pitfall*, pp. 19–44.

Bethlehem J (2009). *Applied Survey Methods: A Statistical Perspective*. John Wiley & Sons.

Bethlehem J, Biffignandi S (2011). *Handbook of Web Surveys*. John Wiley & Sons.

Borg A, Sariyar M (2015). *RecordLinkage: Record Linkage in R*. R package version 0.4-7, URL http://CRAN.R-project.org/package=RecordLinkage.

Buelens B, Daas P, Burger J, Puts M, van den Brakel J (2014). "Selectivity of Big Data." URL http://www.pietdaas.nl/beta/pubs/pubs/Selectivity_Buelens.pdf.

Cavallo A (2012). "Scraped Data and Sticky Prices." *MIT Sloan Research Paper*. URL http://www.mit.edu/%7Eafc/papers/Cavallo-Scraped.pdf.

Cavallo A (2013). "Online and Official Price Indexes: Measuring Argentina's Inflation." *Journal of Monetary Economics*, **60**(2), 152–165.

Central Statistical Office (2014). *Information Society in Poland Statistical Results From the Years 2009-2013 (in Polish)*. Statistical Office in Szczecin, Warsaw, Poland. URL http://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5497/1/7/4/spolecz_inform_w_polsce_2009-2013.pdf.

Choi H, Varian H (2012). "Predicting the Present with Google Trends." *Economic Record*, **88**(s1), 2–9.

Daas P, Puts M (2014a). "Big Data As a Source of Statistical Information." *The Survey Statistician*, **69**, 22–31. URL http://pietdaas.nl/beta/pubs/pubs/Big_data_survey_stat.pdf.

Daas P, Puts M (2014b). "Social Media Sentiment and Consumer Confidence." URL http://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.pdf.

Daas P, Roos M, de Blois C, Hoekstra R, ten Bosch O, Ma Y (2011). "New Data Sources for Statistics: Experiences at Statistics Netherlands." In *Paper for the 2011 European New Technique and Technologies for Statistics conference, February*, pp. 22–24.

Daas P, Roos M, van de Ven M, Neroni J (2012). "Twitter As a Potential Data Source for Statistics." URL http://pietdaas.nl/beta/pubs/pubs/DiscPaper_Twitter.pdf.

Fellegi IP, Sunter AB (1969). "A Theory for Record Linkage." *Journal of the American Statistical Association*, **64**(328), 1183–1210.

Fondeur Y, Karamé F (2013). "Can Google data help predict French youth unemployment?" *Economic Modelling*, **30**, 117–125. ISSN 02649993. doi:10.1016/j.econmod.2012.07.017. URL http://linkinghub.elsevier.com/retrieve/pii/S0264999312002490.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2008). "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature*, **457**(7232), 1012–1014.

Gołata E (2014). "New Paradigm in Statistics and Population Census Quality." European conference on quality in official statistics, URL http://www.q2014.at/fileadmin/user_upload/GOLATA_NEW.pdf.

Hoekstra R, ten Bosch O, Harteveld F (2012). "Automated Data Collection From Web Sources for Official Statistics: First Experiences." *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, **28**(3), 99–111.

Kruskal W, Mosteller F (1979a). "Representative Sampling I: Non-scientific Literature." *International Statistical Review*, **47**, 13–24. URL http://www.jstor.org/stable/1402564.

Kruskal W, Mosteller F (1979b). "Representative Sampling II: Scientific Literature Excluding Statistics." *International Statistical Review*, **47**, 111–123. URL http://www.jstor.org/stable/1402564.

Kruskal W, Mosteller F (1979c). "Representative Sampling III: The Current Statistical Literature." *International Statistical Review*, **47**, 245–265. URL http://www.jstor.org/stable/1402647.

Lang DT (2013). *XML: Tools for Parsing and Generating XML Within R and S-Plus*. R package version 3.98-1.1, URL http://CRAN.R-project.org/package=XML.

Lang DT (2014). *RCurl: General Network (HTTP/FTP/...) Client Interface for R*. R package version 1.95-4.3, URL http://CRAN.R-project.org/package=RCurl.

Miller G (2011). "Social Scientists Wade Into the Tweet Stream." *Science*, **333**(6051), 1814–1815.

Mohorko A, Leeuw Ed, Hox J (2013). "Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time." *Journal of Official Statistics*, **29**(4), 609–622.

National Bank Of Poland (2014a). *Real Estate Market – Quarterly Report*. National Bank of Poland, Finance stability department, Warsaw, Poland. URL http://www.nbp.pl/homen.aspx?f=/en/publikacje/inne/real_estate_market_q.html.

National Bank Of Poland (2014b). *Report On the Situation in the Polish Residential and Commercial Real Estate Market in 2013*. National Bank of Poland, Finance stability department, Warsaw, Poland. URL http://www.nbp.pl/en/publikacje/inne/annual_report_2013.pdf.

Porter AT, Holan SH, Wikle CK, Cressie N (2013). "Spatial Fay-Herriot Models for Small Area Estimation with Functional Covariates." *arXiv preprint arXiv:1303.6668*.

Pratesi M, Giannotti F, Giusti C, Marchetti S, Pedreschi D, Salvati N (2014). "Area Level Sae Models with Measurement Errors in Covariates: An Application to Sample Surveys and Big Data Sources." Small Area Estimation, URL http://sae2014.ue.poznan.pl/SAE2014_book.pdf.

Pratesi M, Pedreschi D, Giannotti F, Marchetti S, Salvati N, Maggino F (2013). "Small Area Model-Based Estimators Using Big Data Sources." NTTS, URL http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_208.pdf.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schouten B, Cobben F, Bethlehem J (2009). "Indicators for the Representativeness of Survey Response." *Survey Methodology*, **35**(1), 101–113.

Shmueli G, Jank W, Bapna R (2005). "Sampling eCommerce Data From the Web: Methodological and Practical Issues." In *ASA Proc. Joint Statistical Meetings*, volume 941, p. 948. URL https://archive.nyu.edu/bitstream/2451/14953/2/USEDBOOK11.pdf.

Vosen S, Schmidt T (2011). "Forecasting Private Consumption: Survey-Based Indicators Vs. Google Trends." *Journal of Forecasting*, **30**(6), 565–578.

Wallgren A, Wallgren B (2014). *Register-based Statistics*. Wiley Series in Survey Methodology, second edition. John Wiley & Sons, Inc. ISBN 9781119942139.

Wanga W, Rothschildb D, Goelb S, Gelman A (2014). "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting. Forthcoming.*

Wickham H (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Springer New York. ISBN 978-0-387-98140-6. URL http://had.co.nz/ggplot2/book.

Wickham H (2015). *Httr: Tools for Working with URLs and HTTP*. R package version 0.6.1, URL http://CRAN.R-project.org/package=httr.

Xu W, Li Z, Cheng C, Zheng T (2012). "Data Mining for Unemployment Rate Prediction Using Search Engine Query Data." *Service Oriented Computing and Applications*, **7**(1), 33–42. ISSN 1863-2386. doi:10.1007/s11761-012-0122-2. URL http://link.springer.com/10.1007/s11761-012-0122-2.

Zhang LC (2011). "A Unit-Error Theory for Register-Based Household Statistics." *Journal of Official Statistics*, **27**(3), 415–432.

Zhang LC (2012). "Topics of statistical theory for register-based statistics and data integration." *Statistica Neerlandica*, **66**(1), 41–63. ISSN 00390402. doi:10.1111/j.1467-9574.2011.00508.x.

Zhang LC (2015). "On Modelling Register Coverage Errors." *Journal of Official Statistics. Forthcoming.*

Affiliation:

Maciej Beręsewicz
Department of Statistics
Poznan University of Economics
61-875 Poznan, Poland
E-mail: maciej.beresewicz@ue.poznan.pl

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 44 *Submitted:* 2014-10-07
April 2015 *Accepted:* 2015-02-18

Austrian Journal of Statistics
April 2015, Volume 44, 59–71.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i2.77

MI Double Feature: Multiple Imputation to Address Nonresponse and Rounding Errors in Income Questions

Jörg Drechsler
Institute for Employment Research

Hans Kiesl
OTH Regensburg

Matthias Speidel
Institute for Employment Research

Abstract

Obtaining reliable income information in surveys is difficult for two reasons. On the one hand, many survey respondents consider income to be sensitive information and thus are reluctant to answer questions regarding their income. If those survey participants that do not provide information on their income are systematically different from the respondents (and there is ample of research indicating that they are) results based only on the observed income values will be misleading. On the other hand, respondents tend to round their income. Especially this second source of error is usually ignored when analyzing the income information.

In a recent paper, Drechsler and Kiesl (2014) illustrated that inferences based on the collected information can be biased if the rounding is ignored and suggested a multiple imputation strategy to account for the rounding in reported income. In this paper we extend their approach to also address the nonresponse problem. We illustrate the approach using the household income variable from the German panel study "Labor Market and Social Security".

Keywords: heaping, measurement error, multiple imputation, nonresponse, poverty rate.

