Journal of Official Statistics
vol. 34, 4 (diciembre 2018)

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 797–809, http://dx.doi.org/10.2478/JOS-2018-0040

Preface

1. New Techniques and Technologies for Official Statistics

Official statistics are steadily moving away from the traditional "one survey – one product" paradigm, with growing efforts to integrate the data sources at hand and include new data sources, be it administrative data or altogether new "big data", such as web scraped data or mobile phone data. Moreover, there is an increased interest in and awareness of the need for reliable processes and systems underpinning an agile production of official statistics and the importance of both producing official statistics and effectively disseminating them with methods that ensure that they reach their intended users.

Research in official statistics, being close to the cutting edge, is a leading indicator of this trend. In recent years, interest in the conference series on New Techniques and Technologies for Statistics (NTTS), organised since 1992, has increased considerably, and NTTS has emerged as a major official statistics research forum. As can be seen in Figure 1, the most recent conference, NTTS 2017, which took place in March 2017 in Brussels, saw a record level of participation. Figure 2 illustrates the diversity of stakeholders brought together by the conference, which included official statisticians (half of the participants represented national statistical institutes, regional statistical institutes, and Eurostat), other parts of the public sector (one-fourth of the participants represented European Union Institutions and national authorities), academia (15% of participants) and the private sector (8%).

Building on the successful experience of the Journal of Official Statistics (JOS) special issue with articles from NTTS 2013 (see Karlberg et al. 2015), the NTTS 2017 Scientific Committee reached out to the JOS editorial board in 2016 to explore the possibility of a JOS special issue based on articles from NTTS 2017. This special issue of the JOS, in which eight articles are presented, represents the final outcome of a highly selective screening and peer review process. First, we would like to thank all authors, including authors of the numerous articles that were unsuccessful in the reviewing process, for considering the JOS as a platform for their work. We would also like to emphasise that this special issue would not have been possible without the 60 referees who kindly agreed to review one (or frequently more versions) of the manuscripts. The manuscripts all benefitted from these constructive reviews, improving greatly in quality.

In this context, we would like to express our deep sorrow at learning that Rein Ahas, Professor in Human Geography at the University of Tartu, passed away unexpectedly on 18 February 2018 at the age of 51. Rein served with us on the NTTS 2017 Scientific Committee, and the very last service that he rendered in this regard was a review of one of the manuscripts for this special issue. In the review, he emphatically insisted on the proper application of scientific standards, constructively providing concrete advice on

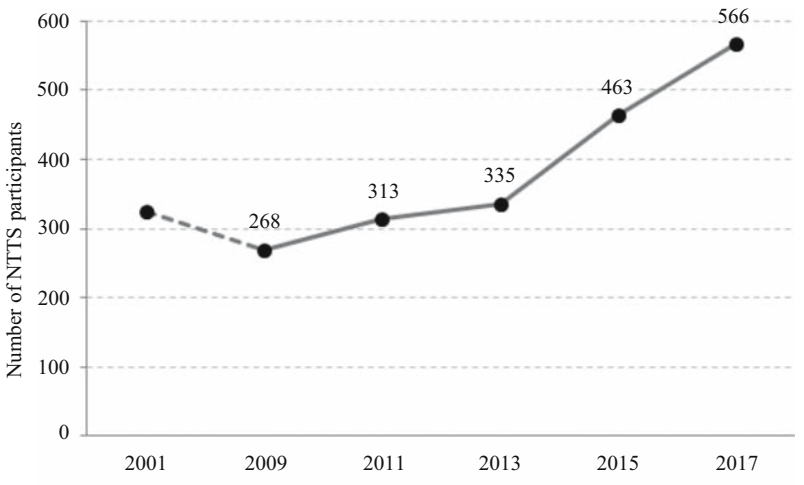

Fig. 1. Number of effectively attending NTTS participants (registrations adjusted for no-shows). Source for 2009-2017: the registration system of each respective conference. For 1992-2001, the sole data source identified to date is a conference report on NTTS 2001.

how the authors could do so in practice. With this, we would like to acknowledge the exceptional service that Rein has given to the NTTS conference series, and offer our deepest condolences to his family.

Appropriate management of core statistical, methodological and IT processes is a challenge that many statistical offices face. Section 2 presents the first two articles of this special issue, which both deal with innovative solutions tackling this fundamental aspect.

In Section 3, we present three articles that all deal with the integration of multiple sources in official statistics, with an emphasis on administrative and registry data, challenges regarding reconciliation (matching and linkage) and the quality control that this entails.

While regular official statistics production based on new big data sources is still in its infancy, there are many experiments underway to investigate its feasibility. Two articles quite far apart on the exploration/application scale are presented in Section 4. One of them has a specific official statistics concept in mind, whereas the other article is of a more exploratory nature.

The eighth and final article of this special issue, presented in Section 5, deals with an innovative approach to dissemination, by means of natural language, underpinned by fuzzy logic. In Section 6, we offer some concluding remarks, focusing on features common to the articles of this special issue.

2. Agile Processes for Statistical Offices

In the first article of this special issue, Salgado, Esteban, Novás, Saldaña and Sanguiao advocate an agile approach to data organisation and process design based on functional modularity. They underline that while it is common practice to see these principles applied in (a) the design of software for statistical production, they are rarely applied consistently when it comes to two other aspects of statistical production, (b) statistical production

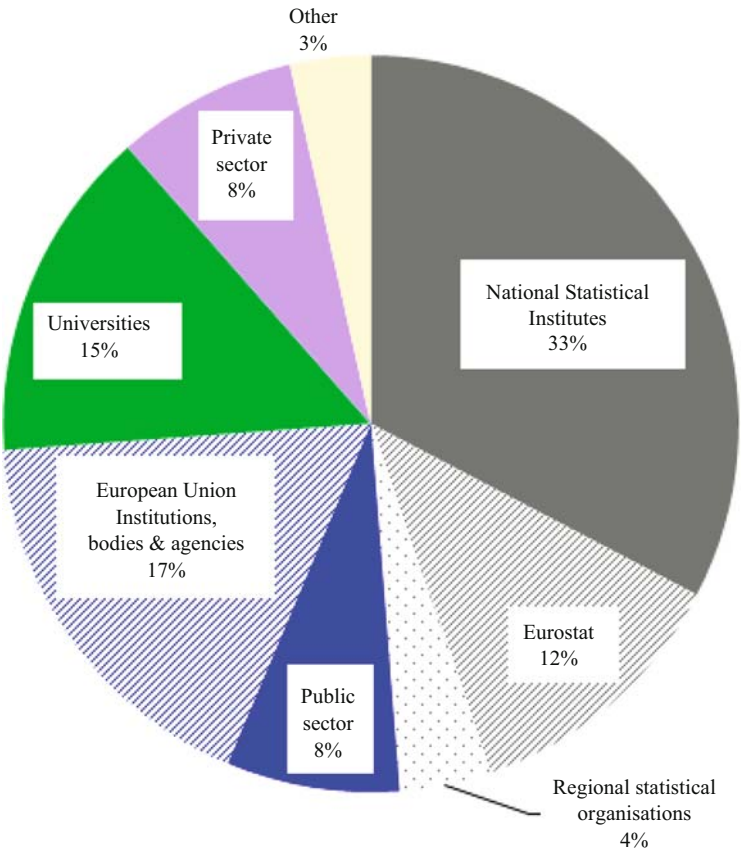

Fig. 2. Participants at NTTS 2017 by type of organisation. Source: registration database (n = 566). Participant categorisation was done manually.

metadata, and (c) statistical methodology. This unfortunate state of affairs is by no means unique to official statistics; in many organisations, stakeholders tend to ignore non-IT aspects of processes, and (implicitly) delegate or outsource key process design decisions to IT services, instead of requiring the business process owners to set up their processes. This results in a lopsided setup, that is, well-defined IT processes that might provide "the right answer to the wrong question", as they may be supporting vaguely or suboptimally defined business processes. Salgado and coauthors argue that the same principles must be applied to *all* of the three aforementioned aspects (a), (b) and (c). This is especially the case, since (as they convincingly argue) official statistics production is a *complex system* (Saltzer and Kaashoek 2009), composed of "(i) a large number of components, (ii) a large number of interconnections between these components, (iii) many irregularities in these interconnections, since the lack of regularity is indeed the rule rather than the exception, (iv) a long description of the system and its related management [. . .] and (v) a team of designers, implementers, and/or maintainers to handle the system".

Their proposed approach is based on object-oriented and functional computation paradigms. "The former comprises a standardised key-value pair abstract data model, where keys are constructed by means of the structural statistical metadata of the

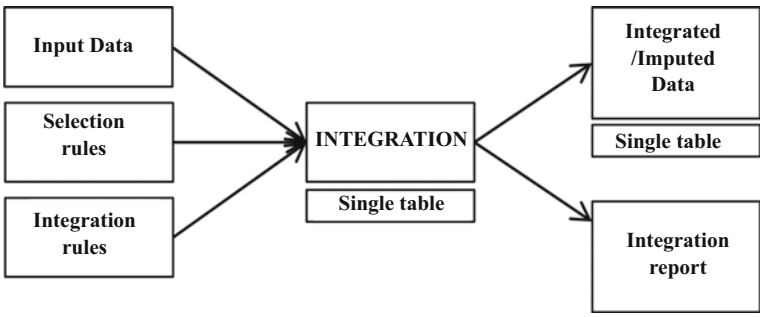

Fig. 3. Abstract information model of a general integration process. Source: Cesaro and Tininini.

production system", whereas the latter relies heavily on "the principles of functional modularity (modularity, data abstraction, hierarchy, and layering) to design production steps." Salgado and coauthors conduct a proof of concept (for the editing process step) in the field of Short-Term Business Statistics, and draw several valuable conclusions on various aspects seeking to emphasise lessons learned that could be transferrable to other statistical offices. In view of their experience, they advocate a non-linear "spiral approach" to software development, and make their case for a "change of mindset to conceive software as constantly evolving instead of as a closed definitive tool [as this] is necessary to industrialise and modernise the statistical production". The authors give practical hints, such as "wrapping" in various ways (developing a simple Excel file to serve as the "front end" for generating XML code; complying with a "SAS only" office policy by running SAS macros that execute R scripts in batch). Importantly, they make all code available in the public domain via GitHub (Esteban et al. 2017).

In the second article of this special issue, Cesaro and Tininini propose a service-oriented architecture (SOA) following the style of lightweight basic integration, and describe how this has been successfully deployed for validation (as well as for prioritisation among multiple sources) for the Italian Statistical Business Registers (SBR). They thoroughly investigate key design choices that drive performance.

While these two special issue articles have different focus, they share many features. Both of them rely on, and discuss, core artefacts developed by the official statistics modernisation community, such as the Generic Statistical Business Process Model (GSBPM), the Generic Statistical Information Model (GSIM) and the Common Statistical Production Architecture (CSPA). They are in favour of high-granularity processes/services (although the level of granularity might differ between the two articles) that are stateless (depending only on the input they receive). They follow the *active metadata* paradigm, letting metadata (including rules, as illustrated in Figure 3) be a part of the input to processes rather than hard-coded into the processes themselves. In short, both articles present solutions that are conducive to shareability and agility.

3. Consolidating and Reconciling Multiple Sourses

In the words of Vâju and Meszaros (2018), statistical authorities "need to produce data faster in a cost effective way, to become more responsive to users' demands, while at the

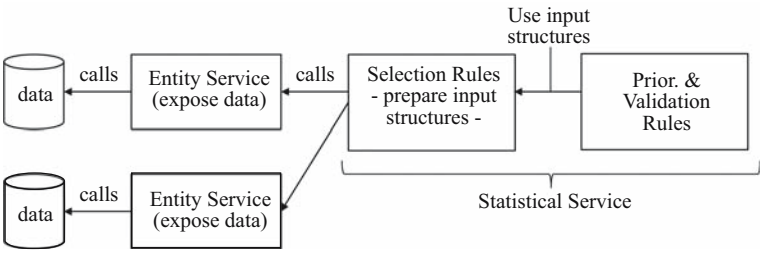

Fig. 4. Logical SOA architecture with Entity services and statistical services: selection rules adapt input data for subsequent processing for prioritization and validation purposes. Source: Cesaro and Tininini.

same time providing high quality output. One way to fulfil this is to make more use of already available data sources, and in particular administrative sources, most typically used in combination with other sources." A key objective of the service developed by Cesaro and Tininini is to support activities fulfilling this objective, by integrating multiple sources, as illustrated in Figure 4. The consolidation and reconciliation of multiple sources is also the topic of three other articles of this special issue, and we will present those in this section.

When one speaks of a set of "administrative data", the typical case that springs to mind is that of a single, monolithic, governmental authority at national level with one common set of rules for data collection. However, in many cases, administrative data are provided by decentralised autonomous administrations (for instance, municipalities that collect data on their inhabitants). Thus, consolidation of multiple sources (each reporting entity being one source) has to take place in order to arrive at one (e.g. national) data set for a certain domain. In the third article of this special issue, van Delden, van der Laan and Prins address this situation – more specifically, how to deal with the heterogeneity in reporting that this can entail; they present a method (illustrated in Figure 5) to detect under- and overreporting by data suppliers for "decentralised administrative data" for the case of change estimates. The method is successfully applied to a case study with administrative hospital data, and the authors conclude by setting out a number of steps concerning adaptations and extensions needed to deploy the methods in official statistics production. While the authors specifically treat the case of decentralised autonomous administrations, the scope of applicability might be even wider. For instance, local offices of a national administration might also have *de facto* developed their own administrative traditions, even if there is a common set of rules nominally applicable to the entire authority.

A reconciliation of multiple sources along another dimension takes place when multiple lists for (largely) the same set of units are brought together. This is the topic of the fourth article of this special issue, wherein Di Consiglio and Tuoto propose a method that is applicable in cases where the goal is to measure the size of a population (partially) enumerated in different lists. Their multiple lists linkage procedure tackles the problem of adjusting population estimates in the presence of linkage errors. With a distribution of estimates close to the one that could be expected without any linkage errors, their proposed class of estimators performs better than the alternatives investigated.

Whereas Di Consiglio and Tuoto consider the reconciliation of more or less complete data sets, each normally covering a very large part of a target population under study,

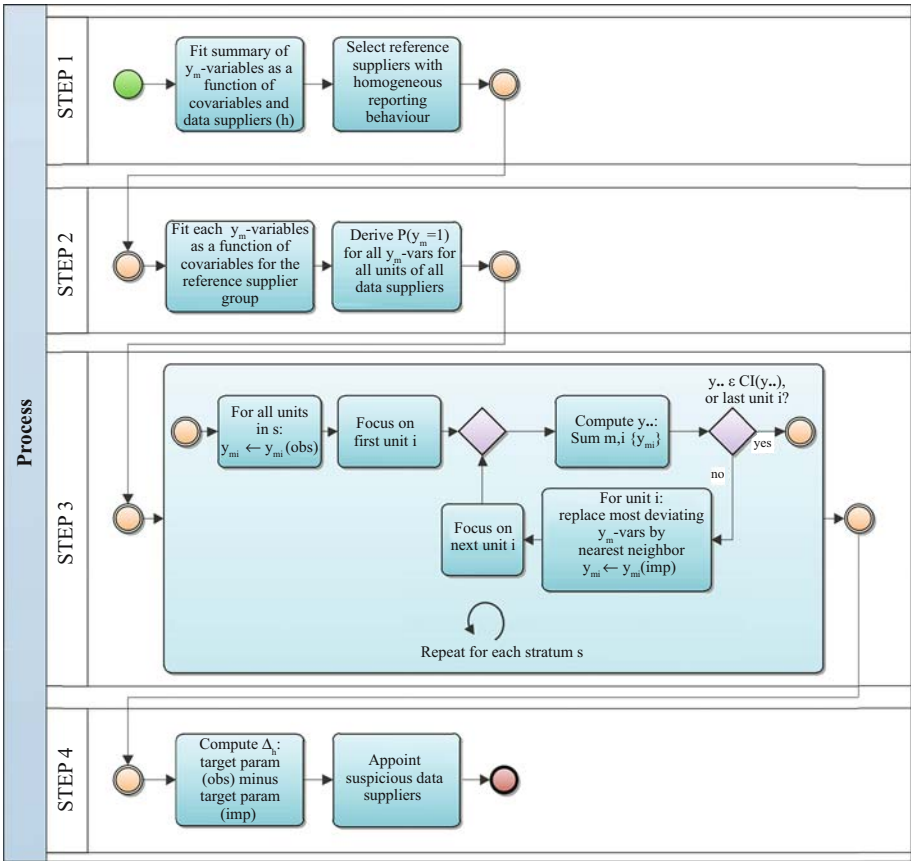

Fig. 5. Flow chart of the four steps of the methodology proposed by van Delden and coauthors.

record linkage can also be conducted for the purpose of adding value to a survey, by bringing in data from administrative records. In the fifth article of this special issue, Gessendorfer, Beste, Drechsler, and Sakshaug note that "Surveys that perform record linkage to administrative records are often required to obtain informed consent from respondents prior to linkage. A major concern is that nonconsent could introduce biases in analyses based on the linked data." This missingness due to nonconsent is illustrated in Figure 6.

To remedy this missing data problem created by nonconsent, they propose that nonconsenters be matched with statistically similar units in the target administrative

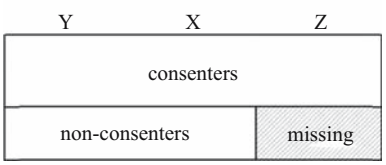

Fig. 6. The missing data situation when a survey data set has been combined with administrative data by means of record linkage (X being variables common to both data sets, Y being variables present in the survey and Z being variables present in the administrative data set). Source: Gessendorfer and coauthors.

database. In an empirical study, they assess the effectiveness of statistical matching for this purpose using data from two German panel surveys that have been linked to an administrative database of the German Federal Employment Agency. Their findings are mixed: the method can be effective in reducing nonconsent biases in marginal distributions, but biases in multivariate estimates can sometimes be worsened. This finding is valuable in itself; it is important that scientific journals do not limit their presentations to "success stories", dooming researchers to re-conduct studies of approaches already found to be inappropriate by others. (As noted by Karlberg and Radermacher (2014), "a lopsided evidence base, with the range of outcomes truncated to exclude failures, is a good example of the classical 'file drawer problem' [Rosenthal 1979]".)

4. Nontraditional Data Sources

While the administrative data sources discussed in the previous section demonstrably come with their own set of challenges, they are comparatively well-structured and well-defined. In contrast, "big data" are much harder to incorporate in statistical production, owing not only to their sheer volume, but also to issues such as their lack of structure and the fact that they are frequently held by third parties. Already five years ago (DGINS 2013), it was acknowledged that big data "represent new opportunities and challenges for Official Statistics" and the European Statistical System and its partners were encouraged to effectively examine the potential of big data sources in that regard. For the better part of the last decade, the official statistics community has been grappling with the challenge of integrating the new big data sources into official statistics production.

At the European level, a multi-pronged action plan and roadmap was established, and within the ESS Vision 2020 (Eurostat 2015), the Big Data (BIGD) project was tasked with implementing it. This has generated several outcomes, including methodological studies (such as the overview by Beręsewicz et al. (2018) of methods for treating selectivity in big data studies), analyses regarding aspects such as legal issues, ethics, quality and IT requirements, and a number of pilot studies (BDES 2018) involving new sources such as web scraping, smart meters, vessel tracking and mobile phone data.

On this latter data source, Vanhoof, Reis, Ploetz and Smoreda note in the sixth article of this special issue that "Mobile phone data are an interesting new data source for official statistics. However, multiple problems and uncertainties need to be solved before these data can inform, support or even become an integral part of statistical production processes." They then proceed to analyse the performance of five home detection algorithms (HDAs), based on mobile phone data characteristics such as amount of activities, amount of distinct days of activities, time of day constraints and space constraints. However, in their study, based on French Call Data Record (CDR) data, it turns out that no matter which HDA is applied, the dissimilarity between ground truth data and the estimated home location is large. While it is unsurprising that this is the case for holiday months (see Figure 7 for an illustration of the situation in August), this dissimilarity remains at a high level for all calendar months. The authors propose remedial actions at three levels: (i) studies at the individual level allowing the simultaneous observation of ground truth data and the related CDR data, (ii) reconciliation at the national level to compensate for the differences in local market shares between operators,

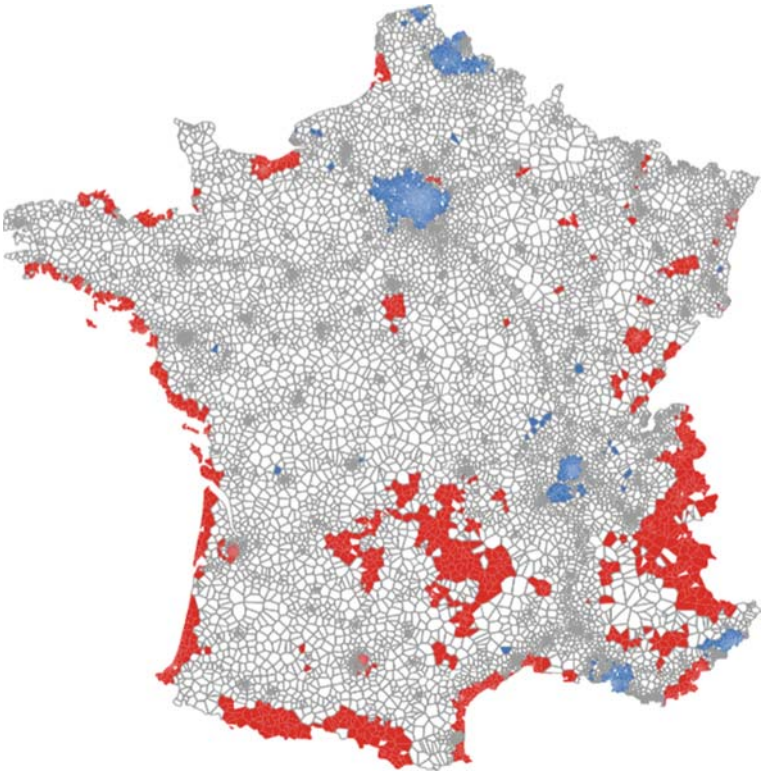

Fig. 7. French hotspots (red) and coldspots (blue) in August based on the ratio between the (mobile phone data based) activity-based home location and the population counts of a validation dataset. Note the effects of the reduced activity level in the capital region and the increased activity level in many typical holidaymaking areas. Source: Vanhoof and coauthors.

and (iii) testing at the international level to ascertain robustness of methods – across countries and over time.

To encourage experimentation with big data sources, one might try approaches that involve a competitive element. Two initiatives in this regard, both launched in 2016, were presented at NTTS 2017. The Big Data for Official Statistics Competition (BDCOMP) was the first official statistics nowcasting competition at EU level with a big data focus (Kovachev et al. 2017), requiring participants to submit a nowcast before the publication of official statistics. The BDCOMP could be said to follow a "marathon" approach, with participating teams delivering monthly submissions over roughly one year's time, whereas the EU Big Data Hackathon (Eurostat 2017a) was more of a "sprint". Over a period of just two days (and nights), participants had to develop solutions to a policy question ("*How would you support the design of policies for reducing mismatch between jobs and skills at regional level in the EU through the use of data?*"). A total of 22 teams from European National Statistical Institutes competed to develop a data analytics tool to support this, and then, on the third day, "pitch" their solutions, each with a time slot of just 6 minutes, to the evaluators. Two independent panels of evaluators (one statistical panel and one industry panel) assessed the contributions according to the criteria of relevance,

methodological soundness, communication, innovativity and replicability. The 1st, 2nd and 3rd prizes were awarded to the teams from Croatia, France and Estonia, respectively (Eurostat 2017b).

Opik, Kirt, and Liiv describe, in the seventh article of this special issue, the 3rd prize winning contribution to the EU Big Data Hackathon. They "present a visual method for representing the complex labour market internal structure from the perspective of similar occupations based on shared skills". Their method, based on graph theory (West 2001), is designed to enable adding extra layers of external information. Moreover, they offer a prototype for a tool allowing users to interact with the visualisation.

To demonstrate their methods and tools, Opik and coauthors conducted a case study in which they analyse the labour market together with the megatrend of automation and computerisation of jobs. Starting out by integrating data sets on job vacancies, they arrive at a graph depicting 2,950 occupations, with links between them based on their occupational similarity, and nodes annotated with megatrend (susceptibility to automation/computerisation) and supply data.

They proceeded to build a user interface that visualises the graph in a way that renders a zoomable and scrollable scalable vector graphics (SVG) document for browsing the graph online (supported by most modern web browsers). It has both a "move and zoom" mode, allowing for exploration, and a "query" mode (illustrated in Figure 8). The source code for the prototype (as well as the prototype itself) is made available on GitHub.

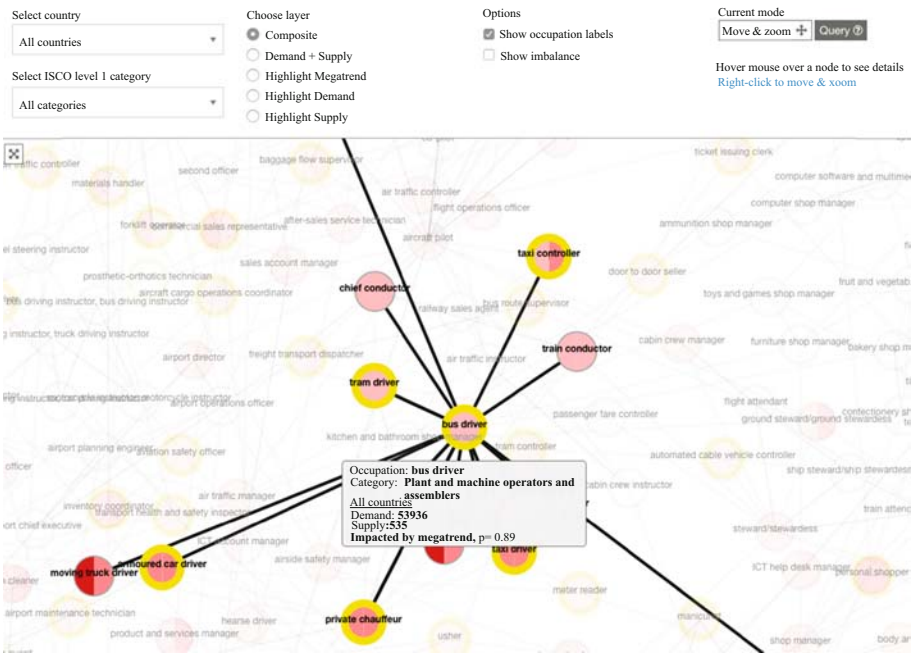

Fig. 8. Screenshot illustrating the query mode activated for the visualisation prototype. When an analyst moves the mouse cursor over a node, a small tooltip with demand and supply numbers is displayed. Hovering also highlights connected jobs and fades out the rest of the graph. Source: Opik and coauthors.

5. New Ways of Disseminating Official Statistics

The official statistics community has become increasingly aware of the importance not only of how official statistics are produced, but also of how they are presented. As an example, two (out of five) key areas of the ESS Vision 2020 (Eurostat 2015) – *focus on users* and *improve dissemination and communication* – concern how to understand user needs, and how to improve user communication. To tackle these issues, the ESS Vision 2020 implementation project *Digital communication, User analytics and Innovative products* (DIGICOM; see Kormann 2016) was launched. Given the simultaneous focus on the use of new big data sources and the communication and visualisation of their outcome, the EU Big Data Hackathon (2017a) was carried out jointly by the BIGD and DIGICOM projects.

While the DIGICOM project has a rich and varied portfolio of activities concerning, for example, user analysis, visualisation, open data dissemination, statistical literacy and gamification (Kormann et al. 2018), there are numerous other innovative activities underway at any given moment in the official statistics research community. A manifestation of this is the final article of this special issue, in which Hudec, Bednárová and Holzinger propose a method for disseminating statistics verbally, by means of natural language. Moreover, and in order to reflect the elasticity of many verbal quantifiers, the authors apply fuzzy logic to allow a "sliding scale" definition of linguistic terms; for attributes related to the values of a unit of the variable under study (such as "high pollution" and "low pollution") as well as quantifiers related to the relative frequencies of units possessing these attributes ("few enterprises", "about half of the enterprises", "most enterprises"). They demonstrate their concept using a test interface interpreting summaries from real municipal statistics data.

6. Outlook for Continued Research and Innovation in Official Statistics

A number of recent trends in statistical research and innovation have been manifested in the articles of this special issue. As pointed out in the article by Salgado and coauthors, the modernisation of statistical production "is to be accomplished under the high pressure of product release calendars within the traditional stove-pipe production model and a decreasing amount of budgeted resources". In Section 2, we have seen approaches that would allow official statistics to innovate in an agile way while under such resource constraints.

A recurrent theme in this special issue is that of learning through sharing. For instance, both Salgado and coauthors and Opik and coauthors embrace this principle by making their source code freely and publicly available on GitHub. Moreover, sharing is also a matter of letting colleagues know what does not work, so that they are already aware of the weaknesses associated with (and the need to improve) certain approaches. In this respect, Gessendorfer and coauthors, in demonstrating the limitations of statistical matching, as well as Vanhoof and coauthors, who point to the weaknesses of home detection algorithms, render valuable services to the official statistics research community.

"Improving the numerical and statistical literacy of citizens, journalists and policymakers will help to increase their awareness and ability to critically assess news, including fake news, and their participation in the democratic process" as noted by the

Power from Statistics panel on statistics in the digital era (Eurostat 2018). As Hudec and coauthors emphasise, the objective of the method that they propose is not to replace existing dissemination with verbal summaries, but rather to provide an alternative way of dissemination, which might be useful to certain categories of users. For instance, this might be a useful complement (to numbers) for persons that are innumerate, or have a low level of statistical literacy. Moreover, visualisation is not the ideal way to disseminate to everyone – visually impaired users, but also users for whom the interpretation of diagrams does not come naturally may be unable to grasp what is being communicated visually. "Reading graphs and charts is far from intuitive", as pointed out by Cairo (2018).

When new, unforeseen political and societal challenges emerge, policymakers and other stakeholders require timely evidence, produced with short lead times. In this context, the article of Opik and coauthors demonstrates the capacity for rapid innovation in the official statistics community. Based on the effectiveness of the "hackathon" approach, we were glad to learn that NTTS 2019, which will take place in Brussels from 12 to 14 March 2019, will also include a hackathon. This is just one of the many elements that contribute to the relevance of the NTTS series of conferences in advancing the research frontiers of official statistics.

Martin Karlberg
Guest Editor

Silvia Biffignandi
Piet J.H. Daas
Loredana Di Consiglio
Anders Holmberg
Risto Lehtonen
Ralf T. Münnich
Boro Nikic
Marianne Paasi
Natalie Shlomo
Roxane Silberman
Ineke Stoop
Guest Associate Editors

7. References

BDES. 2018. *Minutes*. Minutes from the 2018 conference on Big Data for European Statistics (BDES 2018). Available at: https://webgate.ec.europa.eu/fpfis/mwikis/ess-netbigdata/images/c/cf/BDES_2018_05_14-15_Sofia_Minutes.pdf (accessed August 2018).

Beręsewicz, M., R. Lehtonen, F. Reis, L. Di Consiglio, and M. Karlberg. 2018. *An Overview of Methods for Treating Selectivity in Big Data Sources*, Eurostat Statistical Working Paper. Doi: https://doi.org/10.2785/312232.

Cairo, A. 2018. "Uncertainty and graphicacy: How should statisticians, journalists, and designers reveal uncertainty in graphics for public consumption?" In *Power from*

Statistics: data, information and knowledge – Outlook Report 91-102. Publications Office of the European Union. Doi: https://doi.org/10.2785/721672.

DGINS (*Directeurs Généraux des Instituts Nationaux de Statistique*). 2013. *Scheveningen Memorandum – Big Data and Official Statistics*. Adopted by the European Statistical System Committee on 27 September 2013. Available at: https://ec.europa.eu/eurostat/documents/7330775/7339365/Scheveningen-memorandum-27-09-13/2e730cdc-862f-4f27-bb43-2486c30298b6 (accessed August 2018).

Esteban, E., S. Saldaña, and D. Salgado. 2017. *Software Implementation of Optimization-based Selective Editing Techniques at Statistics Spain (INE)*. UNECE Work Session on Statistical Data Editing. The Hague, 24-26 April 2017. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2017/mtg2/Paper_19_Stat-Spain.pdf (accessed August 2018).

Eurostat. 2015. *ESS Vision 2020 – Building the Future of European Statistics*. Luxembourg: Publications Office of the European Union. Doi: https://doi.org/10.2785/078607.

Eurostat. 2017a. *European Big Data Hackathon*. Available at: https://ec.europa.eu/eurostat/cros/EU-BD-Hackathon (accessed August 2018).

Eurostat. 2017b. *Winners of the EU Big Data Hackathon awarded*. Eurostat news release published 17 March 2017. Available at: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/CDN-20170317-1 (accessed August 2018).

Eurostat. 2018. *Power from Statistics: Data, Information and Knowledge – Guidance Report*. Publications Office of the European Union. Doi: https://doi.org/10.2785/302736.

Karlberg, M., S. Biffignandi, P.J.H. Daas, A. Holmberg, B. Hulliger, P. Jacques, R. Lehtonen, R.T. Münnich, N. Shlomo, R. Silberman, and I. Stoop. 2015. "Preface." *Journal of Official Statistics* 31(2) . Doi: https://doi.org/10.1515/jos-2015-0011.

Karlberg, M. and W. Radermacher. 2014. "Discussion: Methodology Architecture – an area under construction." *Statistical Journal of the IAOS* 30(4): 393–398. Doi: https://doi.org/10.3233/SJI-140855.

Kormann, C. 2016. *Innovation for Dissemination in the European Statistical System – the Approach of the DIGICOM Project*. Paper presented at the 2016 UNECE workshop on statistical data dissemination and communication. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.45/2016/Christine_Kormann_paper.pdf (accessed June 2018).

Kormann, C., S. Klasinc, M.J. Vinuesa Angulo, N. Tsiligkaki, A. Ilkova, P. Campos, M. Jankowska, P. Collesi, X. Caruso, and S. Luhmann. 2018. *A European Effort to Explore Games and the Gamification of Official Statistics*. Paper presented at the 2018 Conference of European Statistics Stakeholders. Available at: https://coms.events/cess2018/data/abstracts/en/abstract_0014.html (accessed August 2018).

Kovachev, B., M. Karlberg, B. Nikic, B. Oancea, and P. Righi. 2018. *The Big Data for Official Statistics Competition – Results and Lessons Learned*. Paper presented at NTTS 2017. Available at: https://ec.europa.eu/eurostat/cros/ntts2017programme/data/abstracts/abstract_295.html (accessed August 2018).

Rosenthal, R. 1979. "The file drawer problem and tolerance for null results." *Psychological Bulletin* 86: 638–641. Doi: http://dx.doi.org/10.1037/0033-2909.86.3.638.

Saltzer, J.H., and M.F. Kaashoek. 2009. *Principles of Computer System Design: An Introduction*. MIT Press.

Vâju, S., and M. Meszaros. 2018. *Administrative Data and Quality – Guidelines Towards Better Quality of Administrative Data*. Paper presented at the 2018 European Conference on Quality in Official Statistics (Q2018). Available at: https://www.q2018.pl/wp-content/uploads/Sessions/Session%2037/M%C3%A1ty%C3%A1s%20M%C3%A9sz%C3%A1ros/Session%2037_Matyas%20Meszaros.docx (accessed August 2018).

West, D.B. 2001. *Introduction to Graph Theory*. Prentice Hall.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 811–833, http://dx.doi.org/10.2478/JOS-2018-0041

Data Organisation and Process Design Based on Functional Modularity for a Standard Production Process

David Salgado[1], *M. Elisa Esteban*[1], *Maria Novás*[1], *Soledad Saldaña*[1], *and Luis Sanguiao*[1]

We propose to use the principles of functional modularity to cope with the essential complexity of statistical production processes. Moving up in the direction of international statistical production standards (GSBPM and GSIM), data organisation and process design under a combination of object-oriented and functional computing paradigms are proposed. The former comprises a standardised key-value pair abstract data model where keys are constructed by means of the structural statistical metadata of the production system. The latter makes extensive use of the principles of functional modularity (modularity, data abstraction, hierarchy, and layering) to design production steps. We provide a proof of concept focusing on an optimisation approach to selective editing applied to real survey data in standard production conditions at the Spanish National Statistics Institute. Several R packages have been prototyped implementing these ideas. We also share diverse aspects arising from the practicalities of the implementation.

Key words: Production architecture; key-value pair data model; standardisation functional modularity; process design.

1. Introduction

The modernisation and industrialisation of official statistical production has been at the centre of international and national activity in Official Statistics basically since the turn of the century, with the creation of the High-Level Group for the Modernisation of Official Statistics by the Bureau of the Conference of European Statisticians as a noticeable landmark (HLG-MOS 2017).

Indeed, this group was born with a clear strategic vision (HLG-MOS 2011) to streamline the statistical production by means of "different and better processes and methods tuned to delivering our products at minimal cost with greater flexibility and in cooperation between institutions" so that these "new and better products and services [are produced] more tuned to the way the world is operating today". Many outputs have been produced by the different groups operating under the umbrella of the HLG-MOS; these range from the establishment of diverse production standards (such as the Generic Statistical Business Process Model, GSBPM, the Generic Statistical Information Model, GSIM, the Common Statistical Production Architecture, CSPA, or the Generic Activity Model for Statistical

[1] Spanish National Statistics Institute, Paseo de la Castellana 183, 28046 Madrid, Spain. Emails: david.salgado.fernandez@ine.es, elisa.esteban.segurado@ine.es, maria.novas.filgueira@ine.es, soledad.saldana.diaz@ine.es, and luis.sanguiao.sande@ine.es

Organisations, GAMSO) over the promotion and development of streamlined statistical methods (e.g., UNECE 2017a) to capabilities and communication aspects (UNECE 2017b).

More recently, within the realm of the European Statistical System (ESS hereafter), the future of European Official Statistics is strategically envisaged by the ESS Vision 2020 (Eurostat 2014a) and its implementation portfolio in key projects, such as those focused on the European System of Business Registers (ESBRs), the Common EU Data Validation Policy (VALIDATION), the Shared Services for European Statistics (SERV), and the Digital Dissemination and Communication (DIGICOM), to name a few (Eurostat 2014b).

All these initiatives pose a challenge for statistical offices in their attempt to modernise their production, especially regarding the adoption of these new standards and practices: this is to be accomplished under the high pressure of product release calendars within the traditional stove-pipe production model and a decreasing amount of budgeted resources.

In this article, we want to present the ongoing efforts at the Spanish National Statistics Institute to bring a concrete plan for the modernisation of (a part of) the statistical production process into reality. Our rationale is that an official statistical production system constitutes a clear example of a human-generated complex system. We claim that to cope with this complexity, like with the design of computer systems, the principles of functional modularity are also of great value. These principles must fully integrate statistical production metadata, statistical methodology, and computer software design. These principles are often applied in the construction of software for the production of official statistics, but this is not enough. We claim that these principles must be applied *to fully integrate these three aspects of statistical production*, or else we would fail to cope with the complexity of the process. To illustrate our proposal, we show how we have developed a set of R packages to make a proof of concept that is already being applied in normal production conditions of several Short-Term Business Statistics (STS) at the Spanish National Statistics Institute.

Our proposal is based on two complementary elements. Firstly, for our data architecture, we make use of a key-value pair structure, in which keys are composed by making extensive use of the system of structural metadata. Secondly, adhering closely to the GSBPM and GSIM principles, for our statistical process architecture, we make use of the functional and object-oriented paradigms to incorporate modularity into the statistical methods. As we shall illustrate with the R packages, this paves the way for a natural posterior implementation in software tools. Our central message is thus *to bring modularity by design into the statistical process and the mathematical methodology itself and not just into the construction of computer tools*.

The article is organised as follows. In Section 2 we set up the generic approach, taking us from complexity as an essential trait of statistical production systems to the principles of functional modularity to cope with it. In Section 3 we argue that the international statistical production standards themselves implicitly suggest the use of a combination of the object-oriented and functional paradigms as a basis for building an information architecture. In Section 4, we detail the abstract data model that we propose to use as the central element of our proposed data organisation. Complementarily, in Section 5, we explain our proposed process design, and illustrate, with an example in statistical data editing, the application of modularity principles on a very concrete statistical methodological approach to selective editing. In Section 6, we share diverse aspects

regarding the implementation of this proposal, including the software tools development. We close with conclusions and future prospects in Section 7.

2. Generic Approach: From Complexity to Functional Modularity

The need for modernisation and industrialisation of official statistical production can be immediately argued from the very concept of *complex system*. The key features of a complex system are (Saltzer and Kaashoek 2009) (i) a large number of components, (ii) a large number of interconnections between these components, (iii) many irregularities in these interconnections, since the lack of regularity is the rule rather than the exception, (iv) a long description of the system and its related management (Kolmogorov complexity), and (v) a team of designers, implementers, and/or maintainers to handle the system. It is evident that an official statistical production system is a clear example of a human-generated complex system.

This conclusion can be illustrated and motivated with a simple description of the production of diverse statistical operations at a statistical office. Let us just consider the execution phases of the process. Data collection needs to be carried out in different data collection modes (CAPI, CATI, CAWI, EDI and others) on a number of statistical units (business units, households, or people), usually in the range of tens of thousands for each survey in a mid-sized country like Spain. This is multiplied by the number of variables (data or metadata) associated with each unit. These data must be entered into the system, edited, treated, validated, and curated to produce the corresponding microdata sets. They are further processed to produce the aggregated outputs using the appropriate statistical methods. They are then finally treated for disclosure control and, if necessary, for seasonality and calendar effects adjustment before the due dissemination. Each production step and data and metadata element in the process is interconnected to some other element. For example, a change in a parameter in a validation rule during collection will need to be followed by a post-capture data editing revision and adjusted aggregation procedure (e.g., in variance estimation). Indeed, the interconnections between all elements cannot be described according to a given regularity, thus making explicit the *water-bed effect*: a slight modification of a process step may lead to major consequences in another process step. Given the current setting of the statistical process at production offices, the description of how to produce the statistics for any given survey is not only necessarily long, showing the imbricate set of process steps, but also, hardly standardised. Members of the production staff of two different surveys who carry out the same tasks in the process can seldom be interchanged, despite common standard mathematical procedures underlying the estimation. Moreover, the number of actors in the process to be coordinated, not only for a given statistical operation, but also for the set of surveys conducted at an office (not to mention a whole national or European statistical system) is very high, which introduces evident management challenges.

In our view, the concept of official statistical production as the combination of statistics and complexity lies at the core of the need for the industrialisation of the statistical production process: not only do you need to use sound statistical methodology, you must also cope with this complexity for an efficient production process. Traditionally, in our view, official statistics have been produced in an artisan way, in which each survey was

independently designed and executed. Moreover, in extreme cases, not only have there been diverse (occasionally even incompatible) data and process architectures in different surveys in the same office, but different agents within the same survey have also made use of unconnected architectures, which has rendered management of the whole process virtually impossible. Up to the present, the stove-pipe production model has been extensively followed.

On a more quantitative footing, the inefficiency of this stove-pipe approach can also be justified by the complex nature of the production system itself. As a complex system, it is subjected to the square law of computation (Weinberg 2011) (see also Saltzer and Kaashoek 2009), which in our case can be expressed in terms of resources versus the number of requirements on the system.

A simplified description of how to detect and correct errors in a process step can illustrate and motivate this law. A process step is, basically, a collection of sequential and concurrent production tasks for accomplishing a given objective within the process. We can easily assume that the potential number of errors is proportional to the size of the production step (i.e., to the number of tasks) and that they can occur randomly throughout the step. In principle, in a nonmodular approach, an error is detected after executing the process step, which is then fixed. The process step is then executed again to detect new errors. If the time to find an error is assumed to be proportional to the execution time, the total amount of time to clean the process step will be proportional to the number of errors multiplied by the necessary cleaning time per error. However, the latter is proportional to the number of errors itself. Thus, the total amount of time will be quadratic to the number of errors. This argument shows how a naïve sequential approach to production becomes unmanageable due to the complexity of the system.

Under this square law, it is clear that increasing the number of requirements on the system (due to the incessant demands on Official Statistics, for example new legal regulations, more disaggregated information and so on) will produce a quadratic increase in the demand of resources, which is unattainable. Complexity must be coped with to face these challenges. The need for modernisation derives from the complexity of the global statistical production process.

The bottom line of our proposal is that we believe that the common principles of computer system design jointly known as *functional modularity* (Saltzer and Kaashoek 2009) are of great utility in designing and implementing an efficient official statistical production process. It is worth noting that functional modularity comprises four elements, namely modularity, data abstraction, hierarchy, and layering. These principles should be applied not only to the development of computer tools: *the process itself must be designed along these lines by conjugating statistical metadata, statistical methodology, and software design*.

Modularity is already at the very heart of production standards (such as the GSBPM – see next section), where the production chain is broken down into different subprocesses. However, modularity per se does not help us cope with complexity; we need data abstraction as this allows modules to be designed and implemented independently of each other, except for their interconnecting interface. Statistical processes must be designed independently of each other so that only initial inputs and final outputs uniquely enter into play in the chained execution of a given set of processes. The details of the execution of each subprocess must be transparent throughout the entire process.