1. Introduction

Reliable information on individual and household income is difficult to obtain. Most administrative data sources contain only specific sources of income such as income from earnings or program participation and often only cover a subset of the population (self-employed are usually not included). Thus, most agencies rely on household surveys to collect information on total income. However, inferences based on the collected income information might be biased for two reasons: First, income is considered sensitive information and many survey participants are reluctant to answer questions on their personal income. Second, most respondents do not remember their exact income, especially if they are asked to provide an estimate for their total income including income from earnings, assets, transfers, etc. Respondents often round their income in this case, implicitly incorporating their uncertainty regarding the true value.

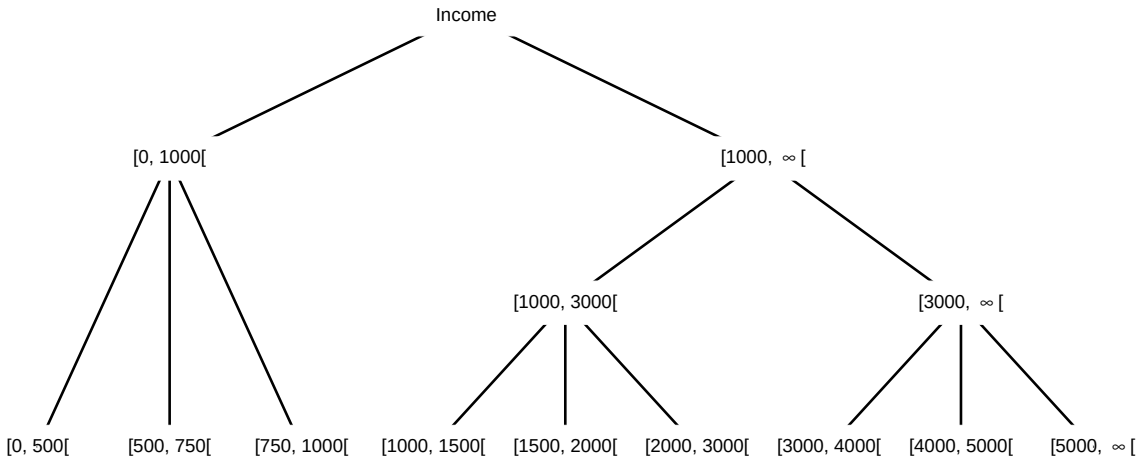


Figure 1: Implied income intervals based on partial income information collected from respondents unwilling to provide their exact income.

Nonresponse can bias inferences if the respondents are systematically different from the nonrespondents. For example, it seems plausible to assume that younger survey respondents are less concerned with confidentiality violations and the protection of sensitive information ("generation Facebook") and thus, their response rates to income questions will be higher. Since income usually increases with age, individuals with lower income will be over-represented among the respondents in this case and the average income of the population will be underestimated if only the observed income values are used.

To reduce the risk of nonresponse bias, many surveys try to obtain at least partial income information for those survey participants that are unwilling or unable to provide exact income information by asking whether the income lies in certain pre-specified intervals. Often subsequent questions further narrow down the interval in which the true income falls. Figure 1 provides an example how (partial) income information is collected in the German panel study "Labor Market and Social Security" (PASS) (Trappmann, Gundert, Wenzig, and Gebhardt 2010). Respondents are first asked for an estimate of their total household income. If they are unwilling or unable to provide this information, the interviewer provides a first threshold (1,000 euros) and asks whether the income is above or below that threshold. Depending on the answer to this question the survey participant is asked to choose from three specific intervals (if the respondent reported an income below 1,000 euros for the first question) or a new threshold (3,000 euros) is provided and the respondent is asked again whether his or her income is above or below this threshold. If the respondent provides an answer to the second threshold question, three different income intervals are offered for both response options and the respondent is asked to pick the interval in which his or her income falls. Figure 1 illustrates the decision steps and the corresponding income intervals that are implied by the responses to each of the questions. The interview process could terminate in any of the nodes of the decision tree. For example, a respondent might refuse to provide the exact income information but might be willing to provide the information that his or her income is larger than 1,000 but less than 3,000 euros. However, he or she might be unwilling to further specify whether the income is in the interval $[1,000,\ 1,500[$ or $[1,500,\ 2,000[$ or $[2,000,\ 3,000[$.

Asking those respondents that are unwilling to provide their exact income for information regarding the interval in which their income falls is a successful strategy to reduce the nonresponse rate. For example, in wave six of the PASS survey, 76.96% of the respondents who are unwilling or unable to provide their exact income provided some information on the interval in which their income falls, reducing the initial nonresponse rate from 4.56% to 1.05%.

Following this procedure, the collected income information consists of exact information for those respondents that are willing to answer the exact income question and interval informa-

Table 1: Percentage of reported monthly household income values that are divisible by a given round number in the PASS survey for the year 2008/2009.

Income divisible by	1,000	500	100	50	10	5
Relative frequency (%)	13.97	23.94	61.57	69.58	80.71	84.13

tion of different lengths for those individuals that answer (some of) the interval questions. Directly obtaining valid inferences from this type of data is not straightforward, especially if refusal to answer any of the income questions should also be taken into account. In this paper we will present an imputation approach that simplifies the analysis of the collected income data. The multiple imputation methodology is not only used to impute the missing values; plausible exact income values are also generated for those respondents that only provided interval information regarding their income. The obtained imputed income data can be analyzed as if the exact income would have been obtained for all respondents. The additional uncertainty implied by the fact that only partial information is available for some of the respondents is correctly reflected through the multiple imputation procedure.

The negative effects of nonresponse are well known. However, the impacts of heaping, i.e., rounding to certain numbers such as multiples of 5, 10, 100, etc., are less studied. Rounding is a common phenomenon in surveys. Most quantitative variables such as questions on expenditure or subjective beliefs (*How likely is it that...*) show some form of rounding (Manski and Molinari 2010). But also questions on timing of events (Huttenlocher, Hedges, and Bradburn 1990) or smoking behavior typically are affected (Wang and Heitjan 2008). In a recent experimental study Ruud, Schunk, and Winter (2013) demonstrated that the amount of rounding increases with the level of uncertainty the respondent feels regarding the quantity he or she is asked for. Regarding questions on income the level of uncertainty is usually very high. Most respondents do not know their income from earnings to the exact euro amount (especially if the earnings before taxes is requested) and exact values for other sources such as monthly income from savings are even more difficult to provide. Thus, it is not surprising that questions on income usually show a large degree of rounding. Table 1 provides the percentage of the reported monthly income values that are divisible by a given round number obtained from the PASS survey for the year 2008/2009 (see Section 4 for a description of the survey). It seems that most of the reported data are rounded to some extent. More than 60% of the reported income values are divisible by 100 and only about 15% of the data are not divisible by 5.

Drechsler and Kiesl (2014) illustrate that heaping in income data can cause substantial bias in important measures such as the poverty rate. They also suggest a strategy for dealing with the problem and demonstrate its merits through simulations and real data applications. The basic idea is to model the rounding behaviour given the reported income value and then to replace the reported value by multiple plausible candidates for the true value that would have been observed if the respondent had not have rounded his or her income. A related idea has been proposed by Heitjan and Rubin (1990) for heaped age data and has later been applied in a number of papers to model the smoking behaviour based on reported cigarette counts (Heitjan 1994; Wang and Heitjan 2008; Wang, Shiffman, Griffith, and Heitjan 2012). The major advantage of the approach is that the imputed values can be treated as true values in any analysis following the imputation, i.e., it is not necessary to develop adjustment methods for each type of analysis separately. The analyst only needs to repeat the analysis of interest on each imputed dataset using standard analysis techniques. The final inferences are obtained using standard multiple imputation combining rules (Rubin 1978, 1987).

In this paper we extend the approach by Drechsler and Kiesl (2014) in order to address (partial) nonresponse and heaping simultaneously. We review the approach of Drechsler and Kiesl (2014) in Section 2 and discuss the necessary extensions to incorporate the interval information and to adjust for nonresponse in Section 3. In Section 4 we illustrate the approach based on data from the PASS survey. The paper concludes with some final remarks.

2. Strategies to adjust for rounding errors

This section discusses the imputation approach suggested by Drechsler and Kiesl (2014) which itself is based on an idea by Heitjan and Rubin (1990). In their paper Heitjan and Rubin (1990) proposed to use multiple imputation to correct for heaped reported age values of young children in Tanzania. The section borrows heavily from Drechsler and Kiesl (2014) and we refer the reader to this paper for a more detailed discussion of the methodology.

To obtain imputed income values that are adjusted for potential rounding, we need two models: one for the true income and one for the rounding behaviour. Following common practice, we model the conditional distribution of the household income Y given some covariates X by a log-normal distribution (see, for example, Clementi and Gallegati (2005) for a motivation for this model):

$$\log(Y)|X \sim N(X'\beta,\ \sigma^2). \tag{1}$$

We only consider rounding to the nearest multiple of c, which corresponds to the rounding function $f_c : x \mapsto c \cdot \lfloor x/c + 1/2 \rfloor$ and which we call rounding of degree c . Other rounding models could be considered: for example, Heitjan and Rubin (1990) suggest a model in which some age values are truncated and not rounded. However, we feel that rounding to the nearest multiple of c is the most plausible rounding strategy for income data. In our model, no rounding at all will be called rounding of degree 0. We assume that there are p possible degrees of rounding $c_1 < ... < c_p$. Typically, the set of c_i's consists of values such as 0, 1, 5, 10, 50, 100. For a given household, our model for the degree of rounding is an ordered probit model, i.e., we assume a normally distributed latent variable G which may (linearly) depend on the logged income $\log(Y)$ and some covariates Z (where some or all components of Z might be in X and vice versa):

$$G|\log(Y), Z \sim N(\gamma_0 + \gamma_1 \cdot \log(Y) + Z'\gamma_2,\ \tau^2)$$

Rounding of degree c_1 occurs, if $G < k_1$; rounding of degree c_i $(1 < i < p)$ occurs, if $G \in [k_{i-1}, k_i[$; rounding of degree c_p occurs, if $G \geq k_{p-1}$. The $p-1$ threshold values $k_1 < k_2 < ... < k_{p-1}$ are unknown model parameters.