Layering and hierarchy are principles applied to design and implement modules to minimise the number of interconnections among their components seeking optimal efficiency. In our proposal, these principles will be translated into organising both data and process architectures into four layers. A bottom layer for the statistical methodology (purely mathematical in many, but not all, cases); a second layer for the finest-grained production tasks upon which more complex activities can be composed (third layer). Finally, a top layer to orchestrate the whole process with these elements will complete the process design. We insist on the idea that this structure *must be applied to the statistical processes themselves, conjugating metadata, mathematics and software design*, not just to the construction of computer tools.

3. From Metadata to Architecture

The starting point for concretising our proposal into data organisation and process design is the interrelationship between the GSBPM and GSIM standards. The GSBPM is an international production standard modelling the statistical production chain in eight phases, each one divided in different production subprocesses. This standard focuses on production activities. Complementarily, the GSIM is another international production standard providing a model for the information objects in the production process. The inspiring interrelationship between the two standards is represented in Figure 1, already originally appearing both in the GSBPM (UNECE, 2013a) and in the GSIM (UNECE, 2013b).

There is also an implicit reference to this interrelationship that appears in the name of the GSBPM level-2 subprocesses (*Design collection*, *Test production system*, *Calculate aggregates* and so on) with the clear structure *action* + *information object*. If several transformations matching Figure 1 are concatenated, where the output of a step is the input of the next one, and if each transformation is associated to each input object, we have the conception of a statistical production process as a sequence of objects defined through their attributes (GSIM-like information objects) and transformed according to their methods (GSBPM-like production tasks).

Our proposal suggests a step forward in this direction by extensively using the principles of functional modularity to substantiate this general view of the combination of both GSBPM and GSIM. Note that these standards do not make any explicit mention of these principles, although their spirit is there. Similarly, in the international DDI standard (DDI 2018) a modular scheme for the successive transformations on both data and metadata sets is provided. Here, we also include these data and metadata under the same modular view.

To implement this dual data-process view under the principles of functional modularity, we firstly need to provide a data organisation scheme to deal with information objects in a standard way. Indeed, the proposed scheme must be valid for all kinds of statistics (social surveys, business statistics, statistics based on administrative registers, and so on). In the

Fig. 1. Interrelationship between GSBPM and GSIM standards (taken from UNECE013a).

next section we present an abstract data model based on key-value pairs in this sense. Indeed, we will define an object class for representing data in any kind of statistical data processing subprocess.

Complementarily, a process design scheme also needs to be provided. We understand that every "unit of statistical production information" is defined through a set of attributes (GSIM-like part) and a collection of statistical transformations (GSBPM-like part). In other words, they are *objects* (Booch et al. 2007). Furthermore, these objects can be thought of as constituting a sequence of transient transformations also combining data and metadata. This enables traceability and auditability of the whole process.

Indeed, this is extremely evocative of well-known computing models (van Roy and Haridi, 2004): the object-oriented and functional paradigms. The application of these paradigms makes each transformation depend only on its object input – it becomes stateless, that is, depending on no previous production step (state, in rigour). A cautious reader may immediately argue whether those steps involving (pseudo)random number generation arise as an exception to this stateless sequence of transient transformations. In full rigour, one can consider the random number generation seed as an internal state of the transformation. However, in the spirit of those statistical methods involving random simulation, we can accept that two processes providing *statistically* similar results can be considered identical under the data organisation and process design we defend here even despite numerical dissimilarities. This is a natural way of implementing referential transparency, that is, a property by which the procedure can be replaced with its corresponding value without changing the behaviour and the result of the whole process. As a consequence, executing a referentially transparent subprocess will always provide the same value for the same input arguments, irrespective of the rest of the process. This is the functional paradigm. As for the object-oriented paradigm, we concentrate on its advantages to model complex objects, and on its characteristics regarding transformations. Thus, transformations are conceived under the functional paradigm and objects are understood and modelled using the object-oriented paradigm.

However, we need to be more concrete about how to combine these paradigms in statistical processes. Let us focus on the recommendations of the METIS group elaborated by their informal task force on metadata flows (ITFMF 2013), in particular, to document each production task by different elements, namely (i) input data, (ii) input parameter, (iii) throughput, (iv) output, and (v) process metric. These recommendations are followed closely in the Generic Statistical Data Editing Models (UNECE 2015). In the present work, we will leave out the fifth element about the metric. We propose the following structure for every data-processing production task. We conceive every data-processing production task as a transforming action on a data set under a set of parameters producing a new data set or a new parameter set. We represent this as

$$\text{OutputData},\ \text{OutputParameters} := \text{Action}(\text{InputData},\ \text{InputParameters})$$

It must be noted that the distinction between data and parameter is somewhat arbitrary, since it depends on the semantic context of the concrete computation. For example, in Predict(InputData, PredictParameters) we compute predicted values for those data in the object InputData according to those parameters specified in the object PredictParameters, for instance, an ARIMA time series model ARIMA(p, d, q). Previously, we would

need to compute the degrees *p, d* and *q*. These can be computed similarly by PredictParameters: = ComputeDegrees(PredictParameters, DegreeParameters), where an initialised parameter object PredictParameters is updated with the computed degrees and where DegreeParameters specifies the parameters needed to compute *p, d* and *q*. Notice how in this second computation PredictParameters acts as an input data object.

This distinction concerning data and parameters can also be discussed in other common settings in standard production conditions. For instance, when joining two data sets, we can consider both data sets as elements of a more complex InputData object and the join resulting from the parameters specified in the corresponding InputParameters object (inner, outer and so on.). In the same vein, adding new records to an existing data set can also be modelled through a complex InputData object with an appropriate InputParameters object. Depending on the traceability and auditability provided to the whole system, the transient transformations can be further conveniently stored specifying timestamps, usernames and so on.

All in all, functional modularity principles can be used to implement this combination of paradigms by setting up a hierarchy of layers from (i) the statistical methodology, over its implementation in (ii) low-level procedures (possibly assembled in libraries) and (iii) high-level procedures thereof, to (iv) a process-orchestrating layer working as a user interface.

Notice how this organisation in layers also coincides with different traditional profiles at statistical offices. Statistical methodology is under mathematicians' and methodologists' responsibility, possibly also with the collaboration of domain experts. This layer focuses on the more abstract and mathematical part of the production system. The second layer implements the methodology as low-level software procedures. It falls under developers' and programmers' responsibility, possibly with the collaboration of programming-skilled methodologists. This layer still maintains a certain degree of abstraction. Concrete applications and production activities are shaped in the third layer under the responsibility of statisticians and survey managers, possibly with the aid of developers. In this layer, the collection of standard low-level procedures is adapted to the concrete needs of each statistical program. Finally, a process orchestrator working as user interface for ease of the human-computer interaction can additionally be put into place. This ease of use allows the management to optimise the production resources by potentially assigning tasks to non-specialists who follow previously specified protocols.

In the following sections we will use concrete surveys conducted at the Spanish National Statistics Institute to illustrate how this information architecture has been partially deployed for the statistical data editing phase. Our first step has been to propose a common data structure for all survey and administrative data sets (thus either InputData or OutputData) based on a standardised abstract data model for any kind of statistics. This is detailed in Section 4.

Next, we have implemented the optimisation-based selective editing techniques formerly developed at the Spanish National Statistics Institute (Arbués et al. 2013) following these principles. This boils down to designing and programming Actions together with different sets of InputParameters (also OutputParameters). We undertake this in Section 5.

4. Data Organisation

We will use the Spanish Retail Trade Survey and Service Sector Indicators Survey, conducted monthly at the Spanish National Statistics Institute, to illustrate the application

of this approach. These are short-term business statistics. Data are collected through paper questionnaires, telephone, fax, email, and CAWI modes. Statistical units are selected according to a stratified simple random sampling design. Target aggregates are mainly Laspeyres indices of both turnover and number of employees, possibly broken down into economic sector code and type of employment contracts, respectively.

In the preceding framework, our first task is to define an abstract data model for all statistical operations. The immediate goals of this model have been the versatility among all kinds of survey or administrative data and fast and easy deployment in the implementation.

The model essentially consists of a key-value pair data model, in which the key is composed by using the structural statistical metadata of the production system. We must distinguish between the data model for storing data in a corporative internal repository (the key is not parsed) and the data model for processing (the key is parsed). For manageability and rapid deployment reasons, in the current implementation the information is stored in plain text files, as explained below. These files are not modified once written. Updated information, if any, is included as a new file (with updated key in the name of the new file; see below). Concurrency issues and many other data architecture details are not considered relevant at this point.

The central element in the data model is the composition of the key for each single datum in the global production system at the office scale (or the whole statistical system scale). The key is composed of the following components:

(i) An alphanumerical code to identify the survey/statistical program.
This alphanumerical code is taken directly from the Spanish National Statistical Plan, where each survey/statistical program is univocally identified. This code references the concrete statistics where this value is generated, processed, and used.

(ii) An alphanumerical code to identify the time period of reference (coincidental with the time period of the corresponding statistics).
An ad-hoc simplified syntax has been put into place to denote the different reference time periods for all statistical operations according to the following table:

Time period	Code
Month	MM, MR
Trimester	TT, TR
Semester	SS, SR
Year	AA, AR

The second character denotes whether it is an ordinary data set or a duplicated data set containing statistical units from the rotated sample. This is especially used in short-term business statistics that use chain-linked Laspeyres indices with rotating panels.

(iii) An identifier to indicate whether they are raw or (partially) edited microdata, paradata, identification data and so on.

The different codes are:

Data file type	Code
Finally validated values	FF
Partially edited values	FD
Raw values	FG
Paradata	FP
Identification variable values	FI
Edit rules (Longitudinal phase)	FL
Edit rules (Cross-sectional phase)	FT

(iv) A version number either with the prefix *P* for provisional or *D* for definitive values.

(v) An identifier for the statistical variable.
This identifier is taken from the system of structural metadata so that each concept measured with a statistical operation in the whole statistical production system is identified with a standard name. For example, the concept of "turnover" is measured in different surveys (industry, retail trade, service sector and so on) and the same identifier *Turnover* is used in every survey. Subtleties in this statistical variable arising from its concrete usage in a survey is further specified using qualifiers (see immediately below).

(vi) A set of qualifiers specifying different attributes (statistical unit ID, geographical code, economic activity code and so on).
Qualifiers are variables that further specify the semantic content of each value. Although from a strict computer point of view, all qualifiers play the same role, this is not the case from a statistical standpoint. There are basically two types of qualifiers, namely, those that allow us to identify the statistical units, and the rest of them. The latter can be further divided into two categories. Firstly, as in the example below, there are qualifiers that amount to codes of standard classifications, such as the NACE, PRODCOM, COICOP and so on. At the Spanish National Statistics Institute, to the extent that it is feasible, international standard classifications are in use, in agreement with the ESS. In parallel, not all qualifiers of this type can be found in standard classifications. In these cases, in agreement with domain experts, the metadata unit puts into place a collection of internal standard classifications for these qualifiers. For example, the number of employees in a business unit is an extensively requested variable, usually broken down according to diverse criteria: by type of contract, by professional situation, and by type of remuneration. These have given rise to classifications with their own codes, which are used as qualifiers in the corresponding key. Secondly, there are qualifiers that are not necessarily understood as part of a classification. For example, the economic activity code of a business unit may change because of a change in its business activity, so that this variable in the population frame should be modified after receiving the updated information during field work. A qualifier (say, IsMod) denoting whether we are referring to the former value (IsMod $= 0$) or the modified value (IsMod $= 1$) must be introduced. This self-evident qualifier

value is not part of a classification. More specific qualifiers can always be used according to the specific process being executed. For example, in statistical data editing qualifiers in terms of population, measurement time, measured unit, and measured element can be properly defined, coded, and used as qualifiers (van der Loo 2015).

The following simplified example clarifies the meaning of these components. Let us consider the validated value of the turnover for a business unit (statistical unit ID 289409300MM) in the Retail Trade Survey (code E30103) in the reference time period of January 2016 in the region of Castilla-La Mancha (geographical code 08), in the economic sector of trade of food and beverages (NACE Rev.2 code 47.11). This value pertains to the first definitive data set for this time period. This is visually depicted in Figure 2. Note that some qualifiers are missing in this simplified example, as structural metadata defining the variable type (integer value expressed in euros).

As stated above, in the current implementation, data are stored in files. Each one is identified by statistical operation code, type of data (finally validated data, raw data, paradata, and so on), reference time period, and definitive or provisional character of the data in the production process. In other words, the common part of the key for a data set is encoded in the name of the corresponding file, where the rest of the key and the values are stored. In each file, each line will keep the standardised identifier and the rest of qualifiers together for each value (e.g., Turnover@@289409300MM47.1108@@9732 in our example). Other implementations are also possible.

A data dictionary is also configured and stored, containing the specifications of each statistical variable: name, description, data type – numeric or alphanumeric, maximal

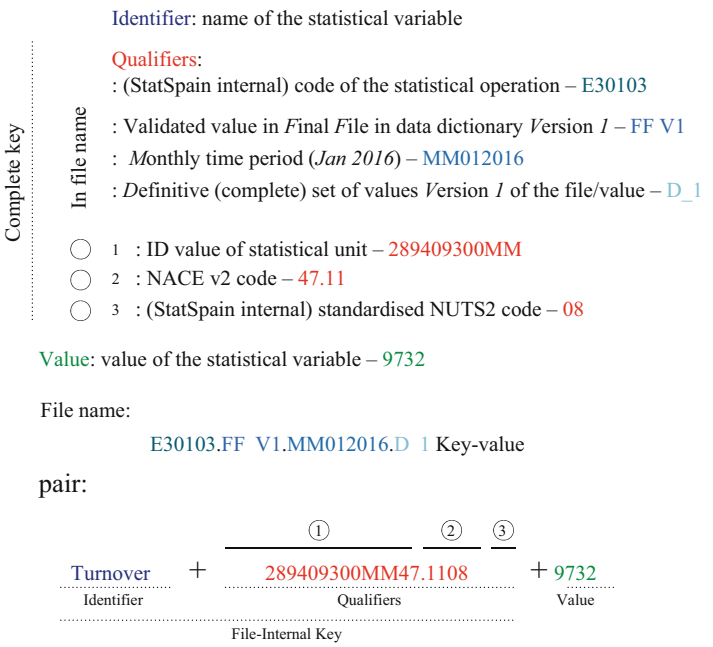

Fig. 2. Example of a key-value pair with a key composed of structural statistical metadata.

length – in terms of number of characters, qualifiers, corresponding domain-used variable names, range of values, and some other technical information for data collection applications. This dictionary allows the user to parse the key to instantiate objects according to a business logic class for all data processing tasks, which is indeed a data frame where the parsed key components are assigned in respective columns together with the corresponding value column. In this way, data are tidy, in the sense of Wickham (2014), for further processing with standardised transformations. Tidy data mean Codd's 3rd normal form so that (i) each variable forms a column, (ii) each observation forms a row, and (iii) each type of observational unit forms a table (see Wickham 2014). This business logic class consists essentially of the data frame and the data dictionary. Data transformations are applied on this class of objects, returning updated objects of the same class.

Immediate benefits are obtained after adopting such a data organisation. Firstly, since every data of every survey/statistical program can be managed in this way, a unique data architecture can be adopted throughout the entire production system of the office. This is a first crucial step towards the suppression of the stove-pipe production model, paving the way for a more efficient architecture. Having a common data architecture allows us to build standardised applications valid for all surveys, thus leading to the rationalisation of resources.

Secondly, these data specifications can be adapted to many actual circumstances in daily production. Let us consider, for instance, the case in which the economic activity code in the example changes along the process because the business unit has changed its activity. The example depicted here is oversimplified for ease of illustration. In practice, the metadata system has dozens of standard classifications for qualifiers (always international when possible) to parameterise each single datum along the process. In particular, we have four classifications that aim to pinpoint (i) the process stage in which the value is generated (design, collection, processing, dissemination, and so on, or a subprocess thereof), (ii) the element of the process which the value is related to (frame population, sample, questionnaire, and so on or a sub-element thereof), (iii) the role of the related actor in the process (statistical unit, interviewer, editing clerk, and so on), and (iv) the type of value (dichotomic variable, excluding variable, percentage, and so on). The evolution of the value along the process can be followed using these qualifiers. The metadata unit has put in place and is maintaining over 70 classifications and growing, as more statistical programs incorporate this architecture. Many classifications are very specific for a given statistical domain, but many others refer to features common to a large number of surveys.

Thirdly, the use of metadata in composing the keys to identify data values paves the way for achieving a standardised production system. In this way, every single datum in the whole production process is parameterised using, so to say, a common system of coordinates. In contrast to the dangerously common opinion of only conceiving metadata as a cumbersome documenting tool independent of production tasks and effective only after production has been executed (socalled *passive metadata* according to Lundell 2013), this data organisation makes use of the metadata system from the very beginning, in which data are generated and provide an interface between data and the user (*active metadata* according to the same author). Notice how this active role of metadata is key in

the sequence of transient transformations along the production process. Every independent transformation on a given data set must be implemented depending only on the input data and input parameters, that is, on the data and metadata contents that transform according to the parameters. If metadata are erroneous, the interface between data and the user is lost, and the process (as a sequence of transformations) cannot be executed.

5. Process Design

The design of the process architecture according to the principles set out in Section 3 is much more complex than the design of the data architecture. To begin with, a standard class of parameters (InputParameter) for all possible statistical methods (Action) is virtually impossible, since there exists a vast number of different statistical techniques. Thus, we will illustrate the application of the functional modularity principles with the concrete example of the optimisation approach to selective editing developed at the Spanish National Statistics Institute (see Arbués et al. 2013).

The division in layers begins by considering the statistical methodology at the bottom of the hierarchy. We will not go deep into the mathematical details and shall focus on the implementation of a very concrete formula to assign local (item) scores to each statistical unit.

The core of selective editing techniques is based on the assignment of a score to each variable to be edited for each statistical unit, thus providing a measure of the degree of suspicion of it containing an influential measurement error. The heuristic approach (de Waal et al. 2011) recommends choosing local (item) score functions such as $s_k = \omega_k |y_k - \hat{y}_k|$, where ω_k stands for the sampling weight of unit k and y_k, $\hat{y}_k$ denote the reported and predicted (expected) values of the variable y under editing, respectively. The main methodological content of the optimisation approach firstly consists of modelling the measurement errors $\epsilon_k = y_k - y_k^{(0)}$ ($y^{(0)}$ denoting the true value) for each unit and computing their first- and second-order moments M_{kl} for each pair of statistical units k and l (business units in our example) and each variable y (turnover and number of employees in our example). These are given by analytical expressions (Arbués et al. 2013):

$$\begin{aligned} M_{kk} &= \sqrt{\frac{2}{\pi}}\,\omega_k\,\hat{\nu}_k\,{}_1F_1\left(-\frac{1}{2};\frac{1}{2};-\frac{(y_k-\hat{y}_k)^2}{2\hat{\nu}_k^2}\right)\zeta_k\left(\frac{y_k-\hat{y}_k}{\hat{\nu}_k}\right), \\ M_{kl} &= 0, \quad k \neq l, \end{aligned} \tag{1}$$

for the loss function $L(a,b) = |a-b|$ and

$$\begin{aligned} m_k &= \omega_k\,\hat{\nu}_k\,\frac{\hat{\sigma}_k^2}{\hat{\sigma}_k^2+\hat{\nu}_k^2}\left(\frac{y_k-\hat{y}_k}{\hat{\nu}_k}\right)\zeta_k\left(\frac{y_k-\hat{y}_k}{\hat{\nu}_k}\right), \\ M_{kk} &= \omega_k^2\,\hat{\nu}_k^2\left(\frac{\hat{\sigma}_k^2}{\hat{\sigma}_k^2+\hat{\nu}_k^2}\right)^2\left[\frac{\hat{\sigma}_k^2+\hat{\nu}_k^2}{\hat{\sigma}_k^2}+\left(\frac{y_k-\hat{y}_k}{\hat{\nu}_k}\right)^2\right]\zeta_k\left(\frac{y_k-\hat{y}_k}{\hat{\nu}_k}\right) \\ M_{kl} &= m_k\,m_l, \quad k \neq l, \end{aligned} \tag{2}$$

for the loss function $L(a,b) = (a - b)^2$, where in both cases

$$\zeta_k(x) = \frac{1}{1 + \frac{1-\hat{p}_k}{\hat{p}_k}\left(\frac{\hat{\nu}_k^2}{\hat{\sigma}_k^2 + \hat{\nu}_k^2}\right)^{-1/2} \exp\left(-\frac{1}{2}\frac{\hat{\sigma}_k^2}{\hat{\sigma}_k^2 + \hat{\nu}_k^2}x^2\right)}$$

Exact details about the derivation of expressions (1) and (2) are given by Arbués et al. (2013). In the first case, when $|\frac{y_k - \hat{y}_k}{\hat{\nu}_k}| \to \infty$, $M_{kk} \to \omega_k |y_k - \hat{y}_k|$, which is the usual expression in the heuristic approach (de Waal et al. 2011) (in this case M_{kk} can be viewed as item scores). Thus, Formulas (1) and (2) can be understood as a rigorous generalisation of the traditional approach to selective editing by using statistical models for measurement errors. The scores also depend on other parameters, such as the probability of reporting an erroneous value and the variability of these errors reported in the past. As a matter of fact, statistical models for the measurement error are behind the diverse parameters in these expressions:

- ω_k denotes the sampling (design) weight of unit k;
- y_k denotes the raw (reported) value of variable y for unit k as collected in the questionnaire;
- $\hat{y}_k$ and $\hat{\nu}_k$ denote the predicted value and its prediction standard deviation for variable y and unit k;
- $F_1(x;y;z)$ stands for the confluent hypergeometric function of the first kind (Pearson et al. 2017), which arises from the choice of the loss function in the underlying optimisation problem;
- $\hat{p}_k$ denotes the estimated probability of measurement error for variable y and unit k, that is, $p_k = \mathbb{P}\left(y_k \neq y_k^{(0)}\right)$, where $y_k^{(0)}$ stands for the true value of variable y;
- $\hat{\sigma}_k^{(0)}$ denotes the estimated standard deviation for the observed measurement error $Q_k = y_k - y_k$.

These quantities can be computed for the whole population or by population cells (e.g., determined by economic sector or geographical region, or both).

Now we consider the second and third layers, in which the statistical methodology is implemented in finer – and coarser – grained production tasks. From the methodology, it is clear that the error moments can be written as functions of diverse parameters $M_{kl} = M_{kl}(y_k, \hat{y}_k, \hat{\nu}_k, \hat{\sigma}_k, \hat{p}_k, \omega_k)$. Now the question arises regarding how to organise this computation in a modular way.

At this point, functional modularity and statistical methodology must be precisely combined. From a strictly computational point of view, there is no distinction between the parameters y_k, $\hat{y}_k$, $\hat{\nu}_k$, $\hat{\sigma}_k$, $\hat{p}_k$, and ω_k. However, from a statistical point of view, this distinction is fundamental for allowing the system to efficiently grow and evolve in the future. Raw values y_k are taken directly from the data collection stage. Independent modules will handle the computation of $\hat{y}_k$ and $\hat{\nu}_k$ (prediction module), $\hat{\sigma}_k$ (observation error estimation module), $\hat{p}_k$ (error probability estimation module), and ω_k (sampling design module). The computation of these parameters will be completely independent of each another and each one will depend exclusively on its input arguments. They will

interact with each other only through their final computed values, so that the computation is transparent.

This organisation in modules is justified by the underlying statistical knowledge. First, there are many prediction methods that potentially could be applied to obtain both $\hat{y}_k$ and $\hat{v}_k$. If new methods need to be added to the system, this can be done without affecting the rest of the computation. This same observation is valid for the remaining modules. Note that this is a simple example in which we are computing a single value with an analytical formula with just six arguments. The consequences of a poor (from a methodological point of view) modular organisation, may produce the opposite effect across the entire production system. This is why functional modularity and statistical methodology must be precisely combined in the design of the production system.

Each module, in turn, makes use of these same principles, so that different methodological aspects of the computation are considered independently. For example, due to missing values or some other reason, predicted values cannot be computed for all statistical units and must be imputed. An independent module for imputation is thus constructed to handle this task independently of any other, and embedding it in the former computation. The architecture is, again, the same:

$$\text{ImputedObject} := \text{Impute}(\text{InputObject}, \text{ImputationParameters}).$$

The whole computation is then constructed as follows. Firstly, the Action element specifying the concrete production task will be denoted by ComputeErrorMoment and will implement either Formula (1) or (2), depending on its arguments.

As InputData we set all elements in Expressions (1) and (2), namely (i) the values of the target variables y (turnover and number of employees in our example), (ii) some other auxiliary variables (e.g., those determining different population domains; economic classification NACE codes, and Spanish geographical classification codes in our example), and (iii) the model parameters $\theta_k = (\hat{y}_k, \hat{v}_k, \hat{\sigma}_k, \hat{p}_k, \omega_k)$ for each variable y and each unit k. These are the parameters for the continuous variable observation-prediction model (Arbués et al. 2013). We will call this InputData data set contObsPredModParam and it is given the key-value pair structure described in the preceding section. These parameters (hence the object contObsPredModParam) must be computed with their respective modules:

- The predicted values $\hat{y}_k$ and their standard deviations $\hat{v}_k$ are computed by initialising the object contObsPredModParam and defining an abstract class PredictionParam for the input parameter. The computation is carried out by updating the object contObsPredModParam:

  $$\text{contObsPredModParam} := \text{ComputePred}(\text{contObsPredModParam}, \text{PredictionParam}).$$

  The concrete statistical method used to compute $\hat{y}_k, \hat{v}_k$ is specified by defining a concrete subclass of PredictionParam. In our example, we have defined four time series models (random walks with regular, seasonal, and regular/seasonal differences and automatic ARIMA models), among which the one with the lowest $\hat{v}_k$ is automatically selected. Any alternative choice (e.g., with machine learning techniques) could easily be implemented by defining the corresponding subclass.

Hierarchy and layering principles are applied by internally constructing routines on the key-value pair data structure in terms of simpler data structures such as vectors. In addition, imputation routines can be embedded in the design of these classes and methods as an attribute of PredictionParam.

- The estimated standard deviation $\hat{\sigma}_k$ of observation errors is computed in the same way:

$$\text{contObsPredModParam} := \text{ComputeObsErrorSTD}(\text{contObsPredModParam}, \text{ObsErrorSTDParam}).$$

In this case, another abstract class ObsErrorSTDParam has been defined. Its concrete subclasses determine the statistical method to be used for the estimation. In our example, we have defined a subclass to estimate σ_k by maximum likelihood, using the historical double sets of raw and validated data. As before, imputation routines can also be embedded in the design of these classes and methods as an attribute of ObsErrorSTDParam.

- The estimated error probabilities $\hat{p}_k$ are also computed in the same way:

$$\text{contObsPredModParam} := \text{ComputeErrorProb}(\text{contObsPredModParam}, \text{ErrorProbParam}).$$

In this case, an abstract class ErrorProbParam is defined. Its concrete subclasses determine the statistical method to be used for the estimation. In our example, we have defined a subclass to estimate p_k by maximum likelihood, using the historical double sets of raw and edited data. Again, as before, imputation routines can also be embedded in the design of these classes and methods as an attribute of ErrorProbParam.

- The sampling weights ω_k are usually computed at an earlier stage of the production process, so that we can simply retrieve them from some other data set in the survey in question. In other cases, if explicitly needed for the editing phase, the computation of the sampling weights can be carried out along similar lines.

Next, as parameters InputParameter in our error moments computation, we essentially need to specify the loss function $L(\cdot,\cdot)$. We will denote this object by ErrorMomentParam.

Finally, the output object OutputData will be denoted by ErrorMoments and is basically an array of error moment matrices $[M_{kl}^{(q)}]$ per population cell (q denotes the turnover and the number of employees in our example). In this way, we already have the content of each object in the expression

$$\text{ErrorMoments} := \text{ComputeErrorMoment}(\text{contObsPredModParam}, \text{ErrorMomentParam})$$

The whole computation at the third (scripting) layer is thus executed just by calling something like

$$\begin{aligned} \text{DD} &:= \text{readFile}(\text{DataDictionaryFile}) \\ \text{contObsPredModParam} &:= \text{buildcontObsPredModParam}(\text{DD}) \\ \text{PredictionParam} &:= \text{buildPredictionParam}(\text{and so on}) \end{aligned}$$

```
contObsPredModParam := ComputePred(contObsPredModParam, PredictionParam)
ObsErrorSTDParam := buildObsErrorSTDParam(and so on)
contObsPredModParam := ComputePred(contObsPredModParam, ObsErrorSTDParam)
ErrorProbParam := buildErrorProbParam(and so on)
contObsPredModParam := ComputePred(contObsPredModParam, ErrorProbParam)
SamplingWParam := buildSamplingWParam(and so on)
contObsPredModParam := ComputePred(contObsPredModParam, SamplingWParam)
ErrorMoments := ComputeErrorMoment(contObsPredModParam, ErrorMomentParam)
```

In the construction of the diverse parameters objects, the same hierarchical scheme can be followed (including e.g., the imputation routines). Notice also the far-reaching consequences on the organisation of work and the production process at a statistical office. Firstly, survey managers and domain experts can work at a scripting level with high-level functions such as ComputePred, ComputeObsErrorSTD, and ComputeErrorProb above. This does not demand extensive IT skills and they can concentrate on the adapted use of these tools to their concrete survey. Indeed, the modularity allows them to seamlessly combine and choose diverse alternatives to compute the parameters and the error moments according to the characteristics of the statistical operation. On the other hand, developers and methodologists (ideally data scientists) can work at a lower level, implementing new statistical methods as new subclasses and overloaded methods. Needless to say, for an optimal design of classes and methods, communication between both layers is recommended. Notice however that both the naming conventions and the division in modules (both functions and libraries) derives directly from the application of the foregoing principles: it is the statistical methodology which should define the borders (interfaces) between the different modules. This paves the way for easy application of standard good practices in software development, supported by a strong mathematical basis. In the current development and implementation of our proposal, we can only offer an empirical view on this particular production stage (editing). However, if these principles are to be applied throughout the process, the different functional modules should similarly interface with one another, thus coping with complexity.

Secondly, this architecture favours software evolution and ease of maintenance over code preservation (Booch et al. 2007). Legacy code is recognised as a heavy ballast in the modernisation of statistical production. We are not providing solutions for the existing legacy code, but this architecture philosophy helps a great deal by not producing legacy code. The code can evolve according to new needs detected in the statistical programs, by defining new subclasses and methods. At the same time, the produced code is easily maintained, since execution statements such as the one above seldom change.

Thirdly, since statistical methods are implemented with an abstraction of concrete statistical operations, the same code at the lower level and very similar at the scripting level is valid for different surveys. This allows us to optimally manage resources among statistical operations, as the methodology and computer tools are standardised.

Fourthly, we would like to comment on the granularity of the services and computer tools. In our example above, by starting with Formulas (1) and (2), we also want to suggest that the statistical methodology should determine the degree of granularity of computer

tools that implement the different methods. In the modular design, the statistical methods themselves should determine the natural borders among modules (hence also their interconnecting interfaces). Furthermore, the internal components of each module should also be structured according to the statistical methodology. Note how, in our example above, each parameter entered into Formulas (1) and (2) is dealt with using an independent method on the object contObsPredModParam, because each parameter can be computed/estimated choosing an adequate statistical method. Should new methodological proposals appear for a concrete computation, these can easily be incorporated without affecting the other software routines (e.g., imputation routines).

Finally, we would like to underline how the scripting philosophy fits perfectly well in the GSDEMs as a processing step, in which input statistical data and input metadata, process details, and transformed statistical data and output metadata are clearly expressed (UNECE 2015). Although we have not yet used this process architecture to manage process metrics, we are convinced that these monitoring parameters can also be computed along similar lines. This may be carried out by complementing each computation or transformation on an input data set with a chosen set of indicators in the output monitoring the transformation.

In conclusion, we must mention that in addition to the foregoing technical, mathematical difficulties, a highly relevant element of the practical implementation of this proposal is the staff reaction to changes in the production system. In the current stage of prototyping in production in a few statistical operations, the role of survey managers has been identified as key, since, in our current production model, they take the decisions on each survey. The gap between statisticians and computer scientists (and their traditional skills) also stands out as an aspect that needs to be addressed further.

6. Implementation: A Proof of Concept

The principles of functional modularity have been applied by designing and developing independent software packages for concrete aspects of this data organisation and process design. There are many aspects of the implementation worth sharing in order to be acquainted with the interplay between theoretical proposals and the practicalities arising in an ongoing production system at a statistical office.

Firstly, since both object-oriented and functional paradigms lie at the core of the proposal, the natural choice for a programming language is one that naturally supports these paradigms, without syntax quirks and twists. Java, C++, R, Python, Scala and many others are candidates that fulfill this condition. Since the user domain is clearly statistical data processing, another requisite is feasible rapid development of trustworthy statistical tools. Finally, a good documenting system of classes, methods, and functions is also desirable, which allows us to document data and parameter inputs, output, and throughput of each element (the process statistical metadata). These considerations led us to choose R (R Core Team 2012; Chambers 2008).

Secondly, the methodology of software development has also been carefully decided. Instead of the more classical waterfall model (see e.g., Palmquist et al. 2013), we have used a spiral approach (Boehm 1988). Thus, instead of compiling specifications, designing, coding, and testing in a linear way, we have incrementally agreed on a first

round of specifications, made a first design implemented on a first version of several R packages, and constructed a first version of the repository with key-value data files for three different short-term business statistics surveys. In this first round, the physical layer (the files themselves), the programming layer (classes, methods, and functions: the R packages), and the scripting layer were constructed. In a second round, apart from bugs and flaws in some functions detected in the testing phase, an important redesign was discovered to be necessary in the classes and methods implementation. The technical reason was that, for performance reasons in order to handle these key-value pair data sets, our packages heavily depend on the package data.table (Dowle and Srinivasan 2016). Formerly we used the S4 system of classes and methods, and the method dispatch, which suspends the lazy evaluation, is thus incompatible with the data.table syntax. We migrated all key-value data packages to the system S3. This affected the second layer, and interestingly enough, it did not affect the scripting layer. Along this line of work, we pursue the production of constantly evolving pieces of software that can adapt quickly and straightforwardly to the needs and changes of production. Again, this change of philosophy is at odds with the traditional culture at a statistical office and requires formidable management efforts in order to implement it at the officewide scale. For example, the idea that computer tools built in this way are not completed and ready for use in production may be risky, since it may lead to rejection of the methodology due to immature tools. These, more agile, methodologies also allow us to make more rational use of scarce resources, since development is incremental. In our view, a mindset change to perceive software as constantly evolving, rather than as a closed definitive tool is necessary for the industrialisation and modernisation of statistical production.

Thirdly, as a byproduct of the preceding methodology, communication between domain experts and survey managers, on the one hand, and developers and methodologists, on the other hand, must be clearly stressed. Although the architecture makes the work of both profiles independent by defining programming and scripting layers, an optimal system design will be achieved when communication between both parts is at a maximum during the development stage. Again, we face a management challenge that may impinge on organisational aspects of the whole statistical office (e.g., does it make sense to differentiate between statistical methodology and statistical software development departments?).

Fourthly, the different actors' computer skills must be taken into account. Two further actions have been taken in this regard to deploy the preceding architecture at the Spanish National Statistics Institute. On the one hand, the file containing the data dictionary is an XML file for machine readability. This technology does not form part of regular computer skills of domain experts and survey managers. Thus, to build this file, we asked these statisticians to record the specifications of each statistical variable in their survey in an Excel file with a prespecified structure. Excel files, although limited when dealing with some data structures, are easily handled. Then, we programmed a specific function, building the data dictionary file automatically from this Excel file.

Fifthly, the statistical computing system used as a standard at the Spanish National Statistics Institute is SAS, and following this institutional policy, computing routines used by survey managers and domain experts must be written in SAS, and not in other languages such as R, Python, Scala, and so on. Thus, the fourth layer, working as a user

interface, has been developed in the form of extremely simplified SAS macros that execute the aforementioned R scripts in batch form. This means that the interaction between the user and the architecture occurs only in SAS (so far, this has only been accomplished to feed and read from the repository; the selective editing routines are executed directly by data collection staff in simplified R scripts). Although the functionality of the system is currently severely reduced and rigidity is increasing, ease of use is noticeable, as the user only needs to specify a few very generic parameters.

Finally, the collection of packages in constant evolution at various stages of maturity are available in GitHub (Esteban et al. 2017a,b,c,d,e,f,g,h,i,j,k,l; Sanguiao 2017). The architecture behind these packages closely follows the statistical methodology of the optimisation approach to selective editing. Thus, it is difficult to give a precise description of what each package does without entering into mathematical content. A summarised description of what each package does can be found in Esteban et al. (2017m). It is important to point out that this division into many different packages that focus on concrete aspects of the statistical process should not be read just as an example of good practices in programming, but also as a consequence of the identification of functional modules according to the underlying statistical methodology.

7. Conclusion and Future Prospects

The main conclusion from this work is that in recognising an official statistical production system as a human-generated complex system, the principles of functional modularity can be used to cope with this complexity of designing both data and process architectures adapted and adaptable to the evolving needs of statistical production. By moving a step in the direction of international standards, we can combine the object-oriented and functional paradigms to define functional modules for the different production tasks whose borders and interacting interfaces are naturally determined by means of the underlying statistical methodology. These principles drive us genuinely to a set of layers in the statistical methodology, over its implementation in lower – and higher – level production tasks and steps to a top-orchestrating user interface.

The data organisation essentially revolves around a key-value pair data model, where keys are composed of statistical metadata. The process architecture implements transformations over information objects, thus combining both paradigms. In our view, these architectures bring relevant benefits to an efficient production system. They provide due roles for the different professional profiles in a statistical office, favour the evolution of software, thus adapting to new needs, lead to complete global parametrisation of every single datum in the process, and lead to standardisation in the production tools in surveys and statistical programs of various types.

Some of the elements presented in preceding sections are connected with the concrete production system at the Spanish National Statistics Institute. Therefore, it is advisable to recognise those elements that are exportable to other offices. Regarding the data architecture, the core element is the use of metadata to identify values. The key-value pair structure could be substituted by alternative data models, such as the SDMX or DDI. Nonetheless, in a deeper stage of analysis, performance issues (among others) should be taken into account in making a choice. In our case, we can process monthly data sets of

around 2 million lines and about 15 qualifiers (around 28,000 business units) to construct their corresponding traditional data matrices in less than two seconds. Regarding the process architecture, the core elements are (i) the application of functional modularity to statistical methods to produce modular computations respecting the natural borders in statistics, (ii) the layers organising the production tasks at different degrees of modularity, (iii) the use of object-oriented modelling for the information objects (both data and parameters), and (iv) the use of the functional paradigm to carry out the chained transformations on these information objects. All other implementation details can be adapted to concrete circumstances.

Nonetheless, our proof of concept reveals relevant challenges ahead. To be more efficient, an agile software development methodology should be preferred over more static methodologies. Also, it is important that the existing gap between methodologists/statisticians and computer scientists/developers must be bridged. All this pushes us to improve communication standards among the different actors (methodologists, computer scientists, domain experts, survey managers, business managers, and so on) within an office. This a remarkable management exercise.

Along this line, as stakeholders in and members of the ESS we recognise that alignment with international initiatives is a strategic matter. Thus, in future revisions and developments, alignment with CSPA services and European standards will be taken into account and pursued. Previously, technical requisites to be CSPA-compliant and to achieve shareability of computer application must be agreed upon by the international community (see, for example, the 2017 meeting report of UNECE 2017a).

8. References

Arbués, I., P. Revilla, and D. Salgado. 2013. "An optimization approach to selective editing." *Journal of Official Statistics* 29: 489–510. Doi: http://dx.doi.org/10.2478/jos-2013-0037.

Boehm, B. 1988. "A spiral model of software development and enhancement." *IEEE Computer* 21(5): 61–72. Doi: http://dx.doi.org/10.1145/12944.12948.

Booch, G., R.A. Maksimchuk, M.W. Eagle, B.J. Young, J. Conallen, and K.A. Houston. 2007. *Object-oriented Analysis and Design with Applications*. Addison-Wesley.

Chambers, J.M. 2008. *Software for Data Analysis*. Springer.

DDI Alliance. 2018. *Data Documentation Initiative 2018*. Available at https://www.ddialliance.org/ (accessed November 05, 2018).

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. Wiley.

Dowle, M. and A. Srinivasan. 2016. *data.table: Extension of 'data.frame'*. Available at https:// CRAN.R-project.org/package=data.table. R package version 1.10.0.

Esteban, E., S. Saldaña, and D. Salgado. 2017a. *RepoTime: Implementation of a notation for time intervals*. Available at https://github.com/david-salgado/RepoTime. R package version 0.2.2.

Esteban, E., S. Saldaña, and D. Salgado. 2017b. *StQ: Tools to manage metadata-incorporated keyvalue pair datasets*. Available at https://github.com/david-salgado/StQ. R package version 0.4.34.

Esteban, E., S. Saldaña, and D. Salgado. 2017c. *RepoReadWrite: Read and write files from/to the microdata repository*. Available at https://github.com/david-salgado/RepoReadWrite. R package version 0.4.5.

Esteban, E., S. Saldaña, and D. Salgado. 2017d. *RepoUtils: Implementation of tools to map and work with repositories*. Available at https://github.com/david-salgado/RepoUtils. R package version 0.1.2.

Esteban, E., S. Saldaña, and D. Salgado. 2017e. *contObsPredModelParam: Class and methods for the parameters of a continuous observation- prediction model*. Available at https://github.com/david-salgado/contObsPredModelParam. R package version 0.0.1.

Esteban, E., S. Saldaña, and D. Salgado. 2017f. *StQPrediction: Define S4 classes and methods to make predictions*. Available at https://github.com/david-salgado/StQPrediction. R package version 0.0.1.

Esteban, E., S. Saldaña, and D. Salgado. 2017g. *StQImputation: Classes and methods to implement different imputation methods upon StQ objects*. Available at https://github.com/david-salgado/StQImputation. R package version 0.0.1.

Esteban, E., S. Saldaña, and D. Salgado. 2017h. *SelEditErrorMoment: Compute the conditional measurement error moments under the optimization approach to selective editing*. Available at https://github.com/david-salgado/SelEditErrorMoment. R package version 0.0.1.

Esteban, E., S. Saldaña, and D. Salgado. 2017i. *SelEditFunctions: Functions for selective editing*. Available at https://github.com/david-salgado/SelEditFunctions. R package version 0.0.1.

Esteban, E., S. Saldaña, and D. Salgado. 2017j. *SelEditUnitPriorit: Classes and methods to implement unit prioritization*. Available at https://github.com/david-salgado/SelEditUnitPriorit. R package version 0.0.1.

Esteban, E., S. Saldaña, and D. Salgado. 2017k. *TSPred: Point and std prediction of time series*. Available at https://github.com/elisa-esteban/TSPred. R package version 0.2.5.

Esteban, E., S. Saldaña, and D. Salgado. 2017l. *BestTSPred: Construction of objects of class BestTSPredParam*. Available at https://github.com/elisa-esteban/BestTSPred. R package version 0.0.1.

Esteban, E., S. Saldaña, and D. Salgado. 2017m. Software implementation of optimization-based selective editing techniques at Statistics Spain (INE). UNECE Work Session on Statistical Data Editing. The Hague, 24–26 April 2017. Available at https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2017/mtg2/Paper_19_StatSpain.pdf (accessed November 05, 2018).

Eurostat. 2014a. ESS Vision 2020. Available at http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020.

Eurostat. 2014b. Vision 2020 Implementation Portfolio. Available at http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020/implementation-portfolio.

HLG-MOS. 2011. "High-Level Group for the Modernisation of Official Statistics. Strategic vision of the High-Level Group for strategic developments in business architecture in Statistics." *Conference of European Statisticians Geneva. 59th Plenary Session*. 14–16 June, 2011. Working Paper 1. Available at https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2011/1.e.pdf.

HLG-MOS. 2017. High-Level Group for the Modernisation of Official Statistics. UN-ECE Statistics Wikis. Available at http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Official+Statistics.

Informal Task Force on Metadata Flows. 2013. "Metadata flows in the GSBPM." *Work Session on Statistical Metadata*. Geneva, 6–8 May, 2013. Working Paper 22. Available at https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2013/WP22.pdf (accessed November 05, 2018).

Lundell, L.-G. 2013. Framework of metadata requirements and roles in the SDWH. ESSnet on microdata linking and data warehousing in production of business statistics. Deliverable 1.1. Available at https://ec.europa.eu/eurostat/cros/content/dwh-sga2-wp1-11-metadata-framework-statistical-data-warehousing-v112-final_en.

Palmquist, M.S., M.A Lapham, S. Miller, T. Chick, and I. Ozkaya. 2013. Parallel worlds: agile and waterfall differences and similarities. Technical Note CMU/SEI-2013-TN-021. Software Engineering Institute. Carnegie Mellon University. Available at http://repository.cmu.edu/cgi/viewcontent.cgi?article=1761&context=sei.

Pearson, J.W., S. Olver, and M.A. Porter. 2017. "Numerical methods for the computation of the confluent and Gauss hypergeometric functions." *Numerical Algorithms* 74: 821–866. Doi: http://dx.doi.org/10.1007/s11075-016-0173-0.

R Core Team. 2012. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Available at http://www.R-project.org.

Saltzer, J.H. and M.F. Kaashoek. 2009. "Principles of computer system design: An Introduction. Morgan Kaufmann, 2009. ISBN: 978-0-12-374957-4.

Sanguiao, L. 2017. *Transformation of Standard Questionnaires*. Available at https://github.com/Luis-Sanguiao/StQT. R package version 0.1.0.9000.

UNECE. 2013a. Generic Statistical Business Process Model. Version 5.0. Available at http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model.

UNECE. 2013b. Generic Statistical Information Model. Version 1.1. Available at https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model.