We assume that given X, $\log(Y)$ and Z are independent, and analogously, given Z, G and X are independent. Under these assumptions $\log(Y)$ and G have the following bivariate normal distribution given X and Z:

$$\log(Y), G|X, Z \sim N(\mu,\ \Omega), \quad \text{where}$$

$$\mu = \begin{pmatrix} X'\beta \\ \gamma_0 + X'\gamma_1\beta + Z'\gamma_2 \end{pmatrix}, \tag{2}$$

$$\Omega = \begin{pmatrix} \sigma^2 & \gamma_1\sigma^2 \\ \gamma_1\sigma^2 & \tau^2 + \gamma_1^2\sigma^2 \end{pmatrix}. \tag{3}$$

To impute true income values based on these models, it is necessary to derive the likelihood for all the unknown parameters $\Psi = (\beta,\ \sigma^2,\ \gamma_1,\ \gamma_2,\ k_1,\ ...,\ k_{p-1})$ (we need to fix γ_0 at 0 and τ^2 at 1 to make the ordered probit model identifiable). Let s_i be the observed income of household i. It can be shown that this likelihood is given as (see Drechsler and Kiesl (2014) for details)

$$\begin{aligned} L(\Psi|s,\ x,\ z) &= \prod_i f(s_i,\ x_i,\ z_i|\Psi) \\ &= \prod_i f(x_i,\ z_i) \cdot \prod_i f(s_i|x_i,\ z_i,\ \Psi) \\ &\propto \prod_i \iint_{A(s_i)} f(g,\ \log(y)|x_i,\ z_i,\ \Psi) d\log(y) dg, \end{aligned} \tag{4}$$

where $A(s_i)$ is the set of $(g,\ \log(y))$ that are consistent with an observed s_i.

Maximizing this likelihood will provide the parameter vector Ψ necessary for the imputations. To approximate a draw from the posterior distribution of $f(\Psi|s,\ x,\ z)$ under the assumption of flat priors for all parameters, we can draw from

$$\Psi^* \sim MVN(\hat{\Psi}_{ML},\ I(\hat{\Psi}_{ML})),$$

where $\hat{\Psi}_{ML}$ contains the maximum likelihood estimates of Ψ, and $I(\hat{\Psi}_{ML})$ is the negative inverse of the Hessian matrix of the log-likelihood with $\hat{\Psi}_{ML}$ plugged in.
To impute exact income values, Drechsler and Kiesl (2014) suggest a simple rejection sampling approach:

1. Draw candidate values for $(\log(y_i)^{imp},\ g_i)$ from a truncated bivariate normal distribution with mean vector (2) and covariance matrix (3) (using parameters from Ψ^*), where the truncation points are given by the maximal possible degree of rounding given the observed income s_i (for example, for an observed income value 850 with possible degrees of rounding 1, 5, 10, 50, 100, 500, and 1,000, $\log(y_i)$ is bounded by log(825) and log(875) and g_i has to be in $]-\infty,\ k_4^*[$).

2. Accept the drawn values if they are consistent with the observed rounded income, i.e., rounding the drawn income value according to the drawn rounding indicator gives the observed income s_i, and impute $\exp(\log(y_i)^{imp})$ as the exact income value.

3. Otherwise draw again.

Repeating this procedure m times provides m imputed datasets that properly reflect the uncertainty from imputation.

3. Extensions for (partial) nonresponse

As discussed in the introduction, many agencies ask respondents who refuse to answer the exact income question whether they would be willing to provide information in which given interval their income falls. This partial information can be used to improve the inferences regarding the income variable. In this paper we suggest to use this partial information when setting up the likelihood and then to impute plausible true income values for each reported income interval. The approach is related to the approach to account for rounding described in the previous section with the only difference that the interval in which the true income must fall is known in advance and does not need to be estimated from the observed data.

Let r_i, $r_i \in \{0,\ 1,\ ...,\ R+1\}$, be a random variable that identifies to which income response group individual i, $i = 1, ..., n$ belongs. Let $r_i = 0$ represent exact income information (which might still be affected by rounding) and let $r_i = 1,\ ...,\ R$ identify the R different income intervals that could be selected from the predefined intervals provided by the agency. For example, according to Figure 1 $R = 13$ in the PASS survey. Finally, let $r_i = R + 1$ represent refusal to provide any income information at all. Let I_i^r be an indicator function that equals 1 if individual i belongs to income response group r and equals 0 otherwise. Let l^r and u^r be the upper and lower bound of the income interval for response group r. We set $l^0 = y = u^0$ and $l^{R+1} = -\infty$ and $u^{R+1} = +\infty$. All other bounds are defined by the income intervals provided by the agency. We extend the definition of s_i to also include all reported income intervals, i.e., s_i is a single value for all individuals that reported the exact income, but is an interval for all individuals that only provided the information in which interval their income falls. The extended likelihood that also takes the interval information into account is given by

$$
\begin{aligned}
L(\Psi|s,\ x,\ z) &= \prod_i f(x_i,\ z_i) \cdot \prod_i f(s_i|x_i,\ z_i,\ \Psi) \qquad (5)\\
&\propto \prod_i \{(\iint_{A(s_i)} f(g,\ \log(y)|x_i,\ z_i,\ \Psi) d\log(y) dg)^{I_i^0}\\
&\cdot \prod_{r=1}^{R+1} [F(\log(u_i^r),\ \mu_i = x_i'\beta,\ \sigma^2) - F(\log(l_i^r),\ \mu_i = x_i'\beta,\ \sigma^2)]^{I_i^r}\}.
\end{aligned}
$$

Once estimates for all parameters are obtained by maximizing the likelihood in (5), imputation of the plausible values for the true income Y is straightforward. The first imputation step is similar to Section 2: Approximate a draw from the posterior distribution of the parameters by drawing from a multivariate normal with mean equal to the maximum likelihood estimates of the parameters and variance equal to the negative inverse of the Hessian matrix of the log-likelihood. The second step depends on the type of data that is imputed. The true income for all exact reporters is imputed as described in Section 2 to account for potential rounding in the reported income values. The true income for the interval respondents is imputed by drawing from a truncated normal distribution $N_t(\mu, \sigma^2)$ with $\mu = X'\beta^*$, $\sigma^2 = (\sigma^*)^2$, where β^* and $(\sigma^*)^2$ are the drawn parameters from step one. The truncation points are given by the bounds of the reported income interval. Finally, imputations for those respondents that refused to provide any information regarding their income are obtained by drawing from a normal distribution with parameters $\mu = X'\beta^*$ and $\sigma^2 = (\sigma^*)^2$.

4. Application to the panel study Labor Market and Social Security

We illustrate the application of our approach using data from the German panel study "Labor Market and Social Security" (PASS). To enable a comparison of our extended approach with the approach of Drechsler and Kiesl (2014) that only focuses on rounding, we use the same models for the income and rounding behaviour and also use the poverty rate to evaluate which impacts the adjustments have on important measures that are regularly computed from income data. The poverty rate is defined as the percentage of persons with an income less than a fixed percentage of the median income. For example, in the European countries the poverty rate is defined as the proportion of persons with an income less than 60% of the median income.

Before presenting the results, we provide a description of the data and a short summary of the imputation models borrowed from Drechsler and Kiesl (2014). The interested reader is referred to this paper for more details.

The PASS survey started in 2006 and conducted yearly ever since, aims at measuring the social effects of labour market reforms. The survey consists of two different samples, each containing roughly 6,000 households. The first sample is drawn from the Federal Employment Agency's register data containing all persons in Germany receiving unemployment benefit for long time unemployment. The second sample is drawn from the MOSAIC database of housing addresses collected by the commercial data provider, microm. This sample is representative for the resident population in Germany. The stratified sampling design for this sample oversamples low-income households. The major benefit of this combination of two different samples lies in the fact that control groups for the benefit recipients can easily be constructed. The panel contains a large number of socio-demographic characteristics (for example, age, gender, marital status, religion, migration background), employment-related characteristics (for example, status of employment, working hours, income from employment, employment history), benefit-related characteristics (for example, benefit history, amount of

Table 2: Covariates included in the income model.

variable	characteristics
household size	5 categories (household sizes> 4 set to "5 or more")
deprivation index	range: 0–21
living space	range: 7–903 square meters
type of household	8 categories
amount of debt	7 categories
income from savings	yes/no
age of respondent	range: 15–99
amount of savings	8 categories (not available for wave 1)
unemployment benefits	yes/no
weight	range: 24.95–186,000

benefits, participation in training measures), and subjective indicators (for example, fears and problems, employment orientation, subjective social position). A detailed description of the survey can be found in Trappmann *et al.* (2010).