UNECE. 2015. Generic Statistical Data Editing Models. Version 1.0. Available at https://statswiki.unece.org/display/kbase/GSDEMs.

UNECE. 2017a. Statistical Data Editing Work Sessions. Available at http://www1.unece.org/stat/platform/display/kbase/UNECE+Work+Sessions+on+Statistical+Data+Editing.

UNECE. 2017b. Capabilities and Communication Group. Available at http://www1.unece.org/stat/platform/display/MCOOFE/Capabilities+and+Communication+Group%3A+Home.

Van der Loo, M. 2015. A formal typology of data validation functions. UNECE Work Session on Statistical Data Editing. Budapest, 14–16 September 2015. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2015/mtg1/WP_5_Netherlands_A_formal_typology_of_data_validation_functions.pdf (accessed November 05, 2018).

Van Roy, P. and S. Haridi. 2004. "Concepts, Techniques, and Models of Computer Programming." MIT Press.

Weinberg, G.M. 2011. "An introduction to General Systems Thinking." Weinberg and Weinberg. ISBN: 978-0-93-263349-1.

Wickham, H. 2014. "Tidy data." *Journal of Statistical Software* 29(10): 1–23. Doi: http://dx.doi.org/10.18637/jss.v059.i10.

Received June 2017
Revised March 2018
Accepted June 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 835–862, http://dx.doi.org/10.2478/JOS-2018-0042

Efficiency and Agility for a Modern Solution of Deterministic Multiple Source Prioritization and Validation Tasks

Annalisa Cesaro[1] *and Leonardo Tininini*[1]

This article focuses on a multiple source prioritization and validation service. We describe a modern rule-based, loosely coupled solution. We follow generalization, efficiency and agility principles in application design. We show benefits and stumbling blocks in micro-service architectural style and in rule-based solutions, where even the selection task is solved through selection rules, which encapsulate the calls to Entity Services, allowing access to input-sources. We allowing the rule-based service efficiency and further local and remote input data selection scenarios for the validation Statistical Service. In particular, data virtualization technologies enable architects to use remote sourcing and further increases agility in data selection issues. Through a wide number of experimental results, we show the necessary level of attention in process implementation, data architectures and resource usage. Agility and efficiency emerge as drivers which possibly sustain the Modernization flexibility impetus. In fact, flexible services may potentially serve multiple scenarios and domains.

Key words: Rule engine for validation and prioritization; highly-performant data management; efficient data parallelism; data virtualization; agile culture.

1. Background

Over the last decade, National Statistical Offices (NSOs) have been investing heavily in new approaches to improve and make more flexible their data supply chains. Modernization efforts in official statistics require the reuse and sharing of methods, components, processes and data repositories.

1.1. Sharing and Reuse Needs: The SOA Impetus

There is a growing trend to introduce Service-Oriented Architectures (SOAs) in official statistics due to their promise of cost efficiency, agility, adaptability and legacy leverage (O'Brien et al. 2008). SOA is based on sharable independent services, which should be (i) efficient, thus being without resource waste and possibly usable for small and large data sets; (ii) agile, thus easily managing constantly evolving scenarios; and (iii) generalized, thus applicable in cross-cutting domains. It is worth noting that several SOA architectural styles exist (Quensel-von Kalben 2017a), for example (i) point to point integration;

[1] Italian National Institute of Statistics (Istat), via Balbo 16, 00184 Roma, Italy. Emails: cesaro@istat.it and Tininini@istat.it

Acknowledgments: Monica Scannapieco for Modernisation discussion, Monica Consalvi and Francesca Alonzi for SBR validation rules, Laura Maglione for enterprise metadata knowledge and legacy validation rules, Marco Passacantilli for java web app refactoring/development.

(ii) platform integration, based on an Enterprise Service Bus, which interconnects mutually interacting components through events or messages management; and (iii) lightweight basic integration, by using autonomous fine-grained micro-services, which are unaware of their position in the process chain/control flow (Fowler 2014; Namiot and Sneps-Sneppe 2014). In the literature, there is still a lack of consensus on what micro-services actually are (Dragoni et al. 2017). A micro-service should maintain focus on providing a single business capability, moreover each micro-service should be operationally independent from others. In such architecture, the only form of communication among services is through their interfaces. Many migration patterns towards a SOA exist in the literature, as outlined by Razavian and Lago (2015) and Khadka et al. (2012).

As highlighted by Quensel-von Kalben (2017b), Enterprise Architecture is a fundamental driver in official statistics modernization actions. The Business Architecture of reference for official statistics is the Generic Statistical Business Process Model (GSBPM), which identifies business functionalities that need to be supported by IT systems. The processes and sub-processes of the GSBPM may rely on different information models (GSBPM v5.0 2017). IT systems, fulfilling the business needs, define the application architecture, which is recommended to be migrated towards a SOA. When SOA is used, the Common Statistical Production Architecture (CSPA) should be taken into account (ESSnet 2015). CSPA states the main design principles to be followed when assessing the design of the business, information and application architectural layers. Briefly, CSPA-compliant services rely on the matching of the service functionalities with one or more GSBPM activities, as well as on non-functional requirements, such as performance (i.e., resource utilization, time behavior and capacity), scalability, security, and language. The abstract information model of reference relies on the Generic Statistical Information Model (GSIM), which could be mapped onto more specialized information models that are in use in official statistics. When used, GSIM enables harmonization in service definition and greater decoupling between statistical domain experts and Information Technology (IT) ones. Before CSPA, Eurostat organized CORE (ESSnet Core Project 2011), which designed a platform to orchestrate GSBPM-compliant Statistical Services: it promoted the idea that the data model – that is, the inputs and outputs of the services – might be described through the GSIM. However, service-sharing across NSOs is a difficult activity and remains a work in progress, as outlined by Quensel-von Kalben (2017a).

1.2. Data Virtualization for Avoiding Silos: A Focus on Performance

Dealing with different types of data sources is also a challenging issue in modern IT systems. Official statistics require many different sources to be integrated in the statistical process. In particular, integration may involve (i) large and unstructured data collections, (ii) performing classical relational, and optimized data management, and for official statistics, either (iii) Resource Description Framework (RDF) data, when standardized and linked open data are used, or (iv) Statistical Data and Metadata eXchange (SDMX) data.

Data virtualization is a relatively new approach to data selection and integration (Pullokkaran 2013) that avoids physically moving data into a single integrated environment and reduces the risk of integration silos (Pullokkaran 2013; Alagiannis

et al. 2012; Karpathiotakis et al. 2015). Several data virtualization patterns exist: (i) data may be passed directly to an execution engine for query processing and then discarded; (ii) data may be cached in memory for subsequent processing and then discarded; or (iii) data may be temporarily written to disk for prompt subsequent processing and then discarded (Idreos et al. 2011; Cheng and Rusu 2015). Multi-source data integration and/or data cleansing and transformation might hence be defined in a logical layer and then applied to data as they are retrieved from the data sources while generating reports (Pullokkaran 2013; Krawatzeck et al. 2015). Databases may be built by launching queries, instead of building databases for launching queries (Karpathiotakis et al. 2015). Such techniques are used nowadays in, for example, the agile Business Intelligence (BI) context (Stodder 2013; Van Der Lans 2013).

Performance is often evaluated in terms of execution time (Karpathiotakis et al. 2015; Alagiannis et al. 2012; Tian et al. 2017) and parallelism can be a rewarding technique in virtual loading (Cheng and Rusu 2015). For an increased performance, service replication in distributed systems may be tackled (Osrael et al. 2006; Chen et al. 2014; Mohamed 2016; Xie et al. 2017). Server resource consumption has to be considered (Xavier et al. 2013) as well. Conversely, when data locality is ensured, data replication and data consistency issues have to be taken into account (Montoya et al. 2017), as well as storage and energy consumption (Milani and Navimipour 2016). The same architectural issues are common in official statistics, where replicated services versus shared services are evaluated in the European Statistical System network (ESSnet) context (Gramaglia 2015).

1.3. Prioritization and Validation Tasks: Rule-Based Solution

Hence, modernization impetuses move official statistics towards SOA in order to react promptly to ever-evolving scenarios and towards heterogeneous data integration to increase the level of quality in relation to some quality components and, possibly, the number of statistical outputs. Such impetuses should be taken into account when deciding IT solutions for a GSBPM activity, and specifically in relation to deterministic prioritization and validation tasks. In the latter case, besides GSBPM, relevant references are: (i) Generic Statistical Data Editing Models (GSDEM) (GSDEM 2015), which is a generic process framework for statistical data editing that focuses on the "Review and Validate" and "Edit and Impute" GSBPM phases; and (ii) specifically on the methodology for data validation (Di Zio et al. 2016), defined in the ESSnet context, which focuses on the "Review and Validate" GSBPM phase. In both cases, rule-based solutions are commonly used as methods for deterministic multiple source prioritization and validation tasks. In data validation "the decisional procedure is generally based on rules expressing the acceptable combinations of values" (Di Zio et al. 2016, 6), and in data editing "edit rules, score functions, correction rules and error localization rules" may be collected (GSDEM 2015, 15). Briefly, when rules are used, they are isolated from the software code, independently managed and customized by expert domain users, and may be accessed by different technological solutions, thus effectively executing the task. The rules are treated as data and not as parts of a source code of a program, thus enhancing rule re-use, sharing and increasing agility. Users with different roles may contribute to specific aspects

(e.g., the robustness of the validation rules in relation to a given task and quality objective, and also the robustness of the rules evaluation IT process).

1.4. Rule Engine

A rule-based infrastructure relies on a rule-processing engine (referred to as rule engine below), which is a component that evaluates which statistical unit meets the condition stated in a rule, and performs the corresponding imputation or correction actions, if any. In general, business solutions based on rules can be found in the literature in different contexts. Several examples can be found in the field of artificial intelligence, for the construction of expert systems (Lavrac 2001), in which expert knowledge in a given domain is represented with structures of the IF-THEN type (rules), which relate information or facts to some action. The architecture of an expert system includes a knowledge base, an inference engine, and a user interface that allows expert reviewers to interact with the facts and rules and maintain the system. Therefore, an expert system does not necessarily require the involvement of a database. The literature in this context define formal frameworks that study rule languages and define rule engines for processing specific language rules by assessing performance (Liang et al. 2009). Rule engines are also used in the literature to expand probabilistic knowledge bases (Chen and Wang 2014; Zhou et al. 2016). Currently, numerous processes of knowledge extraction from unstructured documents have also been proposed (De Sa et al. 2016), such as those available on the web, in order to build structured knowledge bases. Rule-based validation is also supported in the European Statistical System context, in which a formal framework for rule definition has been realized. An example is the Validation and Transformation Language (VTL), defined by an innovation project (Schafer 2015), which aims to be a single reference language for harmonizing the validation approaches among NSOs (Gramaglia 2015; Di Zio et al. 2016; ESSnet ValiDat Integration 2017). Currently, feasibility studies are being carried out in the ESSnet context to assess the possibility of defining converters from VTL to other languages used in national contexts, such as SQL (ESSnet ValiDat Integration 2017).

1.5. The Modern Solution for Rule-Based Prioritization and Validation Tasks

Current IT practices in deterministic rule-based validation tasks are outlined by Quensel-von Kalben (2017b). We adopt a SOA approach, by using a lightweight holistic micro-service (Fowler 2014). It is independently deployable and scalable. User interfaces allow expert staff to manage the rules and view the reports produced in relation to each rule-based task. The rule engine, which is the core of the service, is based on generalization and efficiency principles. It relies on an optimized data schema and on data parallelism in processing. It has sufficient efficiency to manage ever-evolving scenarios (i.e., small as well as big ones), and makes rule-based processing competitive with respect to other technological solutions. Therefore, a question arises whether rule-based tasks may exhibit the data selection agility property. Efficiency and agility in selection could be increased when inserting the input data selection logic within rules (i.e., selection rules) (Karpathiotakis et al. 2016). Therefore, selection rules may encapsulate the input source calls to services, which expose and allow access to input source data (i.e., referred to as

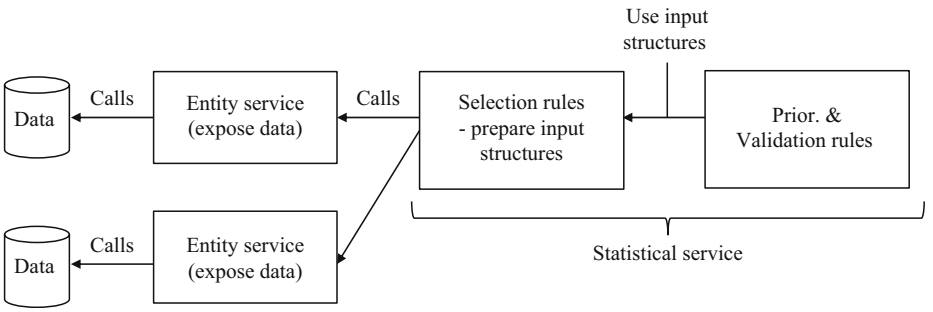

Fig. 1. Logical SOA architecture with Entity Services and Statistical Services: selection rules adapt input data for subsequent processing for prioritization and validation purposes.

Entity Services below) and adapt input data in structures for subsequent processing (i.e., the prioritization and validation rules evaluation). Entity Services may expose single as well as integrated datasets, and local as well as remote data, thus enabling architects to use data virtualization solutions in the input selection task. Entity Services expose the input data required by a Statistical Service, as sketched in Figure 1. Following virtualization principles could further increase agility. Selection rules could also interface with a logical integration layer or manage several source data clients in a transparent way for local and remote sourcing. Such rules allow the evaluation between data-replication architectures versus service-replication ones.

1.6. The Statistical Domain of Reference

The designed service has been used in a widespread manner in the Italian Statistical Business Registers (SBR) context. Briefly, Business Registers are updated yearly by integrating administrative and statistical sources, enabling identification of active statistical units and the estimation of the main structural, economical and identification variables for each unit using a robust methodology. Register data production may require the integration of an ever increasing number of evolving administrative sources and statistical lists, the integration of data from surveys, and integration of data from new unstructured web scale sources. In this evolving context, many deterministic prioritization and validation processes are needed, thus increasing quality standards (e.g., accuracy, comparability, coherence of a statistical output).

The rest of the article is organized as follows. In Section 2, we define the business and information models for deterministic integration and validation tasks. In Section 3 we briefly describe the use cases that concern the statistical user interactions with the IT system for rule and task report customization. We further describe the use cases for task processing, which involve the rule-processing engine. Specifically, we highlight the relevance of efficiency as a driver for providing usability and flexibility, and virtualization as a driver for increasing agility in input data sourcing. In Section 4, we show the benefits of an efficient and scalable rule-engine system and how the input data selection logic may be embedded within rules. We finally assess different technological solutions for remote sourcing. Conclusions are presented in Section 5.

2. Rule-Based Validation and Integration: The Business and Information Models

In this section we define the prioritization and validation service. In particular, we describe its abstract information model and give some hints on rule identification, particularly for selection purposes.

2.1. The Business and Information Models

Rule-based integration and validation tasks may be performed in several GSBPM phases. We refer to a generic modern process, as depicted in Figure 2. An IDentifier (ID) may be matched to any collected unit data in relation to a specific statistical population. By using the identification attributes, the unit may therefore be involved in the production process of a specific statistical output, integrating, possibly, multiple sources using common identifiers and requiring several processing steps. In rule-based integration and validation, specific-domain experts define the integration and validation rules, the necessary input variables from single or multiple sources, and the output variables useful for data validation or correction, which should be transmitted to a statistical output.

The tasks rely on a single base information set or *base table,* whose structure is depicted in Figure 3. Specifically, the *base table* is a statistical data set that is the object of the statistical process "Data prioritization and validation". "It is a collection of values. Conceptual metadata defines the meaning of these data by describing the concepts that are being measured by these data (concepts and definitions) and their practical implementation (value domains and data structure)" (GSDEM 2015, 13). Linkage and integration operations may be required to impute ID keys on source data before starting the integration and validation task. The *base table* fields include: (i) the statistical unit ID; (ii) other relevant secondary IDs, which may relate each unit with others for coupling or aggregation purposes; (iii) the input variables, which represent relevant unit characteristics; and (iv) possibly output variables (i.e., outcomes of the integration or

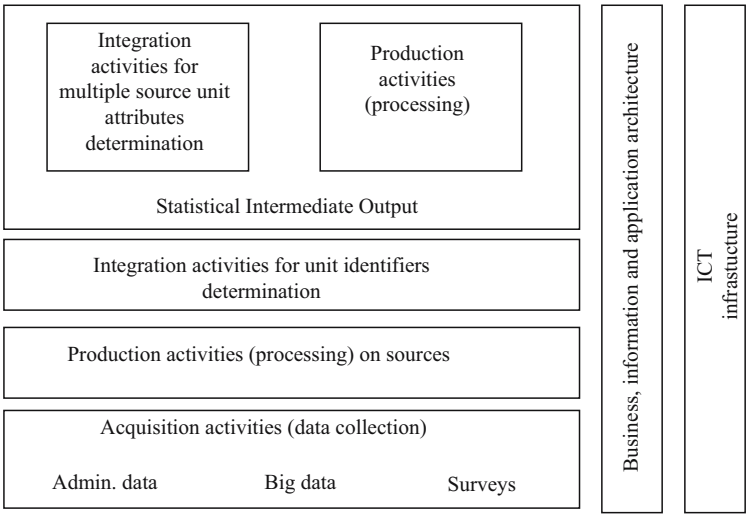

Fig. 2. Generic statistical process, which refers to GSBPM phases and GSIM terminology.

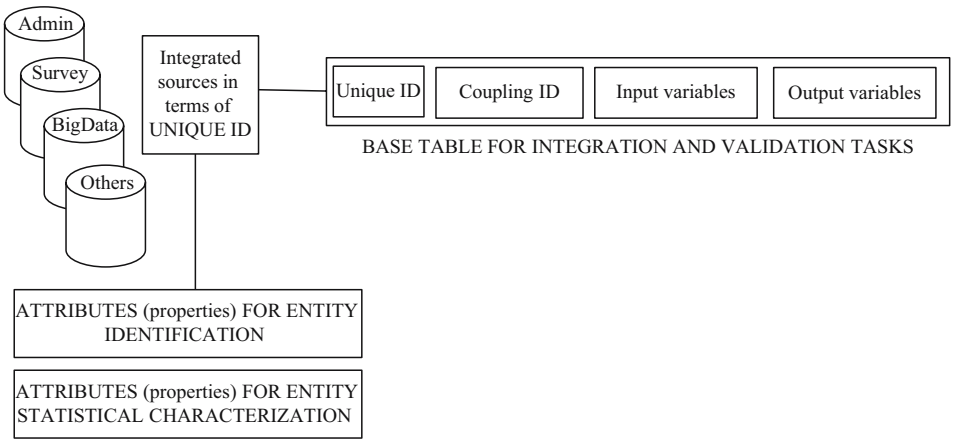

Fig. 3. Generic base table structure.

validation tasks, particularly in case imputation or editing actions are necessary). Rules are expressed in terms of the *base table* fields and any task relies on a single pair < rule set, base table >, that is, data structures and rules should rely on given metadata and be standardized. Metadata definition and common rule definition languages sustain sharing among domain experts. The *base table* is temporary and related to a specific task. Other downstream processes may transform the imputed, validated and/or edited values into a specific statistical output. Three different classes of rules may be specified:

1. *Selection rules*, which retrieve the necessary input data from sources (which may be local as well as remote with respect to the server that processes the validation task), using common ID keys, and define the *base table* input data for the validation task,
2. *Indicator rules*, which determine whether a given condition is met for each unit,
3. *Imputation rules,* which impute a specific value to a given variable under specific conditions.

Generally, GSDEM edit rules (which describe valid or plausible values for *base table* variables or *base table* combination of variables, and detect values presumed to be in error). GSDEM score functions evaluate input data values at unit level and GSDEM error localization rules presumed to be in error without a detectable cause. All these elements may be expressed through indicator rules. Moreover, GSDEM correction rules amend errors and may be expressed through imputation rules. Therefore, rules may exhibit GSDEM review, selection and amendment functions (GSDEM 2015, 8). Moreover, as outlined in Di Zio et al. (2016, 10), several validation levels may exist that review the logical and statistical consistency of the data and that involve more and more input information. Selection rules enable raising the validation level in relation to the business. Rules are expressed in a declarative way. Indicator and imputation rules rely on SQL clauses. Selection rules, when remote sourcing is used, may be based on either SQL and XML. We call them Xrules. The rules, formally specified by the statistical domain experts, are evaluated (i.e., applied) when sources are available. Each rule is uniquely identified (i.e., Unique ID, and Process ID in Figure 4), may or may not exhibit a selective condition

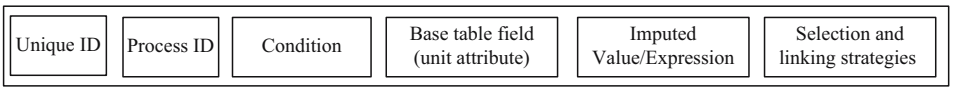

Fig. 4. Generic rule structure.

of validity in relation to the units within the *base table* (i.e., condition in Figure 4), may or may not set a given *base table* field, whose value may be the result of a specific combination of the input variables, or the result of aggregation operations computed on other *base table* fields (i.e., *base table* field unit attribute and imputed value in Figure 4). Xrules may be further based on a triple: (i) the call to a remote Entity Service, whose client has to be available to the calling server, for retrieving source data; (ii) a selection clause for taking into account only relevant information in virtually loaded remote data sets; and (iii) a linking clause for matching virtually loaded units with those within the *base table*. The rule-based integration and validation service may correspond to 5.1 (Integrate), 5.3 (Review and validate), 5.4 (Edit and Impute) and 5.5 (Derive new variables and statistical units) phases of the GSBPM (ESSnet 2015). A few relevant extracts from the GSBPM documentation (ESSnet 2015, 18) are given to assist the reader as follows:

"The 5.1 subprocess integrates data from one or more sources. The input data can be from a mixture of external or internal data sources, and a variety of collection modes, including extracts of administrative data. The result is a harmonized data set. Data integration typically includes:

1. matching/record linkage routines, with the aim of linking data from different sources, where those data refer to the same unit,
2. prioritizing, when two or more sources contain data for the same variable (with potentially different values)."

The integration phases may be performed sequentially. Specifically, we focus on the deterministic prioritization of data from different sources and on linking operations, which rely on unique units' IDs. Probabilistic or more complex record linkage operations are outside the scope of this document. They have to be performed elsewhere in the process chain.

The abstract information objects required by a generic integration service are shown in Figure 5. Integration evaluates conflicting source data and sets a single chosen value in output variables. The task returns the imputed data and a monitoring report.