To model the true income, we assume a log-normal distribution for income conditional on a set of covariates X. Details about the covariates included in the model are contained in Table 2.

All variables are standardized, some sparsely populated categories in X are collapsed and influential outliers are removed to ensure convergence of the maximisation procedure (see Drechsler and Kiesl (2014) for details). For the rounding behaviour, we assume that the tendency to round only depends on the true income.

4.1. Evaluation of the model assumptions

Since the proposed rounding adjustment strategy is purely model based, an evaluation of the model assumptions is essential. We follow the approach of Drechsler and Kiesl (2014) to check whether the model assumptions are reasonable. They suggest to use posterior predictive simulations (Gelman, Carlin, Stern, and Rubin 2004, Chap. 6) for the evaluations since the true income and the rounding behaviour are never observed which complicates the evaluation.

The income model

For the income model evaluation we generate a very large number of imputations for the true income based on the parameters obtained from maximizing the likelihood in (5) at the last iteration of the sequential regression imputation procedure (see Section 4.2 for details). The rounding behaviour is completely ignored here, i.e., imputations are generated for all observations based on the marginal income model described in (1). The obtained imputations can be seen as samples from the posterior predictive distribution of the income for each observation according to the model. To evaluate the model fit we can check whether these posterior distributions cover the observed income values from the original data. Of course many of the observed income values are subject to rounding, so we limit the evaluation to those records for which we can be sure that the reported value is only rounded to the next euro (i.e., all records for which the reported value is only divisible by 1). If the imputation model is correct, the true (observed) income should be covered in the region between the empirical $\alpha/2$ quantile and the $1-\alpha/2$ quantile of the imputed values with a probability of $1-\alpha$. Thus, as a measure for the model fit we calculate the fraction of unrounded income values from the observed data that are covered by this interval computed from the imputed values and compare this fraction to the expected coverage rates. Results based on $m = 1,000$ imputations are presented in Table 3. The empirical coverages are generally close to the nominal coverages: except for wave 2 and 5 the empirical coverages never differ more than

Table 3: Percentage of true income values from the PASS survey that are covered in the defined regions of the posterior distribution of the imputed income values.

Expected	Empirical Coverage (in %)					
Cov. (in %)	wave 1	wave 2	wave 3	wave 4	wave 5	wave 6
99.00	97.65	93.76	97.31	97.19	95.43	96.87
95.00	95.06	91.63	93.34	93.57	92.69	93.66
90.00	91.91	89.00	89.72	89.31	88.55	89.53

Table 4: Percentage of income values that are divisible by a given round number (but not by any of the larger numbers) in the observed PASS data, the unrounded data, and the re-rounded data.

Income divisible by	1	5	10	50	100	500	1,000
Observed income (%)	14.94	4.05	11.58	7.74	37.34	10.29	14.06
Unrounded income (%)	80.05	9.98	7.97	1.00	0.79	0.11	0.10
Re-rounded income (%)	9.67	2.93	12.10	9.49	45.79	10.08	9.94

2.2 percentage points from the nominal coverages. The largest differences are observed for the expected 99% coverage rate for wave 2 (difference of 5.24 percentage points) and wave 5 (3.57 percentage points). But even for these waves the nominal coverages never differ more than 1.5 percentage points from the expected 90% coverage rate. Overall the results indicate a reasonable fit for the income model.

The rounding behaviour model

To evaluate the quality of the rounding behaviour model, we repeatedly re-round the imputed (unrounded) income variable based on the obtained likelihood parameters and compare it to the originally observed data. Specifically, we repeatedly ($m = 100$) generate unrounded income data that are consistent with the original data according to the joint model for income and rounding behaviour. Then, we repeatedly round each of the obtained exact income variables (100 times for each of the generated income variables) according to the rounding probabilities based on the parameters from the rounding behaviour model. Since we have no direct measure for the rounding behaviour we use a proxy for the evaluation. We compare the share of the income values that are divisible by values that are typically used as rounding bases. Table 4 lists these shares for the original data, the re-rounded data (computed as the average across the 10,000 generated datasets) and the unrounded data (computed as the average across the $m = 100$ replicates). Each column reports the percentage of records for which the given number represents the maximum possible rounding base, i.e., these records would not be divisible by any of the larger rounding bases listed in the table. The results are pooled across all waves of the PASS data for readability. Similar results were obtained when looking at each wave individually.

As expected the percentages differ substantially between the observed income and the unrounded income. Most of the values (80.05%) in the unrounded data (second row in the table) are only divisible by one and the percentages decrease quickly as the rounding base increases (note that we assume that values in the unrounded data are always rounded to the nearest euro). This is different for the observed data (first row). Only 14.94% of the data are only divisible by 1 and 37.34% of the records have a maximum rounding base of 100. The divisibility of the re-rounded data (third row) is reasonably close to the observed data. Again, most records are in the category with a maximum rounding base of 100, although the percentage of records that fall into this category is slightly overestimated (45.79%). This overestimation leads to a slight underestimation of the percentage of records that are only divisible by one (9.67%). For most of the remaining categories the percentages based on the

re-rounded data are fairly close to the percentages based on the observed data: the difference in percentage points is less than 1.2 for the rounding bases 5, 10, and 500. The percentage of records with maximum rounding bases of 50 and 1,000 differ somewhat more between the observed and the re-rounded data (1.75 and 4.12 percentage points respectively). Overall the results indicate a reasonable fit of the rounding behaviour model.

4.2. Results

We compare three different approaches to estimate the poverty rates from the six waves of the PASS survey that are available so far. In the first approach we treat the reported income as the true income and only use the information from those respondents that answered the exact income question. To keep the results consistent with the second approach described below, we also exclude the respondents that provided an answer to the exact income question but did not provide an answer for at least one of the covariates listed in Table 2. This approach assumes that the reported income is never rounded and implies that the respondents to the exact income question are not systematically different regarding their income from those that only provide income intervals, completely refuse to provide any information regarding their income, or have missings in the list of covariates, i.e., this approach assumes that the income information is missing completely at random (MCAR) in the terminology of Rubin (1976). In the second approach we use the methodology of Drechsler and Kiesl (2014) to account for the rounding but still only use the data from respondents who provided an answer to the exact income question and all the covariates, i.e., we still assume MCAR. The final approach is the extended approach described in this paper which also takes the information from the interval respondents into account and imputes the missing information in the covariates and missing income information for those survey participants that completely refused to provide any information regarding their income. We note that this approach uses more information to estimate the parameters in the imputation model and only assumes that the income information is missing at random (MAR), i.e., the missingness can be explained by the covariates included in the imputation model.

We apply the models described above separately for each year (the variable *amount of savings* is not available in the first wave of the survey and is thus excluded from the income model in that year). For the third approach the imputation routine for the true income is incorporated into a sequential regression multivariate imputation (SRMI, Raghunathan, Lepkowski, van Hoewyk, and Solenberger (2001)) procedure to impute missing values in any of the covariates. With the SRMI approach missing values in any of the variables are imputed by iteratively drawing from the conditional distributions of each variable given all the other variables. The process of iteratively drawing from the conditional distributions can be viewed as a Gibbs sampler that will converge to draws from the theoretical joint distribution of the data if this joint distribution exists. This is not guaranteed in practice. However, Liu, Gelman, Hill, Su, and Kropko (2013) show that consistent results can still be obtained if the conditional models are correctly specified.

To improve the quality of the imputations we included some additional variables in the imputation models for the covariates. We treated the first 100 iterations of the Gibbs sampler in each wave as the burn-in phase to ensure convergence and stored every 5^{th} iteration after the burn in phase as one imputed dataset. Traceplots of all variable means and variances and the Heidelberger&Welch diagnostic (Heidelberger and Welch 1983) indicated that all Gibbs samplers converged after 90 iterations and autocorrelation plots showed no significant correlation after 3 iterations.

Table 5 presents the poverty rates for the different waves. The estimated poverty rate is based on the disposable income, i.e., the reported income is adjusted for the number of household members and the age of the household members as suggested by the OECD (see, for example, Eurostat (2014a)). The first column contains the number of cases for the available case procedures of approach one and two. The second column contains sample sizes if all missing

Table 5: Estimated poverty rates from the PASS survey (with 95% confidence intervals reported in brackets).

Wave	n_{obs}	n_{imp}	Original data	Rounding adjustment	Nonresponse and rounding adjustment
Wave 1	10,214	12,791	17.29 (15.81;18.77)	16.35 (15.14;17.55)	16.60 (15.48;17.71)
Wave 2	7,311	8,428	16.91 (15.79;18.03)	16.98 (15.69;18.27)	16.39 (15.15;17.63)
Wave 3	8,169	9,534	14.27 (12.28;16.27)	15.40 (13.91;16.90)	15.66 (14.35;16.97)
Wave 4	6,538	7,845	14.89 (13.44;16.35)	14.61 (13.40;15.81)	14.81 (13.61;16.02)
Wave 5	8,623	10,232	16.34 (14.81;17.87)	15.75 (14.41;17.10)	15.82 (14.35;17.29)
Wave 6	8,267	9,508	15.95 (14.49;17.42)	16.27 (14.81;17.72)	15.78 (14.47;17.09)

or partially observed values are imputed. The results based on the original data without any adjustments are presented in the third column while the results for the multiply imputed true income accounting for rounding are included in column 4. The fifth column contains the results based on all data. All imputation results are based on $m = 10$ imputations. The 95% confidence intervals reported in brackets are based on bootstrap variance estimates. We used the normal approximation to compute the confidence intervals based on the estimated variances.