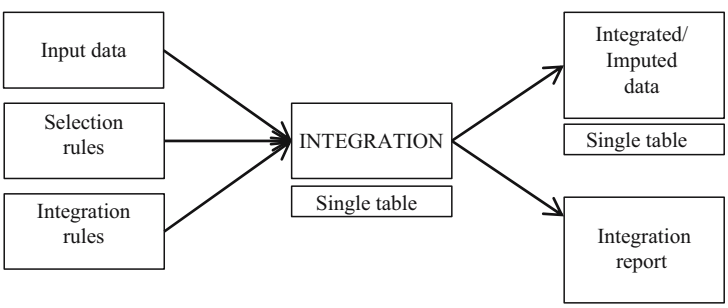

Fig. 5. Abstract information model of a general integration process.

Phase 5.3 "Review and Validate" of the GSBPM document (ESSnet 2015, 18) states;

"This subprocess applies to collected micro-data, and looks at each record to try to identify (and where necessary correct) potential problems, errors and discrepancies such as outliers, item non-response and miscoding. It can also be referred to as input data validation. It may be run iteratively, validating data against predefined edit rules, usually in a set order. It may apply automatic edits, or raise alerts for manual inspection and correction of the data. Reviewing, validating and editing can apply to unit records both from surveys and administrative sources, before and after integration. In certain cases, imputation (phase 5.4) may be used as a form of editing".

With respect to phase 5.4 "Edit and Impute", (ESSnet 2015, 19) it states:

"Where data are missing or unreliable, estimates may be imputed, often using a rule-based approach. Specific steps typically include: (i) the identification of potential errors and gaps; (ii) the selection of data to include or exclude from imputation routines; (iii) imputation using one or more predefined methods for example "hot-deck" or "cold-deck"; (iv) writing the imputed data back to the data set, and flagging them as imputed; and (v) the production of metadata on the imputation process";

And finally, phase 5.5, "Derive new variables and units", (ESSnet 2015, 19) is described as follows:

"This subprocess derives (values for) variables and statistical units that are not explicitly provided in the collection, but are needed to deliver the required outputs. It derives new variables by applying arithmetic formulae to one or more of the variables that are already present in the dataset. This may need to be iterative, as some derived variables may themselves be based on other derived variables. It is therefore important to ensure that variables are derived in the correct order. New statistical units may be derived by aggregating or splitting data for collection units, or by various other estimation methods. Examples include deriving households where the collection units are persons, or enterprises where the collection units are legal units".

Validation rules encapsulate deterministic conditions and actions, and may be expressed as algebraic expressions. In Figure 6, we sketch the main information objects that support the rule-based validation process. Validated data, a validation report, and further automatically edited data, whenever editing is possible through deterministic rules, are the main outcomes of the task. When manual editing is necessary, the output reports assist the expert user.

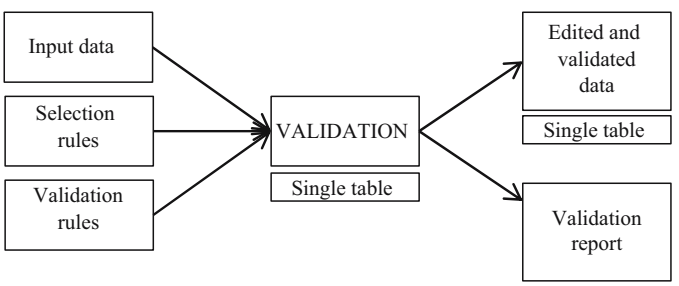

Fig. 6. Abstract information model of a general validation process.

2.2. *Hints for Rule Identification*

The logical flow of these processes has already been described. All units of the reference statistical population are imported into a single *base table*. Selection rules, through *existence* operators, enable the user to relate the *base table* with local or remote sources, storing temporarily input values to be used for validation and/or transformation tasks. Expert statisticians express their knowledge in terms of task rules, which are then applied on the input data. It is out of scope of the present document to show classical prioritization or validation rules, which evaluate different fields of the *base table* for detecting wrong or misleading values: data structures and rules depend on the specific domain and should be shared, based on metadata and possibly standardized. For details, the reader may refer to Di Zio et al. (2016) and GSDEM (2015) for an indepth analysis. We remark that, in practice, a large part of the deterministic rule requirements are satisfied through the use of logical, existence and inclusion operators. Moreover, the same rules could also be described in VTL for sharing and documentation purposes.

A sample selection rule is shown in Figure 7a. More complex queries may be required, for example when aggregation or algebraic functions must be performed on disjoint groups of units. A sketch of a generic aggregation rule on disjoint groups of units, identified by a given key, SETID, is shown in Figure 7b. These types of rules may also be used to redistribute some variables under certain conditions and to compute derived variables, for proportional inference, source prioritization, and integration and correction purposes, as explicitly outlined in Subsection 2.1. An example of a distribution rule may be as follows: a first derived variable sums up other variable values in relation to disjoint groups of the statistical units.

A second variable computes the number of units actively involved in any group, and finally a third variable represents the proportional imputation of the summed-up variable, that is, the proportional distribution of the aggregated variable in equal parts on the units actively involved in the group.

3. Micro-Service Architecture, Agile Cooperation, Efficiency and Data Virtualization for Service Reuse and Sharing

In this section, we highlight generalization, service efficiency and agility as application design principles sustaining flexibility, and hence service reuse and sharing. In particular, the micro-service architectural pattern, the agile cooperation, and data virtualization solutions may sustain service agility. An extensive performance assessment may evaluate service efficiency. The promoted design pattern, which relies on service autonomy in evolution and deployment and on efficiency in processing, has enabled widespread usage

Sample selection rule (a)	Sample aggregation rule (b)
Update BaseTable base Set field1 = (select genericField from SourceTable source where base.ID = source.ID) Where exists(select 1 fromSourceTable source where base.ID = source.ID)	Update BaseTable base Set field1 = (select sum(genericField) from BaseTable source where base.SETID = source.SETID) Where exists(select 1 fromBaseTable source where base.SETID = source.SETID)

Fig. 7. (a) Sample selection rule. (b) Sample aggregate rule on disjoint set of rules.

of the prioritization and validation service in the SBR domain. The highlighted principles are furthermore CSPA-compliant, and potentially facilitate service-sharing even across domains and organizations. CSPA principles sustain an inclusive design of components. Inclusion with respect to various domains, ever-evolving scenarios and input/output communications. When service flexibility is increased, a service might potentially be used in a widespread manner.

3.1. Micro-Service Architecture and Agile Cooperation

The implemented service relies on the micro-service architectural pattern: it is a lightweight basic one and relies on its own web user interfaces, a self-contained schema and an efficient rule-processing engine. All functionalities and data of a specific business capability are realized by an independent service, which can be deployed on specific hosts (Fowler 2014). A natural distribution of the workload sustains system efficiency and availability. In the case of increasing load, micro-service relocation and/or replication on a cluster, or in the cloud, may assure scalability. The deterministic integration and validation tasks have been vertically solved in an autonomous and stateless manner. The service takes the input data, processes the rules and generates the output reports. When it fails, it can be restarted without any dependence on previous states. The service is unaware of its position in the process chain/control flow. The service, which provides holistic system functionality, is independently deployable both with respect to user interfaces and to processing components, and may be independently progressed in both cases. In this section, we specifically describe the interfaces for user interaction (i.e., the use cases that involve web user interfaces). We describe data management and processing issues (i.e., the use cases that involve different actors with regard to the service users and relate specifically with task execution) in the following sections. Micro-service pattern sustains short agile software development cycles in a stepwise manner, and by involving users, also in the functional design phase and in the acceptance testing phases. Such development pattern could decrease difficulties in service reuse. In particular, through the developed web application, the domain expert users may manage and customize the task rules, view the output reports (i.e., the outcomes of the executed tasks), as well as download specific information sets in relation to the statistical units, whose characteristics met the condition of a rule during processing, and possibly manage metadata reporting. The user interfaces rely on a specialized Java application, based on a software architectural pattern similar to that used in the context of the COmmon Reference Environment (CORE) Project – Eurostat (Scannapieco et al. 2011). An expert user may insert new rules and modify or delete existing ones through specific web functionalities. Any rule is equipped with a customizable text field used for documentation purposes: it might also contain the equivalent VTL description, thus using a *lingua franca* to describe it. As already outlined, rule-based integration and validation provides an effective decoupling between the domain expert work and the IT work, providing flexibility in rules definition. Web functionalities for executing and monitoring rule-based processing could further increase the ability of users to run the service with less reliance on IT experts. A generic output report shows the number of units that met the rule conditions during rule-processing, and, if necessary, it may decompose such total number into subsets by classifying the units in

Report Title – ID PROCESS							
Id Rule	Level_rule	Condition_rule	Description_rule	Count_all	Subcount_col1	Subcount_col2	…
1	1	Example: SQL where clause	It could be the VTL description of the rule or a natural language one	501	101	203	197
……	……	……	……	……	……	……	……
Downloadable Lists Section – ID PROCESS & ID REPORT							

«Application Frequency» Values For Each Rule

Fig. 8. A skeleton sample output report (final check in SBR context): the overall number of units for which the rule condition triggered is shown. Such value is decomposed in relation to four disjoint classes of the involved units.

relation to relevant classes of data (e.g., in SBR validation context, the report may classify the outcomes in relation to not-active legal units and active legal units in relation to specific ranges of employees). Different reports, which sub-classify data differently, may exist for the same homogeneous set of rules. The report may be consulted as well as downloaded by expert users for further analysis, thus enabling the users to check which statistical units contributed to the count. Each unit in the downloaded lists may be equipped with a customizable information set (i.e., other relevant variables at unit level). Figure 8 shows a typical output report. Each row corresponds to a single rule. The total rule-validity frequency with respect to the involved statistical units is shown. Furthermore, such frequency value is decomposed in relation to disjoint classes of units.

3.2. Data Schemas Solutions for Highly-Performant Data Management

A data schema for supporting general rule-based processes has been designed. Generalization is preserved by parameterizing the concepts of interest, namely (i) the specific process/task associated with a fixed *base table* structure and rule set; (ii) the statistical units typology; (iii) the specific output report; and (iv) the downloadable custom information set. Generally, integration and validation tasks may involve millions of units, or more, and may also be based on hundreds of control/processing rules. In this scenario, each server, which hosts the rule-processing engine, may manage simultaneously numerous integration and validation tasks. Therefore, a highly-performant data management, which relies on efficiency principles, should be used. In classical theory, BI applications need to evaluate the tradeoff between minimization of data duplication and the increase in data duplication for reducing the cost of expensive integration operations among separated data structures (Pullokkaran 2013). In rule-based tasks, a question arises whether a schema, where duplication is carried on, may improve performance in processing and downloading by allowing each task to rely on a single data structure that maintains all the necessary information for a single task.

Specifically, the association between a valid rule and a statistical unit, whose characteristics met the rule condition during processing, may be stored in a separate data structure (i.e., a join table), as sketched in Figure 9, the same for each task, or in the *base*

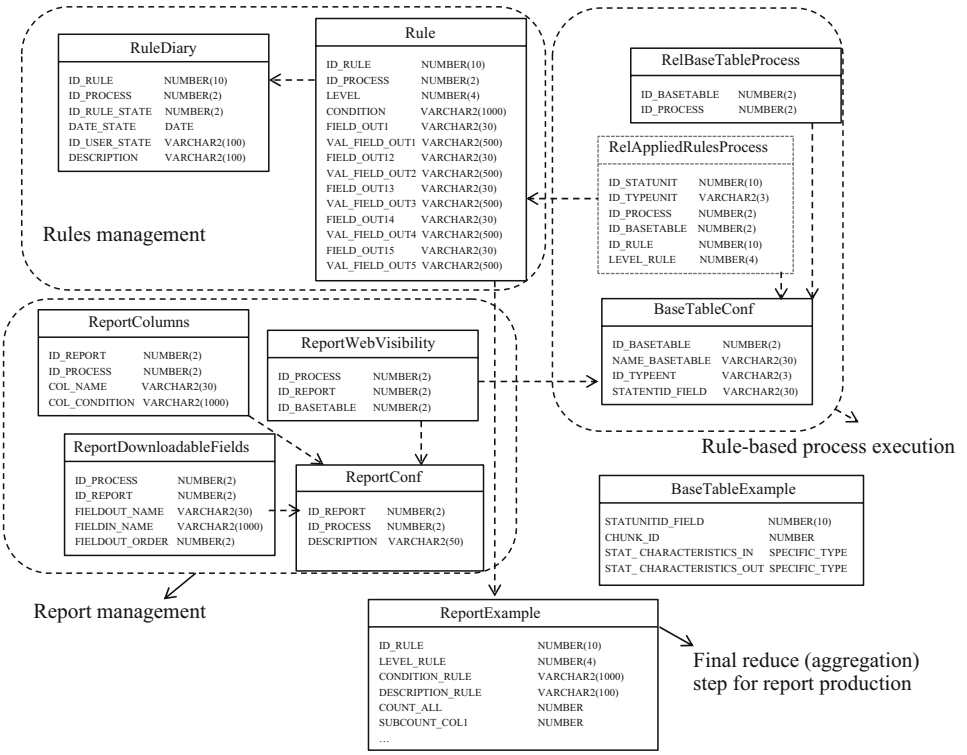

Fig. 9. Skeleton ER for the schema with less duplication.

table itself, which is specific and different for each task. The common join table may soon accommodate tens of billions of records. It should be organized to provide data proximity: specifically processing and downloading operations should access only opportunely partitioned and close data. The improvement in the selection queries time behavior positively affects overall download performance, and in addition, enhances the user experience.

3.3. The Parallel Rule-Engine

The designed integration/validation processes are *de facto* heavily parallelizable: they work on a specific population of statistical units, retrieve all the needed information and execute the set of rules of interest in relation to each unit or to disjoint grouped unit sets. Therefore, data may be divided into many fine grained similar tasks (i.e., which realize the same operations, i.e., the same rules, on disjoint sets of data) and output reports in relation to each subset of data may be aggregated into a single final report. Parallelism may refer to task parallelism as described by Subhlok et al. (1993) and data parallelism, which focuses on distributing the data across different parallel computing nodes. We explore the latter technique for developing an efficient rule engine solution. Parallelism is generally used in clustered systems, where performance preservation is granted by saving server resources and by giving equal conditions to all tasks in which a job is massively executed in parallel (Ananthanarayanan 2013; Ananthanarayanan et al. 2013). In order to solve such problems,

recent studies propose several techniques to determine the impact of parallelism on the amount of resources (Delimitrou and Kozyrakis 2014). Platform dependent algorithms for managing the parallel execution are called cache-aware algorithms (Prokop 1999). They are particularly relevant in database management server heavy load conditions, when the benefits of parallelism start to fade. The proposed rule-processing engine solution for integration and validation tasks in official statistics domains makes use of data parallelism as follows. Data are divided into consistent subsets (i.e., *base table* bands or chunks associated to similar mini (i.e., sub-) tasks); a flow of similar mini tasks is hence provided to a given number of active server processes. The rule-processing strategy is depicted in Figure 10. Through such a mechanism, only a subset of the whole set of data is managed by the database management server in a given time unit and, if the mini tasks are solved in an efficient manner, the cache stress is constrained.

Specifically, the number of parallel servers and the chunk dimension may be set optimally, thus saving database management server resource consumption. We will show system benefits in terms of resource consumption in the following section.

3.4. Selection Logic Embedded in Rules

In this section, we focus on the input data selection sub-task. When designing a service for solving multiple sources prioritization and validation tasks, an important issue is how to retrieve input data for performing these tasks. Our Statistical Service may be hosted in specialized servers, possibly scalable ones. However, input data may be spread in several intra/inter NSOs remote systems, which could be based on different technologies. Dedicated integration solutions, which load *ad-hoc* the necessary inputs in the same homogeneous enviroment, may bring to data silos and data labyrinth (Van Der Lans 2013). Moreover, data locality may require the definition of expensive processes (Goede 2011) and a careful design of data replication-based architectures. As outlined in Section 1, in order to avoid such silos, the literature proposes data virtualization for ever-changing integration needs (Krawatzeck et al. 2015). Remote source-independent selection becomes a key element for ensuring agile data integration. In the European context such solutions

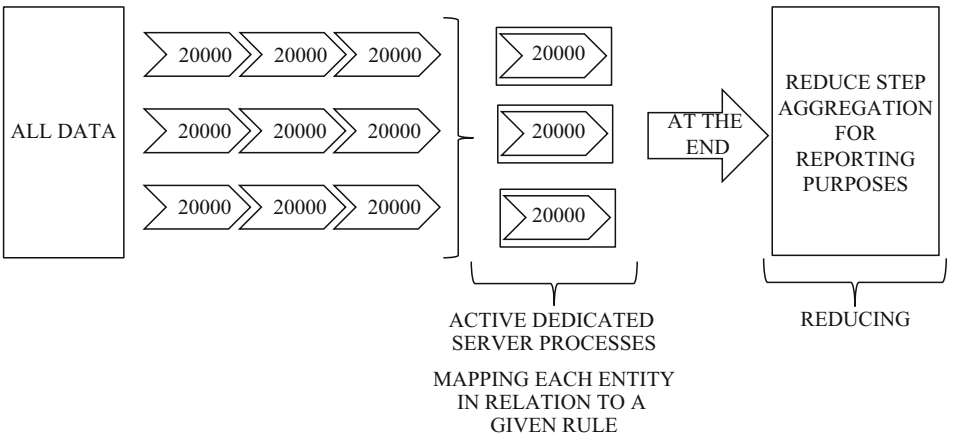

Fig. 10. Data parallelism strategy.

have also been increasingly explored for modernizing inter/intra NSO production (Gramaglia 2015). In these cases, security and confidentiality issues (Yu et al. 2010; Zissis and Lekkas 2012), in relation to "Not only Auth-entication and Authorization" (NoAuth) issues, are outside the scope of this document. However, particularly when Entity and Statistical services are exposed outside a single trusted domain (e.g., in inter-NSOs scenarios or in case web-linked external resources are used in the data production chain), they should be taken into account and analysed in-depth. Specifically, we have explored the usage of the implemented rule engine for the selection issues. Data parallelism may increase efficiency in selection, thus opening up various data architectures, where data may be stored locally with respect to the server that hosts the rule engine, as well as remotely. In fact, remote sourcing may involve multiple servers. A large amount of exchanged data stresses the multiple servers resources and may be a stumbling block in using such a solution. Remote sourcing may benefit from a more efficient data exchange. We encapsulate source data calls within the rules. Each different data source consists of a different remote call. In such a scenario the server, which hosts the rule-processing engine, must use the data source callable functions (i.e., clients). The available clients represent our remote integration set, which may be virtually integrated in a temporary *base table*. Conversely, selection rules might interface with a separate single integration layer. Different technologies enable architects to promote remote sourcing. We consider two different solutions, thus outlining robust considerations in relation to data virtualization use in the statistical context. Specifically, distributed connectivity may be provided by using database connectivity technologies in homogeneous and heterogeneous environments, possibly by relying on gateway agents and drivers. Otherwise, distributed connectivity may be provided by using web services, thus providing the highest level of interoperability in heterogeneous environments. Each different component framework, using wrapping, may be exposed through web services. In particular, we assess the following technologies: database links in an Oracle homogeneous environment, and Simple Object Access Protocol (SOAP) web services, exchanging data using eXtended Markup Language (XML) format, which is a standard for data and message exchange over the internet.

4. Highly-Performant Data Management, Efficiency in Processing and Virtualization in Selection Assessment

In this section, we assess the robustness and efficiency of the designed service by tackling the open issues arising from the previous sections. In particular, our experimental results show how engineered highly-performant data management together with fine-tuned parallelism techniques may substantially improve the flexible, inclusive usage and therefore the level of reuse of a prioritization and validation service. We also compare the aforementioned input data selection solutions by assessing local data with respect to remote data access. In the latter case, we assess the usage of database connectivity technology with respect to web service technology. Data locality is obviously the best choice in terms of processing times. However, data locality may require static loading processes. The cost of the data replication architecture should also be taken into account. Parallelism can mitigate remote-sourcing worse performance; likewise, service replication architectures, which rely on scalable services, should be assessed.

4.1. How Relevant is a Performant Data Management for Having Sharable Services?

In this section we evaluate different data schema solutions. We validate the highly-performant one in relation to the service operations. We show the benefits of data duplication reduction and of optimized data schemas in task processing. In particular, in Figure 11 we show the unitary execution times in milliseconds (i.e., the ratio between the execution time of a process and the number of statistical units involved) of a single rule-based process by using the less-duplication (i.e., partitioned with less duplication) schema and the more-duplication one, as outlined in Subsection 3.2, when the number of involved units increases. The association table between applied rules and statistical units accommodates the data of a single process. The tested process is an SBR validation process based on the evaluation of about 400 rules. The results clearly show that the data schema with less duplication (double dashed line) outperforms the other one (solid line).

Figure 11 also shows that the unitary execution time of a rule-based process remains almost constant as the number of involved units increases (in normal database load conditions). The execution times curve is linear in the number of statistical units (obviously, when the load conditions increase and physical server resources become scarce, the performance decreases and the curve changes its shape). We measured similar processing times in relation to the less-duplication schema, even in the case of eight simultaneous validation processes, operating on different *base table*s, but on similar data and rules, and resulting in about 2,000,000,000 records in the common join table. A highly-performant data management increases the flexibility in service use. In the next section, we assess a further improvement in performance by enabling the system to use scalable processing techniques.

4.2. The Parallel Engine

In this section we show the benefits of using parallelism. Opportune settings of parallel execution parameters may produce efficiency gains in terms of both execution time and

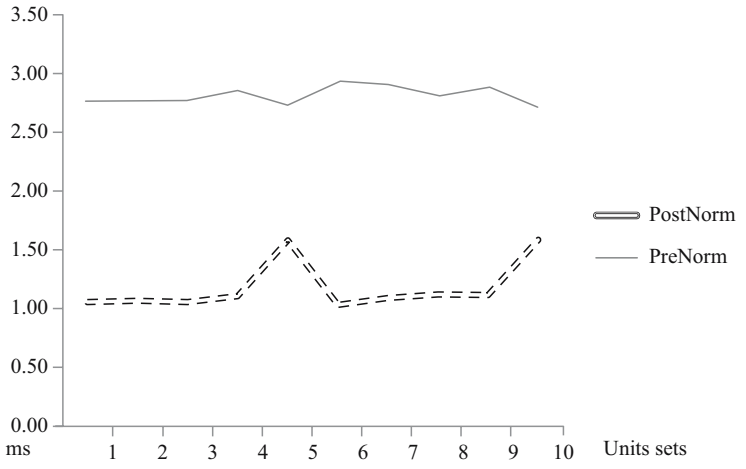

Fig. 11. Unitary execution times (in ms) (Y-axis) of ten executions (X-axis) of a rule-based process when a data schema with more duplication (solid line) and a schema with less duplication (double dashed line) are used. On X-axis the number of involved units increases from about 1,370,000 up to 13,700,000 units (ten executions).

server resource consumption. We assess typical performance measures of an Oracle database management server: the task execution time, the consistent gets, the Process Global memory Area consumption (PGA) and CPU time consumption.