Generally, the impacts of the different adjustment methods are modest. Given the large amount of uncertainty in the estimates, the 95% confidence intervals mostly overlap. Still, there is some evidence that the impact from rounding is stronger than the impact due to (partial) nonresponse in most years. While the differences between the poverty rates based on the unadjusted point estimates and the estimates that account for the rounding (column three compared to column four) range from -1.13 to $+0.94$ percentage points, the differences between the adjusted estimates and the estimates that also account for the nonresponse (column four and column five) only range from -0.26 to $+0.59$ percentage points. The nonresponse adjustments only have a stronger impact in waves 2 and 6 in which the poverty rate hardly changes between the naïve direct estimate and the adjusted estimate. The smaller impact of the nonresponse is to be expected given that only 13–20% of the records are imputed to adjust for nonresponse compared to approximately 85% of the records that are imputed for rounding adjustments. Still, the differences in the poverty rates albeit small indicate that income is not missing completely at random and ignoring the nonresponse results in biased inferences.

5. Conclusions and Outlook

Obtaining reliable income information from surveys is notoriously difficult. Income is considered sensitive information and survey respondents often find it difficult to remember their exact income. In this paper we suggested a strategy to address two common potential sources of bias: nonresponse and rounding. Our multiple imputation approach tackles both problems simultaneously and provides a simple tool to incorporate interval information when making inference based on the collected data. The application to the PASS survey showed that adjusting for these two factors can have a direct impact on politically important measures such

as the poverty rate. We found that rounding has a higher impact on the results than nonresponse at least for our study. The changes in the poverty rates that we found in our empirical evaluation are modest although an increase of the poverty rate by 1.4% as observed for wave 3 of the PASS survey would likely cause some political discussions. We believe that the main reason for the relatively small changes lies in the robustness of the poverty measure which is based on the median of the income distribution. It would be an interesting area of future research to evaluate the impacts on less robust measures such as the income quintile share ratio (see, for example, Eurostat (2014b)) which computes the ratio of the 80% and the 20% quantile of the income distribution as a measure of income inequality.

Of course the adjustments proposed in this paper are based on several assumptions and it is important to critically review these assumptions. First, the correction methods are based on models and the underlying model assumptions need to be evaluated. Alternative models for the income distribution have been suggested in the literature. For example, Graf and Nedyalkova (2013) suggested to model the income distribution using the generalized beta distribution of the second kind. However, it is not straightforward to incorporate covariates in this model. Furthermore, we feel that our model evaluations in Section 4.1 indicate a good fit of the log-linear model for the conditional income distribution. Second, we assume that the income information is missing at random (MAR), i.e., the nonresponse can be explained by the variables included in the imputation model. This is a crucial assumption in most imputation models and this assumption can never be tested based on the observed data. We believe that the covariates in our model such as age of the respondent, deprivation index, or household size should help to explain the nonresponse in the data. However, if the MAR assumption does not hold, results from our imputation strategy will be biased and imputation models such as the non-ignorable models proposed in Little and Rubin (2002, Chap. 15) need to be considered. Finally, nonresponse and rounding might not be the only sources of bias in the data. Several studies found that individuals with low earnings tend to overreport their income while individuals with high income tend to underreport their income (see, for example, Pischke (1995)). Incorporating this additional measurement error into the adjustment strategy would be an interesting area of future research.

Acknowledgements: We thank one anonymous referee and the editor for thoughtful suggestions which helped to improve the paper. This work was partially supported by the DFG grants DR 831/2-1 and KI 1368/1-1.

References

Clementi F, Gallegati M (2005). "Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States." In A Chatterjee, S Yarlagadda, B Chakrabarti (eds.), *Econophysics of wealth distributions*, pp. 3–14. Milan: Springer.

Drechsler J, Kiesl H (2014). "Beat the Heap – An Imputation Strategy for Valid Inferences from Rounded Income Data." *IAB Discussion Paper 2/2014*, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany].

Eurostat (2014a). "Glossary: Equivalised Disposable Income - Statistics Explained (2014/11/07)." http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:Equivalised_disposable_income.

Eurostat (2014b). "Glossary: Income Quintile Share Ratio - Statistics Explained (2014/11/07)." http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:S80/S20_ratio.

Gelman A, Carlin J, Stern H, Rubin D (2004). *Bayesian Data Analysis*. Second edition. London: Chapman and Hall.

Graf M, Nedyalkova D (2013). "Modeling of Income and Indicators of Poverty and Social Exclusion Using the Generalized Beta Distribution of the Second Kind." *Review of Income and Wealth, online first*.

Heidelberger P, Welch P (1983). "Simulation Run Length Control in the Presence of an Initial Transient." *Operations Research*, **31**, 1109–1144.

Heitjan D (1994). "Ignorability in General Incomplete-Data Models." *Biometrika*, **81**, 701–708.

Heitjan D, Rubin D (1990). "Inference from Coarse Data Via Multiple Imputation with Application to Age Heaping." *Journal of the American Statistical Association*, **85**, 304–314.

Huttenlocher J, Hedges LV, Bradburn NM (1990). "Reports of Elapsed Time: Bounding and Rounding Processes in Estimation." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**(2), 196–213.

Little RJA, Rubin DB (2002). *Statistical Analysis with Missing Data*. Second edition. New York: John Wiley and Sons.

Liu J, Gelman A, Hill J, Su YS, Kropko J (2013). "On the Stationary Distribution of Iterative Imputations." *Biometrika*, p. (online first).

Manski C, Molinari F (2010). "Rounding Probabilistic Expectations in Surveys." *Journal of Business & Economic Statistics*, **28**, 219–231.

Pischke JS (1995). "Measurement Error and Earnings Dynamics: Some Estimates from the PSID Validation Study." *Journal of Business & Economic Statistics*, **13**(3), 305–314.

Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P (2001). "A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models." *Survey Methodology*, **27**, 85–96.

Rubin DB (1976). "Inference and Missing Data." *Biometrika*, **63**, 581–590.

Rubin DB (1978). "Multiple imputations in sample surveys." In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 20–34. Alexandria, VA: American Statistical Association.

Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Ruud PA, Schunk D, Winter JK (2013). "Uncertainty Causes Rounding: An Experimental Study." *Experimental Economics*, pp. 1–23.

Trappmann M, Gundert S, Wenzig C, Gebhardt D (2010). "PASS: A Household Panel Survey for Research on Unemployment and Poverty." *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften*, **130**, 609–622.

Wang H, Heitjan D (2008). "Modeling Heaping in Self-Reported Cigarette Counts." *Statistics in medicine*, **27**(19), 3789–3804.

Wang H, Shiffman S, Griffith SD, Heitjan DF (2012). "Truth and Memory: Linking Instantaneous and Retrospective Self-Reported Cigarette Consumption." *The annals of applied statistics*, **6**(4), 1689–1706.

Affiliation:

Jörg Drechsler
Department for Statistical Methods
Institute for Employment Research
Regensburger Straße 104
90478 Nürnberg, Germany
E-mail: joerg.drechsler@iab.de

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 44
April 2015

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2014-09-30
Accepted: 2014-11-21

Austrian Journal of Statistics
April 2015, Volume 44, 73–83.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i2.62

You Don't Teach, Students Learn: A Report on a Project on Statistical Literacy in Ireland

Eoin MacCuirc
CSO Ireland

Abstract

In 2007, with the aim of improving statistical literacy and effective use of statistics, the Central Statistics Office (CSO) in Ireland launched an Education Outreach Programme. To achieve these objectives, the CSO has fostered key academic partnerships at a national and international level. The CSO tailored projects developed under the umbrella of the Education Outreach Programme to target key audiences using statistics. This paper outlines a number of key lessons learned in the Irish Education Outreach Programme with illustrations drawn from the Irish experience to date.

Keywords: statistical literacy, outreach, strategic partnerships.

"Mathematics scares and depresses most of us, but politicians, journalists and everyone in power use numbers all the time to bamboozle us. Most maths is really simple - as easy as 2+2 in fact. Better still it can be understood without any jargon, any formulas - and in fact not even many numbers. Most of it is common sense, and by using a few really simple principles one can quickly see when maths, statistics and numbers are being abused to play tricks - or create policies - which can waste millions of pounds. It is liberating to understand when numbers are telling the truth or being used to lie, whether it is health scares, the costs of government policies, the supposed risks of certain activities or the real burden of taxes." (Blastland and Dilnot 2008)

1. Introduction

In 2007, the Central Statistics Office (CSO) in Ireland launched their outreach programme '*Investing in the Future*'. The CSO recognised it could not address all of the educational 'gaps' in statistical literacy and that each 'gap' might require a bespoke approach, so the Education Outreach Programme targeted three limited or specific cohorts:

1. Primary and Post-Primary Education;
2. Third level and Continued Professional Development; and
3. Media and Oireachtas (The National Parliament or Oireachtas in Ireland consists of the President and two Houses: Dáil Éireann (House of Representatives) and Seanad Éireann (the Senate)).

This paper outlines some simple lessons learned by the CSO since the programme began in 2007 to improve statistical literacy in Ireland.

In 2007, the Senior Management Board of CSO found itself deliberating the role of a national statistics office in improving statistical literacy. One of the five CSO corporate goals '*Increased awareness and effective use of our statistics*' (Central Statistics Office 2008) inferred that statistical education would be prudent. A proposal, before the Board, argued that CSO should develop an 'Education Outreach Programme' that collaborated with the Department of Education and Science (Department of Education and Science 2005) and other recognised national and international experts to improve statistical literacy. The concept of statistical literacy used in the proposal leaned heavily on Smith's presentation about "Data Literacy" at ACCOLEDS conference in Vancouver, Canada in 2002, which stated that statistical literacy should comprise being able to:

- *Understand and interpret statistical data;*
- *critically evaluate statistical information and data-related arguments;*
- *use the information in context of daily life; and*
- *discuss or communicate one's reactions.*

The proposal argued that the programme should be designed to promote the statistical outputs of the office. This would be done by '*Creating real life projects that enhance the learning process and nurture the life skills needed in our knowledge based society, while supporting the Irish education system in developing its youth as future policy makers, entrepreneurs and statisticians'*. The Board endorsed the proposal to establish an education outreach programme, viewing it as a long term investment.

2. Lessons learned

2.1. Many hands make light work – collaboration is key

The Central Statistical Office Ireland (CSO) is a National Statistic Office; its areas of expertise are in gathering, producing, analysing and disseminating information. The CSO is not prima facie an institution that educates. Yet, it is important that the information provided by the CSO is used correctly. The CSO favours more educated users, using its information. Hence, the development of the CSO Education Outreach Programme.

To improve statistical literacy in Ireland the CSO has collaborated with many organisations nationally and internationally (a detailed list of collaborations is outlined in Appendix 1). Cooperation and collaboration across a number of different projects has introduced CSO to new networks of talented, enthusiastic and committed professionals with aims that overlap with our own. By building on this enthusiasm and harnessing these diverse skills, CSO has achieved so much more than it could do on its own.

To get a picture of this collaboration, take for example, the CSO sponsored apps4gaps competition. The CSO coordinates and part funds the project having overall ownership of the apps4gaps competition. Insight (https://www.insight-centre.org/), maintains the website and provides technical backup.
SFI Discover (http://www.sfi.ie/discover-science-engineering-dse/) and Open Government Partnership Ireland (http://www.ogpireland.ie/) provide funding. The Department of Education and Skills (DES) (Department of Education and Science 2005) and the Department of Education Northern Ireland (DENI), write to every school in Ireland to encourage school participation. Coder Dojo (https://coderdojo.com/), CensusAtSchool (http://www.ncca.ie/en/AboutUs/) and Project Maths (National Council for Curriculum

and Assessment 2012) encourage participation and support students to enter the project. All organisations promote the competition.

Collaboration has been critical adding to the strength and depth, of the reach, of the CSO Education Outreach Programme.

2.2. Engage and make it engaging

Sometimes, there is a tendency to make statistical education complicated and tedious. Having been a secondary school mathematics and accountancy teacher, a founding parent of a Waldorf Steiner School, an education and activities coordinator in a homeless shelters, a teacher of diverse topics from kinesiology, to organic gardening and permaculture design, a member of the Cork City UNESCO Global Network of Learning Cities implementation group, besides my work in the CSO Education Outreach Programme, I have considerable experience of being both a teacher and a student in many disciplines. As a student, my most profound learning experiences have come from being engaged in a subject by learned, passionate, open and inspiring teachers. From contract law, through building swales, to the intricacies of the somatid cycle, I have been engaged and touched by the most surprising topics. In the presence of a great teacher, I have discovered that almost any subject can be enthralling. Palmer (2009) describes this teaching beautifully in his book The Courage to Teach.

Good teaching cannot be reduced to technique; good teaching comes from the identity and integrity of the teacher in every class I teach, my ability to connect with my students, and to connect them with the subject, depends less on the methods I use than on the degree to which I know and trust my selfhood – and am willing to make it available and vulnerable in the service of learning.

Commitment 8. Enhancing Quality in Learning from the United Nations Educational, Scientific and Cultural Organisation report on the International Conference on Learning Cities states (UNESCO 2014):

In many cities, there is a disparity between the numbers of people participating in education and learning and those who succeed in mastering relevant, portable skills and competences. Quality is, therefore, of utmost importance.

It goes on to state:

In developing learning cities, we attach great importance to enhancing quality in learning by: (here I have picked two of the five points)

• promoting a paradigm shift from teaching to learning, and from the mere acquisition of information to the development of creativity and learning skills;

• fostering a learner-friendly environment in which learners have, as far as practicable, ownership of their own learning;

I have found the same to be true of many cases of mathematics and statistical education. Many students go through the education system, but, few achieve mastery in mathematics and statistics. I have a number of suggestions in this regard.

2.3. Keep it simple

In all cases, whether speaking to primary school students or government ministers, it is important that the information imparted is understood. In education outreach you don't teach, students learn. You are the presenter of the message. It is important that the message presented is at the appropriate level and presented in a way that make it simple to understand. When in doubt, one should simplify, so everybody learns.

In *The Tiger That Isn't: Seeing Through a World of Numbers* (Blastland and Dilnot 2008), the authors state that size is personal and that big numbers should be made clear to everybody.

For example, in presenting statistics involving big numbers to a group, one could ask the participants what is a million? Agree that a million is the number 1,000,000 and show the number. Then ask the participants how long would it take to count to one million? If it takes a second to count off each number, how long would that take, in hours, days? One could take some guesses from the participants and then inform the participants it takes over eleven and a half days to reach one million. One could follow this up by asking what is a billion? Agree a billion is the number 1,000,000,000, again showing the number. Asking again how long would it take to count to one billion? If it takes a second to count off each number, how many hours, days would it take? Again one could take some guesses from the participants. It takes over thirty-one and a half years. Participants are then clear that a million and a billion are vastly different numbers. Asking participants to picture a baby boy eleven days old and a thirty-one year old man/woman and the span of time they have lived. It is totally different. People get a clear picture of the difference between the numbers, so when data is presented in millions and billions, now, there is real understanding.

Big numbers are important, so I suggest, one should present them in a simple way that everybody can understand.

A great teacher, through their mastery, presents complex subject matter simply, in a way that students can learn.

2.4. A good example

A good example facilitates learning. In keeping with the billion theme above one can refer to the United Nations Development Report (Malik 2014).

One could ask participants can they estimate world hunger? Recent media headlines state two billion poor and one billion hungry. The FAO report on Global Food Losses and Food Waste (Gustavsson and Sonesson 2011) estimates 1.3 billion tons of food loss and food waste annually in the world.

These two pieces of information can lead to lots of questions, discussion and learning. Why are one billion people hungry? Why with all this food waste and food loss are people going hungry? This in turn can then lead to a fruitful discussion on policy making, the complexity of global issues, the value of different data sources, global awareness and informed policy making. Having a good example that is easily understood and engaging can enhance learning. Big issues are ideal in this context.

A good example is a great tool for teaching and learning.

2.5. Make it personal

Bringing a personal perspective promotes deeper learning. Engaging with CensusAtSchool data creates this opportunity. Here (http://www.censusatschool.ie/images/phases/phase_11_questionnaire_en.pdf is an interesting example from CensusAtSchool Phase 11 (2011/2012). The question (9b) asked "Please state what you had for breakfast this morning". Respondents ticked a selection of breakfast items or "I did not have breakfast".

Looking at the national CensusAtSchool data broken down by sex and age, the percentage of boys who did not have breakfast remained practically the same throughout the secondary school cycle. However, looking at the girls' data the percentage of girls not having breakfast spiked, at 15 years old, before going back to normal. Of the 5,224 records, 15.8% of the girls as opposed to 10.1% of the boys did not have breakfast.

Some engaging questions might be: Why would girls at a certain age stop eating breakfast? Does skipping breakfast assist in weight loss? Is there evidence that eating breakfast combats obesity and diabetes in young people? What does the data for our school/class show?

Again, what makes CensusAtSchool interesting for students is that they are looking at data about themselves and their peers. Students can see patterns in the data for their class/school

and for a randomised sample of the data from other schools. The Database Interrogation Tool (http://www.censusatschool.ie/en/get-data/datatool) and the Random Data Selector (http://www.censusatschool.ie/get-data/random-data-selector) are provided to assist in the learning process.

A video (http://vimeo.com/110872135) has been produced to assist teachers and students in using this tool. The video shows how engaging CensusAtSchool can be both for teachers and students.

One should be careful when dealing with potentially sensitive topics with younger children e.g. student weight has not been asked on the Irish CensusAtSchool questionnaire to date.