Briefly, consistent gets represent the number of logical read requests to get data from the memory area shared by all the processes. PGA represents the single process dedicated memory area size and it is a dynamic, limited part of the overall shared one. In Figures 12 and 13 we show the time behavior (i.e., execution time curve) of a big validation process (around 13,000,000 statistical units and 400 validation rules) in relation to an increasing number of simultaneous active server processes when parallelism is used. In Figure 12, the dashed line shows the execution time of a chunked validation task when a single server is allocated, divided by the X-axis number of servers. The latter refers to an ideal parallel process, since it represents the ideal situation when several parallel processes solve the same task on disjoint subsets of data. The dotted line refers to the actual experienced execution time of the chunked validation task when allocating from one to five active parallel servers. The solid line shows the execution time of an equivalent non-parallel process, which may be computed by multiplying the actual execution time of the parallel validation task (i.e., the dotted line values) by the number of allocated parallel servers (i.e., X-axis numbers). Parallelism is effective in reducing the processing time of the task and close to the ideal case in relation to the single server time. Furthermore, in Figure 13a we show the scalability level of the parallel executions and their robustness in relation to several load scenarios.

In particular, in Figure 13a and 13b *Parallel_P1*, *Parallel_P2*, *Parallel_P3* represent three simultaneous "big" parallel similar validation processes, which solve the validation task in relation to the same input data, units and rules. The word "big" refers to a validation process with the same set of rules (approximately 400) and related to the same statistical population, about 13,000,000 statistical units, although working on different *base tables* and metadata. *Parallel_NoP* is a single parallel execution, processed in a separate test session when two other big, non-parallel similar (as before) validation processes are active

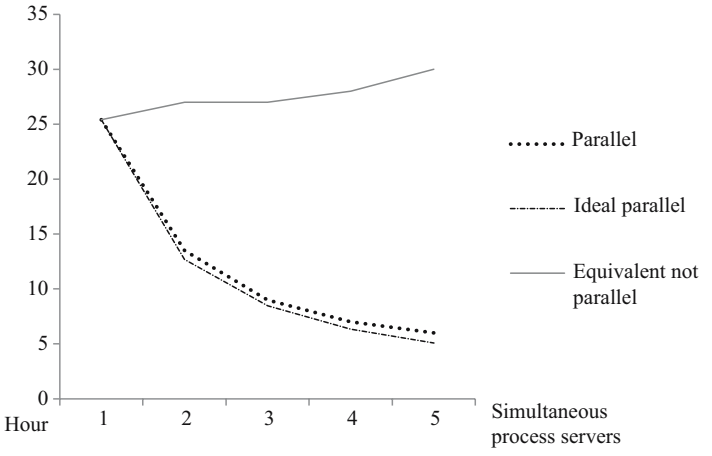

Fig. 12. Real rule-based process executed in a parallel fashion in relation to an increasing number of active allocated server processes with respect to an ideal, thus optimal, execution of a rule-based process in parallel fashion.

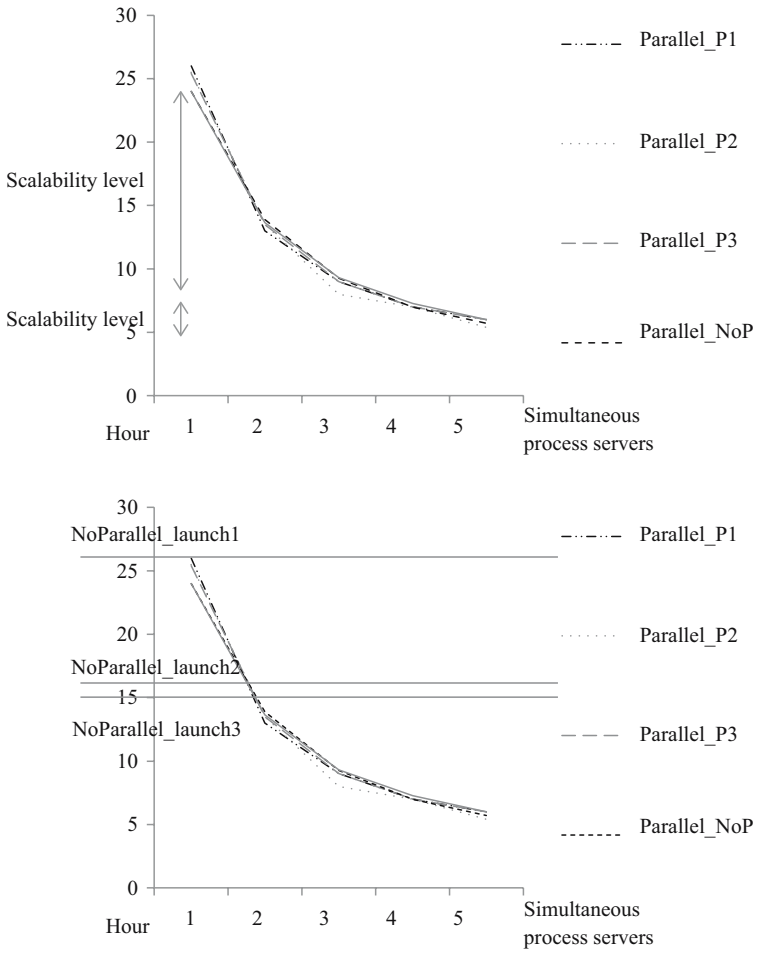

Fig. 13. (a) upper, scalability level and speed up in relation to the active allocated server processes; (b) lower, several parallel execution of the same (same data, same rules) rule-based process by increasing the number of allocated parallel servers. The straight lines aim to point the execution times of three contemporary executions of the same rule-based process, processed in a non-parallel fashion (obviously not varying in relation to the number of allocated servers).

at the same time. *NoParallel_launch1*, *NoParallel_launch2*, *NoParallel_launch3* are three simultaneous executions of three big similar validation tasks, when parallelism and chunking of data are not used. The scalability level substantially decreases when the numbers of allocated simultaneous servers grows. The execution time gain is therefore smaller and smaller in relation to an increasing number of active parallel servers. The minimum number of active servers in relation to a target speedup and to the server resources preservation is a rewarding processing choice. Each active server process manages a mini task in a given time unit: the fewer active mini task/processes we allocate, the lesser resources we simultaneously consume. Moreover, parallel executions performance does not vary in relation to different load scenarios. The performance of the parallel processes is quite stable, even when other "big" processes (parallel or non-parallel) are executed. On the other hand, one of the three simultaneous non-parallel big

Table 1. Local data: mean performance indices in relation to six non-parallel executions of a selection process on 28 different sources (i.e., the process is composed by 28 different selection rules) and related to an increasing set of involved statistical units (from 10,000 to 60,000) and in relation to the corresponding parallel executions (with four and six servers).

	LOCAL DATA			
	Mean single chunk (A)	Ideal single chunk	Mean not-parallel execution (B)	B/A
Consistent gets	509058	346726	3120532	6.13
CPU	631	4883	15675	24.84
PGA	2696481	494110	4446995	1.65

processes experienced decrease in performance. In Figure 13b, when just one active server is allocated for parallel execution, by letting it manage the overall flow of chunked validation mini-tasks, performance is worse than the single non-parallel execution. The parallel processing performance is driven by the single unit overhead, introduced by chunking. Each single chunk of data introduces an overhead in processing due to a not ideal consumption of resources. Therefore, in Table 1 we show the server resource consumption in terms of consistent gets, CPU time consumption in centiseconds and PGA consumption in bytes in relation to a single processing task in case of non-parallel execution and to a single processing mini task in case parallelism is used.

Specifically, we evaluate performance indices for a selection task (i.e., processing of the selection rules for retrieving input data for a specific integration and validation task), which is related to 10,000, 20,000, 30,000, 40,000, 50,000 and 60,000 statistical units, and which has been processed in a parallel (both four and six contemporary active servers and mean chunk dimension is about 4,000 units) and non-parallel fashion. In Table 1, the mean values are obtained by averaging both the non-parallel and the parallel executions. B/A outlines, in a given time unit, the over-consumption in relation to a specific index of performance of a mean single/unique non-parallel execution with respect to a mean single parallel mini task execution. It shows a possible degree of parallelism (i.e., number of simultaneous active mini tasks) we may choose for consuming in a given time unit less resources in relation to the single/unique non-parallel execution of the same overall process. Consistent gets and PGA refer to the server memory cache. PGA seems to be the more critical parameter. Benefits in time execution might require an over-consumption of single process server memory. The scalability issue becomes a relevant topic.

In Figure 14, we show the PGA used and consistent gets for one single statistical unit when processing selection rules in parallel (black line in the figure) and in non-parallel (grey line in the figure) fashion in relation to an increasing number of involved units, as before. The overhead of the parallel process in unitary terms determines the performance decrease between the non-parallel single/unique execution and the chunked parallel exeuction when only one active server is allocated for managing the overall flow of chunked mini tasks, as shown in Figure 13b. Therefore, optimization in data schema design, described in Subsection 3.2, is even more relevant in relation to the unitary over-consumption of server resources, thus increasing the benefits of data parallelism. The consistent gets overhead appears linear in the selection dimension, while PGA

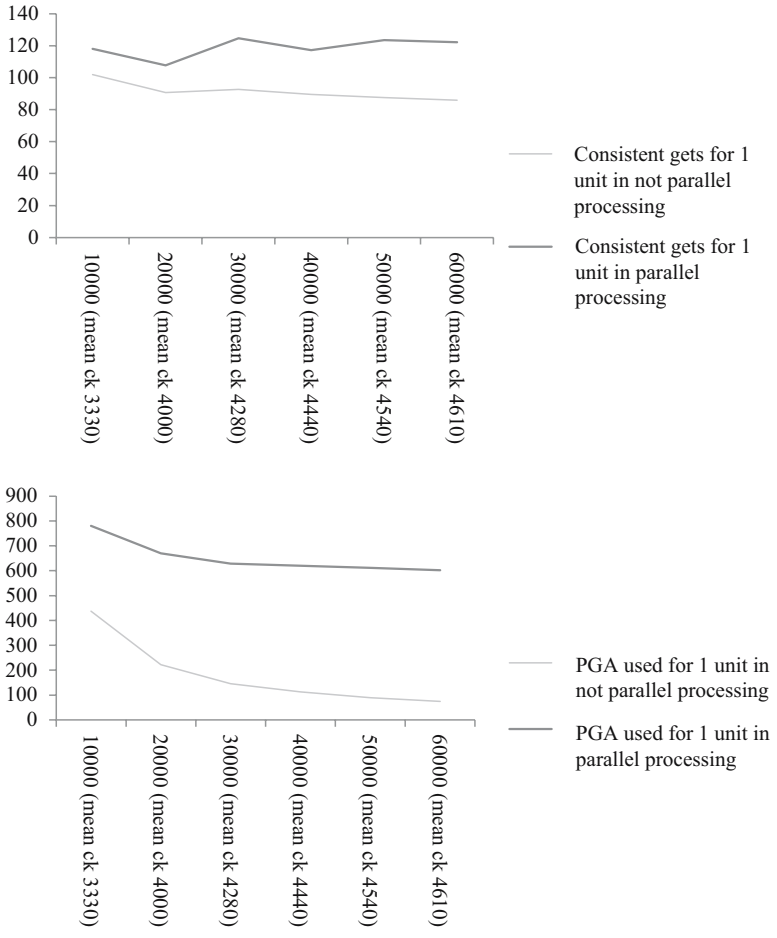

Fig. 14. Unitary PGA used in bytes and unitary consistent gets in number of data block reads for a selection process which is performed in parallel and in non-parallel fashion.

consumption does not seem to exhibit the same trend. Linear trends produce a fixed single server performance decrease, which guides the performance of the overall parallel processing, whatever chunk dimension we choice.

4.3. Data Loading for Prioritization and Validation Tasks: Where Can We Place Input Data?

Nowadays, data-driven architectures have been increasingly imposed to migrate old IT systems or build new ones in a SOA perspective. We explore the use of the developed efficient rule engine for input selection purposes by encapsulating within selection rules the input data calls. In this section, therefore, we assess the data architectures presented in Subsection 3.4. Specifically, we compare local data (with respect to the server that hosts the rule-processing engine component) and remote data sourcing in terms of selection time and server resource consumption by using the same performance measures presented in the previous section. In the case of remote sourcing, we focus on a distributed database

scenario, by comparing database connectivity technology for exchanging data (i.e., referred to as JDBC/SQL below), and a web-serviced data scenario (i.e., referred to as SOAP/XML below), by using web services for exchanging data in XML format. The latter solution enables architects to use interoperable Entity Services for sourcing any Statistical Services. In Figure 15, we compare selection times in six different scenarios and in relation to an increasing number of statistical units. On the X-axis we have different sets of involved units from 10,000 up to 60,000. In Figure 15, the *dblink* curve refers to a non-parallel execution of a selection task, which is composed of 28 selection rules, when a JDBC/SQL selection is adopted. The *dblink p4* one refers to the execution of the selection process in a parallel fashion, by using four active simultaneous process servers, when database connectivity technology is used.

The *dblink p6* curve refers to a processing scenario with six simultaneous active servers. The curve labelled with *xml* represents the non-parallel processing of the same selection tasks when data are exchanged through XML. The *xml p4* curve refers to the parallel processing of the selection tasks with four active servers and finally the *xml p6* curve refers to the parallel processing of the selection tasks with six active simultaneous servers. The curve that is labelled with *local* refers to the same tasks as before when source data are stored locally with respect to the server hosting the rule-processing engine and parallelism and data chunking are not used. Local data obviously exhibits the best performance, but the static loading issue remains. XML data exchange time is affected by the XML serialization and deserialization processes and by the selection of virtually loaded data, before importing them into the *base table*. Therefore, it exhibits the worst performance in relation to the rule-based selection tasks, although parallel execution may partially mitigate the performance decrease. The knowledge of the above curve may help architects to choose the most suitable selection scenario, possibly adopting the most interoperable data exchange both in intra-NSO and in inter-NSOs context, when the overhead is acceptable or may be suitably managed. We assess server resources consumption in

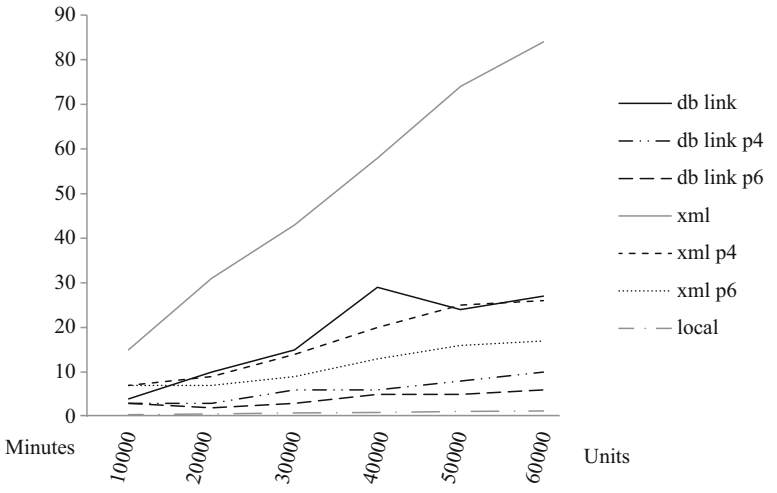

Fig. 15. Selection times in case of database selection (three scenarios: non-parallel, parallel with four servers and parallel with six servers) and in case of xml selection (three scenarios: non-parallel, parallel with four servers and parallel with six servers).

Table 2. Specifically, CPU time, PGA consumption and shared cache accesses (i.e., consistent gets) are presented.

The benefits of XML data exchanges are counterbalanced by higher resource consumption in terms of cache stress and CPU time. The SOAP/XML overhead in resource consumption is always greater than the JDBC/SQL one. Agile database deployment and service scalability, which is further recommended in a micro-service architecture, might surely be relevant drivers for enabling architects to choose data virtualization solutions based on XML data exchange. Entity Services and Statistical Services should be easily scalable. Container-based architectures (Xavier et al. 2013), which moreover ensure the technology neutrality property in service development, seem promising in providing agility in database deployment. They sustain scalable service architectures and should be explored, thus supporting a future-proof manner Statistical Service sharing. When remote sourcing is used, the overhead in resource consumption, for example, in terms of memory, is higher than when data locality is granted. However, in SOAP/XML scenario the degree of parallelism we may set, consuming less resources with respect to the corresponding execution when data parallelism and data chunking are not used, is higher than in JDBC/SQL one. Specifically, due to data parallelism, the speed-up gain may be substantial in the case of XML data exchange, as Figure 15 shows, thus enabling architects to choose such selection scenario in a managed way. Data virtualization benefits may be achieved when a performant data exchange is ensured. Data parallelism and selection rules sustain selection performance. Scalable service architectures may manage any over-consumption of involved servers resources.

Table 2. Mean performance index values with regard to the parallel and non-parallel executions, whose processing times have been shown in Figure 14.

JDBC/SQL DATA			
	Mean single chunk (A)	Mean not-parallel execution (B)	B/A
Consistent gets	510323	287752	0.56
CPU	2920	46259	15.84
PGA	2961426	5222504	1.76
SOAP/XML DATA			
	Mean single chunk (A)	Mean not-parallel execution (B)	B/A
Consistent gets	904636	5158917	5.70
CPU	24067	699154	20.52
PGA	18359343	27679507	1.51
OVERHEAD RESOURCES BETWEEN SOAP/XML AND JDBC/SQL %			
	Mean single chunk (A)	Mean not-parallel execution (B)	
Consistent gets	43	94	-
CPU	91	93	-
PGA	83	81	-

5. Conclusions and Future Work

5.1. Summary of the Proposed Solution

In this article we describe a multiple source rule-based prioritization and validation service, successfully developed in the Italian SBR context. Rule-based solutions provide decoupling between the domain experts and the IT experts, sustain agile rules evolution and simplify sharing. We assess a micro-service solution, which may be easily inserted into a production chain and is an affordable migration path towards a SOA. It also facilitates the agile cooperation between domain expert users and IT ones. We specifically promote optimized data management and efficient data processing, using data parallelism techniques, in deterministic rule-based tasks. We further assess selection rules, which encapsulate the input-source calls to Entity Services. The latter might expose single as well as integrated datasets, local as well as remote data, thus enabling architects to use data virtualization solutions in the input selection task. Data locality obviously shows best performance, although the static data loading problem remains, and data replication architectures and consistency issues should be carefully taken into account. Remote sourcing, when needed, requires attention in physical server dimensioning and scalability issues.

5.2. Highlighted Key Principles

Modernization impetuses move official statistics towards reuse and sharing of methods and components. Specifically, it moves official statistics towards SOA, in order to react promptly to ever-evolving scenarios and towards heterogeneous data integration, in order to increase the level of quality in relation to some quality components and, possibly, the number of the statistical outputs. Such impetuses should be taken into account when deciding IT solutions for a GSBPM activity. We aim to promote an inclusive application design pattern which enables reuse and sharing in a modern way. Specifically, we promote the following, CSPA compliant, principles. The "single capability principle", which is a functional foundation of micro-service architecture, ensures the minimization of costs for new or changed requirements. The "technological neutrality principle", which does not impose a specific development, integration or deployment platform. The micro-service architecture does not rely on a given technological platform, but rather on stateless, autonomous and self-contained data schema, web-user interfaces, and processing components, which may be independently deployed. We further highlight the importance of non-functional requirements. In particular, performance assessment and efficiency evaluation are relevant drivers for an inclusive reuse of the service. Data virtualization is another relevant driver which increases agility. It enables a SOA where Statistical Services may call remote Entity Services to consume input, eventually integrated, data.

5.3. Future Work

Care should be taken in implementing the micro-service architectural pattern: it introduces some extra administrative overhead, in particular for deployment, administration, monitoring and security. When data virtualization is used in a single trusted domain,

Auth-entication and Auth-orization (Auth) policies and techniques are well defined and easily taken into account in service usage. However, when interfacing various inter-organization domains, federated Auth policies should be engineered and NoAuth issues, such as the confidentiality one, should be taken into account as strict nonfunctional requirements which simplify service sharing. Security issues could be stumbling blocks in sharing. Even technological issues may represent a stumbling block in service sharing. We promote technological neutrality in development. Containerization packages single services and complex applications in autonomous containers, which exhibit great isolation capabilities and are portable across different technological environments. While imposing the platform-as-a-service paradigm, nowadays it seems to preserve the tech neutrality property and to sustain a stepwise migration pattern towards a Statistical SOA. Future work should therefore explore the latter solution due to its promise of simplifying platform configuration, and offering lightweight runtimes, which sustain orchestration, scheduling, scalability and security issues at container level. Likewise, work which further promotes system efficiency could, as proven, have further positive impacts in relation to service reuse and sharing, and further work on data virtualization, by assessing selection rules which interface with a single integration layer, which manages different data sources, may increase the service agility (Karpathiotakis et al. 2015).

Briefly, from the experience in SBR domain, we may highlight that agility and efficiency grant service reuse and sharing in relation to multiple source prioritization and validation tasks. A systematic performance assessment in relation to resource utilization, time behavior and capacity may evaluate efficiency in processing and communication; while rule-based solutions, micro-service patterns, data virtualization and cooperation in development may provide service agility. Future exploration of container-based architectures seems promising in granting other non-functional requirements, such as scalability, security, maintainability and service portability in a technological neutral perspective.

6. References

Alagiannis, I., R. Borovica, M. Branco, S. Idreos, and A. Ailamaki. 2012. "NoDB: Efficient Query Execution on Raw Data Files." In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data: 241–252. Scottsdale, Arizona, U.S.A. May 20–24, 2012. Doi: http://dx.doi.org/10.1145/2213836.2213864.

Ananthanarayanan, G. 2013. *Optimizing Parallel Job Performance in Data-Intensive Clusters*. Diss. University of California, Technical report Berkeley EECS. Available at: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2014/ (accessed November 2017).

Ananthanarayanan, G., A. Ghodsi, S. Shenker, and I. Stoica. 2013. "Effective Straggler Mitigation: Attack of the Clones." *In NSDI* 13: 185–198. ISBN: 978-1-931971-00-3. Available at https://www.usenix.org/conference/nsdi13/technical-sessions/presentation/ananthanarayanan (accessed November 2017).

Chen, Y. and D.Z. Wang. 2014. "Knowledge Expansion Over Probabilistic Knowledge Bases." In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data: 649–660. Snowbird, Utah, U.S.A. June 22–27, 2014. Doi: http://dx.doi.org/10.1145/2588555.2610516.

Chen, T., R. Bahsoon, and A.R.H. Tawil. 2014. "Scalable Service-Oriented Replication with Flexible Consistency Guarantee in the Cloud." *Information Sciences* 264: 349–370. Doi: http://dx.doi.org/10.1016/j.ins.2013.11.024.

Cheng, Yu, and F. Rusu. 2015. "Scanraw: A Database Meta-Operator for Parallel In-Situ Processing and Loading." *ACM Transactions on Database Systems (TODS)* 40(3): 19. Doi: http://dx.doi.org/10.1145/2818181.