2.6. Make it fun

Discussing data, it can be fun to get participants to actively engage. Sometimes data can be surprising, the birthday paradox is a good example. Eastaway (2008) gives the example of a wedding party where there are about 50 guests old and young. The question "I wonder what the chance is that there are two people in the wedding party who have the same birthday as each other, not the same birth date, just the same birthday, like 5 May or something". The answer is a surprise to most students. With 50 people at the party then the chance of a birthday coincidence is 97%.

From the birthday paradox the chance of a matching birthday with 20 people in the group is 43% going up to 70% at 30 people and 97% at 50 people; the 50/50 chance is 23 people.

For a smaller group, one could bet the group that two people have a birthday within one day of each other. So, if one group member had a birthday on January 8^{th}, a match would happen, if someone else in the group had a birthday on January 7^{th}, 8^{th} or 9^{th}. 13 people in this case are needed for a 50% chance of a match and 28 people for a 95% chance of a match. A good bet in a classroom situation.

Most students, unaware of the paradox, are willing to place a bet that two people don't share the same birthday. Having won the bet, this can be a fun way of introducing probability.

3. Tools that enhance the education outreach process

3.1. Competitions

The CSO's experience with the John Hooper Medal for Statistics and the apps4 gaps competition (www.apps4gaps.ie has been documented in a recent CSO paper presented at the International Conference on Teaching Statistics of the International Association for Statistical Education (IASE) in Arizona in 2014. These competitions have proved very successful as a vehicle for engaging students and the public and in bringing statistics to a wider and younger audience. The winner of the John Hooper Medal for Statistics represents Ireland at the biannual International Statistical Literacy Project (ISLP) (A list of ISLP participating countries is detailed in Appendix 2, giving an indication of the countries' engagement in the competition).

3.2. CensusAtSchool

CensusAtSchool is where the CSO education outreach programme began in 2007 and it is going from strength to strength. The Irish 2014/2015 CensusAtSchool questionnaire (see http://www.censusatschool.ie/images/phases/2014-15-questionnaire.pdf) celebrates family in keeping with the UN International Year of the Family Farm 2014 and the twentieth anniversary of the International Year of the Family in 1994. The CensusAtSchool website is a vehicle to promote much of CSO's education outreach work at secondary school level, with over 100,000 hits on the questionnaires (http://www.censusatschool.ie/en/take-part/

questionnaires) alone. CensusAtSchool questionnaire responses by country/region are detailed in Appendix 3, giving an indication of international participation. The figures in Appendix 3 are an update on the figures in (Davies 2011) about the further uses of CensusAtSchool and the ExperimentsAtSchool projects, to improve collaborative teaching and learning, statistical thinking and literacy for learners and teachers.

3.3. Professional diploma in official statistics for policy evaluation

The *Professional Diploma of Official Statistics for Policy Evaluation* (see http://www.ipa.ie/index.php?lang=en&p=page&id=363) was launched in 2012 by the CSO in cooperation with the Institute of Public Administration (IPA) (Institute of Public Administration 2014) and University College Dublin (UCD). The diploma is a one-year, part-time programme and is targeted at the public service, specifically those who use (or should use) data to formulate or assess policy. The course is designed as a practical 'hands-on' course where students are shown how to access and interpret official statistics. Considerable emphasis is also placed on presenting and visualising statistics so that useful policy relevant information can be derived. This course has proven very successful and is now in its third year.

The diploma introduces students to important Irish and international official statistics that will help them better understand the structure and trend of Irish and international economies, societies and environments and their respective inter-dependencies.

The diploma is not designed as a quantitative methods or technical statistics course but rather to teach an appreciation of statistics and how they can be used to find and present key messages. The aim of the diploma is to encourage sound evidence based policy making and sound evaluation of existing policy. It is hoped the course will take the fear and mystery out of official statistics.

27 people successfully graduated from this course in 2013, 36 graduated in 2014 and over 50 students are currently registered for the 2014-2015 academic year.

The statistical educators in New Zealand continue to inspire on what is possible in this field. Forbes (2014) outlines the great work that continues to be done in New Zealand. Finally a really exciting development is Chris Wild's (Wild 2014) ground breaking work, culminating with the University of Auckland launching Data to Insight: An Introduction to Data Analysis, a Massive Online Open Course (MOOC).

4. Seminar series

The CSO currently runs two seminar series: A Business Statistics Seminar series and an Administrative Data Seminar series along with occasional ad-hoc seminars (see for further details on these seminars: http://www.cso.ie/en/newsandevents/eventsconferencesseminars/). The philosophy underpinning these series has four central pillars:

- to make users and potential users aware of all the data already available ;
- to demonstrate how these data could be used by providing case studies or illustrations of analyses;
- to improve our relationships and develop a network of researchers, policy makers, academics and other stakeholders; and
- to market new products or datasets.

The seminar series has been running since 2008 and continues to promote awareness and use of statistics.

The seminars provide a forum where CSO staff, data users, respondents and policy makers can meet face to face to discuss matters of mutual interest. The forum showcases new and

interesting work by CSO staff, but also includes work from other researchers and policy experts. All are encouraged to raise issues about data and demonstrate existing and new ways that statistics can enrich us all. The presentations are hosted on the CSO website. Attendance varies depending on content, for example, 93 people attended the latest Administrative Data Seminar in February 2014, representing a broad range of data users and suppliers from academia, government departments and agencies (see the piechart in Figure 1 for attendance by guest type).

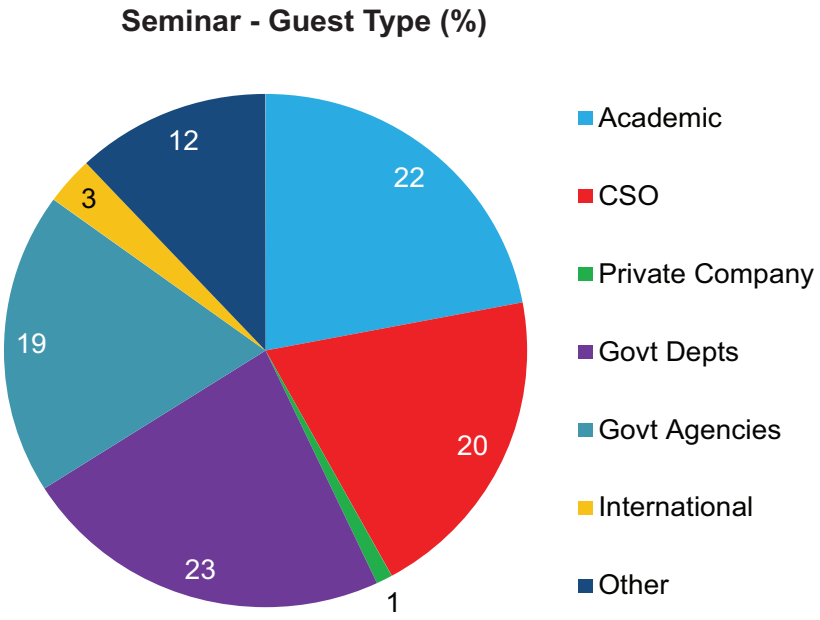


Figure 1: Attendance by guest type at the seminar series.

4.1. Stories, video and interactive visualisation

In October 2014, the CSO organised an in-house training course on visual storytelling for professionals. Helen Kuyper of 24/7 Storytelling.com led the group in exploring how to tell stories with numbers. Chris Wild's MOOC begs the question: are there newer ways to engage with numbers and statistical literacy. A large number of National Statistics Institutes are exploring the issue of visual engagement. There are countless examples of infographics[1], video[2] and interactive visualisation[3].

Census Charlie (see Figure 2), was a character in the CSO Census 2011 story, part of the education resources for primary school teachers.

People are consuming information more visually through a myriad of devices and through an ever expanding list of media, this presents opportunities to improve engagement with data and

[1]UK Labour Statistics: http://www.ons.gov.uk/ons/rel/lms/labour-market-statistics/may-2013/sty-employment.html, New Zealand Census: http://www.stats.govt.nz/Census/2013-census/profile-and-summary-reports/2013-census-infographic-chch.aspx, Irish Transport Omnibus: http://www.cso.ie/en/releasesandpublications/ep/p-tranom/transportomnibus2013/#.VKLAgs4A

[2]Hans Rosling (2006) TED, Monterey California: https://www.youtube.com/watch?v=hVimVzgtD6w, Australia Census Spotlight (2011): http://spotlight.abs.gov.au/Flash/, Irish Census (2011) https://www.youtube.com/watch?v=lrYN3yhbZXY

[3]Family Spending UK (2013): http://neighbourhood.statistics.gov.uk/HTMLDocs/dvc203/index.html, Statistics Sweden, Statistical Atlas (2012): http://www.scb.se/Kartor/Statistikatlas42KN201205English/index.html#story=0, Job Churn Ireland (2013): http://www.cso.ie/visual/pyramids/jobschurnexplorer/n2pyramid/visualise_by_age_and_gender.html

Figure 2: Census Charlie – a character in the CSO Census 2011 story

in turn improve statistical literacy. Campos (2013) and the Portuguese Statistical Society's *Exploristica* itinerant exhibition, is a very interesting project in this regard. The CSO plans to translate *Exploristica* for an Irish audience in 2015.

5. Conclusion

The outreach programme '*Investing for the Future*' has two broad aims: (1) to promote the availability and appropriate use of official statistics and (2) to improve the general standard of statistical literacy among all cohorts of CSO data users. This programme is viewed as a long term investment, but an investment worth making, if it encourages the rational and sensible use of information to formulate and assess public policy and life decisions.

There are a number of key conclusions from the CSO outreach programme. Firstly, collaboration is critical to success. Secondly, different people and age-groups learn in different ways - no single solution will reach everyone.

From its humble beginnings in 2007 the programme has gradually expanded. New and interesting partnerships and collaborations have been formed. New courses and seminars, new competitions, new websites and new tools have been created, encouraging people to engage with CSO data and official statistics. The programme continues to evolve.

The three most important lessons learned to date are engage, engage, engage.

References

Blastland M, Dilnot A (2008). *The Tiger That Isn't: Seeing Through a World of Numbers*. Profile books. ISBN 9781846681110.

Campos P (2013). "Exploristica - Adventures in Statistics: a New Itinerant Exhibition for Teaching and Learning Statistics." URL http://www.statistics.gov.hk/wsc/CPS103-P13-S.pdf.

Central Statistics Office (2008). *Statement of Strategy, 2004-2006 : Statistics for a Modern Ireland*. Central Statistics Office Ireland.

Davies N (2011). "Developments of AtSchool Projects for Improving Collaborative Teaching and Learning in Statistics." *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, **27**(3), 205–227.

Department of Education and Science (2005). *A Brief Description of the Irish Education System*. Communications Unit, Department of Education and Science, Ireland.

Eastaway R (2008). *How Many Socks Make a Pair?: Surprisingly Interesting Everyday Maths*. JR Books Limited. ISBN 9781906217594.

Forbes S (2014). "The coming of age of statistics education in New Zealand, and its influence internationally." *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, **22**(2), 1–19.

Gustavsson J, Sonesson U (2011). "Global Food Losses and Food Waste." *Technical report*, FAO, Rome, Italy.

Institute of Public Administration (2014). "Annual Report 2013." *Technical report*, Institute of Public Administration, Dublin, Ireland.

Malik K (2014). "Human Development Report 2014. Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience." *Technical report*, United Nations Development Programme (UNDP), New York, USA.

National Council for Curriculum and Assessment (2012). "PROJECT MATHS. Responding to Current Debate." URL http://www.projectmaths.ie/about/.

Palmer P (2009). *The Courage to Teach: Exploring the Inner Landscape of a Teacher's Life*. Wiley. ISBN 9780470469279.

UNESCO (ed.) (2014). *Lifelong Learning for All: Inclusion, Prosperity and Sustainability in Cities. International Conference on Learning Cities*. UNESCO Institute for Lifelong Learning, Beijing, China. ISBN 978-92-820-1184-3. Conference Report.

Wild C (2014). "Middleware for Middle Earth." In *ICOTS9*.

Appendices

Appendix 1 – The CSO education outreach programme. A list of collaborations.

To improve statistical literacy in Ireland the CSO has collaborated with many organisations nationally and internationally. In primary and post primary education, the CSO has collaborated with:

(a) the Department of Education and Skills (DES) (Department of Education and Science 2005) and the Department of Education Northern Ireland (DENI), to promote and develop statistical literacy.

(b) Project Maths (National Council for Curriculum and Assessment 2012) – a project team dedicated to revising the post primary mathematics curriculum, changing what students learn in mathematics, how students learn mathematics and how students mathematics skills are assessed. Project Maths work with CSO on all our post primary projects

(c) the Professional Development Service for Teachers, in particular the PDST Technology in Education team, to promote the role of technology in mathematics and statistical literacy

(d) CensusAtSchool (http://www.censusatschool.org.uk/international-projects) internationally and the Royal Statistical Society Centre for Statistical Education (RSSCSE), who helped the CSO to set up the Irish CensusAtSchool website (www.censusatschool.ie) with our Irish partners: Project Maths, PDST Technology in Education and the National Council for Curriculum and Assessment

(e) HEAnet Ireland's National Education and Research Network, who act as web hosts for the Irish CensusAtSchool data

(f) CoderDojo (https://coderdojo.com/about/), an open source, volunteer led, global movement of free coding clubs for young people.

At third level and for Continued Professional Development the CSO has collaborated with:

(a) The Institute for Public Administration, where The Professional Diploma of Official Statistics for Policy Evaluation was launched in 2012

(b) University College Dublin, who awards a level 8 National Framework Qualification special purpose diploma worth 20 credits to those who successfully complete the diploma

(c) Dublin Castle, which currently hosts the CSO seminar series (http://www.cso.ie/en/newsandevents/conferencesseminars/)

(d) the International Statistical Institute and other international statistical bodies including the International Association of Statistical Education (IASE), the International Association of Official Statistics (IAOS) and the International Statistical Literacy Programme (ISLP).

(e) Insight: The Centre for Data Analytics has worked with the CSO on our linked open data projects and the apps4gaps competition

(f) The Science Foundation of Ireland Discover Programme (http://www.sfi.ie/discover-science-engineering-dse/), which has promoted and part funded the apps4gaps competition (www.apps4gaps.ie)

(g) The Open Government Partnership Ireland (http://www.ogpireland.ie/), who part fund the apps4gaps competition and promote CSO open data initiatives.

Appendix 2 - international statistical literacy project (ISLP) – poster competition (2012-2013). A list of participating countries

Here are all the countries that organised the ISLP poster competition 2012−2013. The numbers represents the participating students from each country.

Argentina	46	Indonesia	2	Russia	448
Australia	41	Ireland	1,281	Slovakia	21
Bangladesh	12	Italy	98	South Korea	16
Bhutan	5	Japan	3,849	Spain	2
Brazil	20	Kazakhstan	5	Sweden	85
Bulgaria	3	Kuwait	10	Togo	3
Czech Republic	530	Mexico	11	Ukraine	18
Finland	233	New Zealand	38	United Arab Emirates	5
Hungary	65	Poland	115	United Kingdom	95
India	3	Portugal	123	USA	8

Source: International Statistical Literacy Project (ISLP)

http://iase-web.org/islp/Poster_Competition_20122013.php?p=Participating_countries

5.1. Appendix 3 – CensusAtSchool questionnaire

Responses by country/region to 2013.

Country/Region	Years	Responses	Responses Remarks
United Kingdom	2000 - 2013	247,983	Run annually
South Africa	2001, 2009	1,500,152	Partial sample (over 3m took part in 2001)
Queensland	2001 - 2003	30,789	Became part of Australia-wide project
New Zealand	2003 - 2013	121,606	Run every two years
Canada	2003 - 2013	206,756	Statistics Society of Canada hosting from 2012
Australia	2005 - 2013	225,923	Run four times
South Australia	2003	21,557	Became part of Australia-wide project
Ireland	2009 - 2013	22,637	Run annually
Japan	2009 - 2013	635	13 schools
USA	2010 - 2013	7,967	39 States, 364 registered teachers
Total		**2,386,005**	

Source:

International Centre for Statistical Education, Plymouth University, United Kingdom icse@plymouth.ac.uk

Affiliation:

Eoin MacCuirc
Central Statistics Office, Ireland
E-mail: eoin.mccuirc@cso.ie

News and Announcements

Matthias Templ

Arbeitslosenzahlen über Jahre falsch berechnet! So titelte die Kronen Zeitung am 20.03.2015. Eine Negativschlagzeile nach der anderen, obwohl die Schätzungen verbessert werden. Es stellt sich grundsätzlich die Frage inwieweit die Verwendung von statistischen Methoden zur Non-Response Analyse, Datenintegration, Record Linkage und Kalibrierung zur Verbesserung von Schätzungen den Medien vermittelt werden können?

Last year, IASC submitted a proposal to the International Statistical Institute (ISI) to receive financial support from the World Bank Trust Funds for Statistical Capacity Building (WBTFSCB). One of the proposed projects, an R course to be held in Tanzania by Peter Filzmoser and Matthias Templ in Dar Es Salaam, was successfully selected for funding. The course was given on February 12-16, 2015. For a nice picture and further information, have a look at `http://www.iasc-isi.org/announcements/279`.

According to the cover image at the title page: the analysis of sound and instruments is an applied area of statistics. Our special guest editor – Gerhard Nachtmann – can play a song on it, see e.g. his contributions *"Experimental demonstration of the effect of wall vibrations on the radiated sound of the horn and a search for possible explanations."*, *"Bell vibrations and radiated sound of brass wind instruments – is there an audible correlation?"* or *"More experimental evidence favouring the hypothesis of significant wall vibration influence on radiated horn sound"*. Practical applications are given almost weekly, e.g. with BOKU Blaskapelle (`http://blaskapelle.boku.ac.at/`) and their bandmaster Gerhard Nachtmann.

The former Department of Statistics and Probability Theory at the Vienna University of Technology have been merged with the Institue of Mathematical Economics from the same university. The new institue is now called Institute of Statistics and Mathematical Methods in Economics. The new institute contains seven research groups, whereas three of them are dedicated to statistics:

- Computational Statistics (head of research unit: Peter Filzmoser)
- Mathematical Stochastics (head of research unit: Karl Grill)
- Applied Statistics (head of research unit: Rudolf Dutter)

More information can be found on `https://swm.tuwien.ac.at/`.

ISSN 1026–597X

Contents