Delimitrou, C. and C. Kozyrakis. 2014. "Quasar: Resource-Efficient and Qos-Aware Cluster Management." In *ACM SIGPLAN Notices* 49(4): 127–144. ACM. Doi: http://dx.doi.org/10.1145/2644865.2541941.

De Sa, C., A. Ratner, C. Ré, J. Shin, F. Wang, S. Wu, and C. Zhang. 2016. "DeepDive: Declarative Knowledge Base Construction." *ACM SIGMOD Record* 45(1): 60–67. Doi: http://dx.doi.org/10.1145/3060586.

Di Zio, M., N. Fursova, T. Gelsema, S. Giessing, U. Guarnera, J. Petrauskienė, L. Quenselvon Kalben, M. Scanu, K.O. ten Bosch, M. van der Loo, and K. Walsdorfer. 2016. "Methodology for Data Validation." ESSNET ValiDat Foundation. Available at https://ec.europa.eu/eurostat/cros/system/files/methodology_for_data_validation_v1.0_rev-2016-06_final.pdf (accessed November 2017).

Dragoni, N., M. Mazzara, S. Giallorenzo, F. Montesi, A. Lluch Lafuente, R. Mustafin, and L. Safina. 2017. "Microservices: Yesterday, Today, and Tomorrow. In Present and Ulterior Software Engineering." Springer Berlin Heidelberg. Doi: http://dx.doi.org/10.1007/978-3-319-67425-4_12.

ESSnet. 2015. "Enterprise Architecture Reference Framework." Available at https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework_en (accessed November 2017).

ESSnet Core Project. 2011. "Common Reference Environment." Available at https://ec.europa.eu/eurostat/cros/content/core_en (accessed November 2017).

ESSnet ValiDat Integration. 2017. "Harmonising Data Validation Approaches in the ESS." Available at https://ec.europa.eu/eurostat/cros/content/essnet-validat-integration_en (accessed November 2017).

Fowler, M. 2014. "A definition of this new architectural term". Available at http://martinfowler.com/articles/microservices.html (accessed November 2017).

Goede, R. 2011. "Agile Data Warehousing: The Suitability of Scrum as Development Methodology." In Proceedings of the 5th IADIS Multi Conference on Computer Science and Information Systems (MCCSIS'2011): 51–58. Rome, Italy. 20–26 July 2011. Available at http://ims.mii.lt/ims/konferenciju_medziaga/MCCSIS/I_WAC_TNS_2011.pdf#page=72 (accessed November 2017).

Gramaglia, L. 2015. "Towards a European Validation Architecture." ESSNET ValiDat Foundation. Available at https://ec.europa.eu/eurostat/cros/content/workshop_en (accessed November 2017).

GSBPMv5.0. 2017. The Generic Statistical Business Process Model. Available at https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0 (accessed November 2017).

GSDEM. 2015. *The Generic Statistical Data Editing Models*. Available at https://statswiki.unece.org/display/sde/GSDEMs (accessed November 2017).

Idreos, S., I. Alagiannis, R. Johnson, and A. Ailamaki. 2011. "Here are my data files. here are my queries. where are my results?" In Proceedings of 5th Biennial Conference on

Innovative Data Systems Research (No. EPFL-CONF-161489). Asilomar, California, U.S.A., January 9–12, 2011. Available at http://cidrdb.org/cidr2011/Papers/CIDR11_Paper7.pdf (accessed November 2017).

Karpathiotakis, M., I. Alagiannis, T. Heinis, M. Branco, and A. Ailamaki. 2015. "Just-In-Time Data Virtualization: Lightweight Data Management with ViDa." In Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR) (No. EPFL-CONF-203677). Asilomar, California, U.S.A., January 4–7, 2015. Available at https://infoscience.epfl.ch/record/203677/files/vida-cidr.pdf (accessed November 2017).

Karpathiotakis, M., A. Ioannis, and A. Anastasia. 2016. "Fast Queries Over Heterogeneous Data Through Engine Customization." *Proceedings of the VLDB Endowment* 9(12): 972–983. Doi: http://dx.doi.org/10.14778/2994509.2994516.

Khadka, R., A. Saedi, A. Idu, J. Hage, and S. Jansen. 2012. "Legacy to SOA evolution: A systematic literature review. Migrating Legacy Applications": *Challenges in Service Oriented Architecture and Cloud Computing Environments*: 40. Doi: http://dx.doi.org/10.4018/978-1-4666-2488-7.ch003.

Krawatzeck, R., B. Dinter, and D.A. Pham Thi. 2015. "How to make business intelligence agile: The Agile BI actions catalog." In System Sciences (HICSS), 2015 48th Hawaii International Conference on: 4762–4771. 5–8 January 2015. Hawaii, U.S.A. IEEE. Doi: http://dx.doi.org/10.1109/HICSS.2015.566.

Lavrac, N. 2001. "Data Mining and Decision Support: A note on the issues of their integration and their relation to Expert Systems." In the workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning IDDM. Available at http://kt.ijs.si/Branax/IDDM-2001_submissions/Lavrac.pdf (accessed November 2017).

Liang, S., P. Fodor, H. Wan, and M. Kifer. 2009. "OpenRuleBench: An analysis of the performance of rule engines." In Proceedings of the 18th international conference on World Wide Web: 601–610. Madrid, Spain. April 20–24, 2009. ACM. Doi: http://dx.doi.org/10.1145/1526709.1526790.

Milani, B.A. and N.J. Navimipour. 2016. "A Comprehensive Review of the Data Replication Techniques in the Cloud Environments: Major Trends and Future Directions." *Journal of Network and Computer Applications* 64: 229–238. Doi: http://dx.doi.org/10.1016/j.jnca.2016.02.005.

Mohamed, M.F. 2016. "Service Replication Taxonomy in Distributed Environments." *Service Oriented Computing and Applications* 10(3): 317–336. Doi: 10.1007/s11761-015-0189-7.

Montoya, G., H. Skaf-Mollia, P. Molli, and M.-E. Vidal. 2017. "Decomposing Federated Queries in Presence of Replicated Fragments." *Web Semantics: Science, Services and Agents on the World Wide Web* 42: 1–18. Doi: http://dx.doi.org/10.1016/j.websem.2016.12.001.

Namiot, D. and M. Sneps-Sneppe. 2014. "On Micro-Services Architecture." *International Journal of Open Information Technologies* 2(9): 24–27. Available at http://injoit.org/index.php/j1/article/view/139 (accessed November 2017).

O'Brien, L., P. Brebner, and J. Gray. 2008. "Business transformation to SOA: aspects of the migration and performance and QoS issues." In Proceedings of the 2nd international

workshop on Systems development in SOA environments: 35–40. Leipzig, Germany. May 10–18, 2008. ACM. Doi: http://dx.doi.org/10.1145/1370916.1370925.

Osrael, J., L. Froihofer, and K.M. Goeschka. 2006. "What Service Replication Middleware Can Learn from Object Replication Middleware." In Proceedings of the 1st workshop on Middleware for Service Oriented Computing (MW4SOC 2006): 18–23. Melbourne, Australia. November 27 – December 01, 2006. ACM. Doi: http://dx.doi.org/10.1145/1169091.1169094.

Prokop, H. 1999. *Cache-oblivious algorithms*. Doctoral dissertation, Massachusetts Institute of Technology. Available at http://supertech.csail.mit.edu/papers/Prokop99.pdf (accessed November 2017).

Pullokkaran, L.J. 2013. *Analysis of Data Virtualization and Enterprise Data Standardization in Business Intelligence*. Doctoral dissertation, Massachusetts Institute of Technology. Available at http://hdl.handle.net/1721.1/90703 (accessed November 2017).

Quensel-von Kalben, L. 2017a. "SERV – Adopting Common Statistical Production Architecture (CSPA) in Europe." *NTTS 2017*. Doi: http://dx.doi.org/10.2901/EUROSTAT.C2017.001.

Quensel-von Kalben, L. 2017b. "Validation, shared services and enterprise architecture: how it fits." *UNECE SDE*. Available at https://www.unece.org/index.php?id=43887 (accessed November 2017).

Razavian, M. and P. Lago. 2015. "A Systematic Literature Review on SOA Migration." *Journal of Software: Evolution and Process* 27(5): 337–372. Doi: http://dx.doi.org/10.1002/smr.1712.

Scannapieco, M., L. Tosco, C. Vaccari, and A. Virgillito. 2011. "A Common Reference Architecture for National Statistical Institutes: the CORA Project." *NTTS 2011*. Doi: http://dx.doi.org/10.2901/Eurostat.C2011.001.

Schafer, M. 2015. A study on VTL. *A Study on the Validation and Transformation Language*. Available at https://ec.europa.eu/eurostat/cros/content/essnet-validation-study-vtl-final_en (accessed November 2017).

Stodder, D. 2013. *Achieving Greater Agility with Business Intelligence*. TDWI Best Practices Report, First Quarter. Available at http://info.attivio.com/rs/attivio/images/TDWI-and-Attivio-Best-Practices-Report-Achieving-Greater-Agility-with-Business-Intelligence-Q1-2013.pdf (accessed November 2017).

Subhlok, J., J.M. Stichnoth, D.R. O'Hallaron, and T. Gross. 1993. "Exploiting Task and Data Parallelism on a Multicomputer." In *ACM SIGPLAN Notices* 28(7): 13–22. ACM. Doi: http://dx.doi.org/10.1145/173284.155334.

Tian, Y., I. Alagiannis, E. Liarou, A. Ailamaki, P. Michiardi, and M. Vukolić. 2017. "DiNoDB: an Interactive-speed Query Engine for Ad-hoc Queries on Temporary Data." *IEEE Transactions on Big Data*. Doi: http://dx.doi.org/10.1109/TBDATA.2016.2637356.

Van Der Lans, R.F. 2013. *Creating an Agile Data Integration Platform using Data Virtualization*. R20 consultancy technical whitepaper. Available at http://stonebond.com/wp-content/uploads/2014/02/Rick-Van-Der-Lans-Whitepaper-May-2013.pdf (accessed November 2017).

Xavier, M.G., M. Neves, F. Rossi, T. Ferreto, T. Lange, and C. de Rose. 2013. "Performance evaluation of container-based virtualization for high performance

computing environments. Parallel, Distributed and Network-Based Processing (PDP)." 2013 21st Euromicro International Conference on. IEEE. Belfast, United Kingdom, 27 Februari–1 March 2013. Doi: http://dx.doi.org/10.1109/PDP.2013.41.

Xie, G., G. Zeng, Y. Chen, Y. Bai, Z. Zhou, R. Li, and K. Li. 2017. "Minimizing Redundancy to Satisfy Reliability Requirement for a Parallel Application on Heterogeneous Service-oriented Systems." *IEEE Transactions on Services Computing*. Doi: http://dx.doi.org/10.1109/TSC.2017.2665552.

Yu, S., C. Wang, K. Ren, and W. Lou. 2010. "Achieving Secure, Scalable, and Fine-Grained Data Access Control in Cloud Computing." *In Infocom, 2010 proceedings IEEE*: 1–9. Ieee. Doi: http://dx.doi.org/10.1109/INFCOM.2010.5462174.

Zhou, X., Y. Chen, and D.Z. Wang. 2016. "ArchimedesOne: Query Processing Over Probabilistic Knowledge Bases." *Proceedings of the VLDB Endowment* 9(13): 1461–1464. Doi: http://dx.doi.org/10.14778/3007263.3007284.

Zissis, D. and D. Lekkas. 2012. "Addressing Cloud Computing Security Issues." *Future Generation Computer Systems* 28(3): 583–592. Doi: http://dx.doi.org/10.1016/j.future.2010.12.006.

Received June 2017
Revised April 2018
Accepted July 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 863–888, http://dx.doi.org/10.2478/JOS-2018-0043

Detecting Reporting Errors in Data from Decentralised Autonomous Administrations with an Application to Hospital Data

Arnout van Delden[1], *Jan van der Laan*[1], *and Annemarie Prins*[2]

Administrative data sources are increasingly used by National Statistical Institutes to compile statistics. These sources may be based on decentralised autonomous administrations, for instance municipalities that deliver data on their inhabitants. One issue that may arise when using these decentralised administrative data is that categorical variables are underreported by some of the data suppliers, for instance to avoid administrative burden. Under certain conditions overreporting may also occur.

When statistical output on changes is estimated from decentralised administrative data, the question may arise whether those changes are affected by shifts in reporting frequencies. For instance, in a case study on hospital data, the values from certain data suppliers may have been affected by changes in reporting frequencies. We present an automatic procedure to detect suspicious data suppliers in decentralised administrative data in which shifts in reporting behaviour are likely to have affected the estimated output. The procedure is based on a predictive mean matching approach, where part of the original data values are replaced by imputed values obtained from a selected reference group. The method is successfully applied to a case study with administrative hospital data.

Key words: Administrative data; measurement errors; predictive mean matching; reporting errors; selective editing.

1. Introduction

Use of administrative data in official statistics offers several advantages over survey data, such as observations for a larger fraction of the target population, reduced data collection costs and lower response burden. Therefore, administrative data is increasingly used by National Statistical Institutes (NSIs) to compile statistics, either as a sole data source or in combination with other sources. Administrative data here refers to data collected by an organisation external to the statistical office for administrative purposes, thus not targeted for use in official statistics (UNECE 2011). When the statistical population, unit and variable definitions coincide with those for the administrative data source, estimation of the statistical output is straightforward. For instance, the total number of persons receiving

[1] Statistics Netherlands, Department of Process Development and Methodology, Henri Faasdreef 312, P.O. Box 24500, 2490 HA The Hague, The Netherlands. Emails: a.vandelden@cbs.nl, and dj.vanderlaan@cbs.nl
[2] Netherlands Institute for Health Services Research (Nivel), P.O.Box 1568, 3500 BN Utrecht, The Netherlands. a.prins@nivel.nl

Acknowledgments: The authors would like to thank the Associate Editor, four anonymous referees, Peter van der Heijden and Peter-Paul de Wolf for their useful comments and suggestions, which have led to a significant improvement of this article. The authors thank Erik van Bracht for drawing the flow chart.

an unemployment benefit is easily derived from the corresponding administrative data. However, when population, unit or variable definitions do not coincide, or when the purpose of the register holder clearly differs from the intended statistical use, methodological issues may arise (Bakker and Daas 2012; Wallgren and Wallgren 2014).

One of the issues that might occur with administrative data is that the registered values differ from the true ones (as defined by the statistical office), resulting in measurement errors. This happens especially with variables that are not of crucial importance to the owner of the data set. For instance, enterprises might register their reported value added tax data as being monthly, whereas in fact it concerns four-week values (Van Delden and Scholtus 2017). The tax office tolerates deviations in monthly values, especially for smaller enterprises, as long as the yearly amount of tax paid is correct. Also, Statistics Netherlands (CBS) uses administrative fire brigade data. Fire brigades need to register the variable "did the fire cause any environmental damage". They underreport any occurrence of environmental damage, because this way they avoid having to register a number of subsequent variables, such as an estimated cost of the environmental damage (Berenschot 2012). From research on questionnaires, it is also known that respondents learn to shorten questionnaire duration by underreporting events (Backor et al. 2007; Shields and To, 2005; Silberstein and Jacobs, 1989). In the case of surveys, there is a lot of literature available on reporting errors, for instance reporting errors might occur when asking sensitive questions, or as a result of socially desirable behaviour (Tourangeau et al. 2010; Tourangeau and Yan 2007). In the case of administrative data, numerous studies on measurement errors have been done (e.g., Groen 2012; Oberski et al. 2017 and references therein), but to the best of our knowledge, the role of the administrative practice of data suppliers on these measurement errors has hardly been given any attention.

Some of the administrative data sets used in official statistics are obtained through decentralised data collection. For example, population data and social benefit data are registered by municipalities. Similar examples concern administrative data sets provided by fire brigades (on fires), by schools (on pupils), by hospitals (on patients), by local authorities (on building activities), by employers (on salary information of employees) and by courts (on legal proceedings). These decentralised administrations will be referred to as "data suppliers" in this article and the corresponding administrative data will be referred to as "decentralised administrative data". Each of these decentralised administrations may have their own administrative practices (Brackstone 1987), resulting in measurement errors that vary with the data supplier. For instance, employers in the Netherlands vary in the intensity of reporting employees' overtime. In surveys, a similar phenomenon occurs with personal interviewing, where interviewer-dependent measurement errors may occur (West and Blom 2017). For instance, homeless respondents reported drug use more frequently in the presence of male interviewers (see West and Blom 2017, 189 and references therein).

From an official statistics point of view, preventing measurement errors in administrative data is desirable, for instance by unifying and improving the "fields" that the administrators have to fill in, or the questions that they have to respond to. Nonetheless, there are at least two obstacles to achieving such improvements in practice. The first obstacle is that local administrations may have different administrative systems (software). This is, for instance, the case with employers reporting salary data for

employees, with hospital data (see Section 2) and with financial administrative data of municipalities. A second obstacle is that local administrations are autonomous and act rather independently of the statistical offices that receive the data. The best that the NSIs can do is to discuss quality issues with them and request improvements; the NSI cannot prescribe any changes to the administrative systems. Before such a discussion can be held, the NSI should have serious indications that measurement errors occur. The present article therefore focusses on the detection of reporting errors in data of decentralised, autonomous administrations.

In order to avoid biased outcomes, NSIs usually correct influential measurement errors in a data editing process. Automatic error correction methods are applied when correct values can be deduced from other variables, or when records are not influential (De Waal et al. 2011). Otherwise, selective manual data editing will be used, and, if needed, respondents are contacted. Selective editing methods aim to identify units with a high risk of influential errors, where an "influential error" is defined as one "that has a considerable effect on the publication figures" (De Waal et al. 2011). Our approach resembles that of selective editing. However, when under- or overreporting of one or more variables occurs in decentralised administrative data, correcting those data may not always be easy. Applying the methodology described in the current article, we are able to detect the data suppliers with measurement errors, but we cannot precisely detect which of the units within a data supplier contain errors.

The data suppliers responsible for those decentralised administrative data will not be able to determine which of the values are incorrect, nor to provide the "correct" values for individual records in the data. The problem is that the correct data either have not been registered, or can only be obtained with considerable effort. Municipalities, for instance, might not be able to identify which students have moved out of their parents' homes and which have not. Hospitals may not be able to see the complete set of diseases of their patients, if not all of them have been registered. In such a situation, the best option is to analyse which of the data suppliers have relatively many measurement errors. Subsequently, one can contact "suspicious" data suppliers and motivate them to improve their administrative processes in order to reduce the number of errors in future data deliveries.

Detecting under- or overreporting in a large number of variables in decentralised administrative data may be especially difficult in the case of level estimates. An option to analyse reporting behaviour in the case of level estimates makes use of a second independent source. In the present article, we aim to detect changes in reporting behaviour between two time periods. More specifically, we aim to develop an automatic procedure to detect suspicious data suppliers in decentralised administrative data in which shifts in reporting behaviour are likely to have affected a targeted change estimate. We apply our method to hospital data.

The remainder of the article is organised as follows. Section 2 gives some background information on the hospital data and the potential reporting errors therein. Section 3 describes the methodology used to select data suppliers with deviating reporting behaviour. How this methodology is applied to the case study is described in Section 4. Section 5 presents the results of the case study. Finally, Section 6 discusses the outcomes.

2. Background of the Case Study

At CBS, Dutch Medical Registration data (LBZ) is used concerning hospital stays of patients. This data set contains patient-related information, such as age and sex, and diagnosis-related variables, such as main diagnosis and comorbidities, which are other diagnoses describing the medical condition of the patient (Elixhauser et al. 1998). These data are compiled by the hospitals to provide a clinical data set that can be used by medical researchers. At each hospital, LBZ data is registered by coders using the administrative data system of the hospital and patient files (Van den Bosch et al. 2010). The LBZ data set is not targeted for use in official statistics, and fulfils the UNECE (2011) description of administrative data. We therefore refer to LBZ data as administrative data in the remainder of this article.

Since 2011, CBS has been responsible for computing the yearly Hospital Standardised Mortality Ratio (HSMR) for Dutch hospitals using LBZ data. The HSMR aims to measure differences in quality of hospital care and its computation was initiated in the United Kingdom by Jarman et al. (1999). Nowadays, it is being computed in a number of countries, such as the United States, Canada, the United Kingdom (Bottle et al. 2011), Australia and the Netherlands. Mortality is taken as a measure of hospital care, since several studies have shown that mortality correlates with quality of hospital care (e.g., Pitches et al. 2007). The HSMR of a hospital is computed as the ratio of observed to expected mortality, normalised to 100 (over all hospitals in a year). The HSMR includes an expected mortality to remove differences between hospitals that are caused by differences in patient populations. The expected mortality is estimated from a logistic regression model that includes a large number of background variables (Israëls et al. 2012).

In the hypothetical situation that we send the same patient to all Dutch hospitals, we would like the hospitals to register the same values for all patient- and disease-related variables. In practice, this is indeed the case in the Netherlands for variables such as age and sex. However, for a number of other variables, differences in reporting frequency were found among hospitals. The largest differences were found for the variables comorbidity and urgency of admission (Jarman, 2008; Pieter et al. 2010; Van der Laan 2013). According to Van den Bosch et al. (2010), reasons for these differences in reporting frequency are time pressure due to a limited number of coders, interpretation differences of the coding rules, and late delivery of patient files. Furthermore, the average number of coders per admission, and consequently, the time typically spent on LBZ registration (Van den Bosch et al. 2010) varies between hospitals. Van der Laan (2013) showed a sharp increase in the average number of reported comorbidities in some hospitals in 2008–2010, where 2010 was the year when the HSMR would become publicly available. Since such a large shift in the patient population of the hospitals in such a short time seems unlikely, this suggests that some hospitals changed their comorbidity reporting (Van der Laan 2013). Note that increased comorbidity reporting – everything else being the same – leads to a decreased HSMR. An increase in the average number of comorbidities by 0.1 led to an estimated HSMR decrease of five points (Van der Laan 2013), implying improved hospital care. This makes the data interesting as a case study. Another quality issue in the hospital data is that hospitals sometimes use the wrong codes (misclassifications) when reporting the main diagnosis or the comorbidities of patients, see for instance Harteloh et al. (2010)

and Quan et al. (2008). Although this is an important quality issue, estimating these misclassifications is beyond the scope of the present article.

In the present study, we focus on the effect of reporting behaviour on estimated changes. In the HSMR case study, for instance, many hours of manual analysis are being spent to clarify whether some hospitals with a changed HSMR have been affected by changes in intensity of comorbidity reporting. Cases of "suspicious results", such as a large change in the average number of comorbidities per hospital stay, are reported by the staff to the data holder, Dutch Hospital Data, which releases the outcomes. Currently, the average number of comorbidities per hospital stay is used as a first simple quality indicator for reporting differences between hospitals. However, applying this simple indicator on previous years showed two serious shortcomings. The first one is that it does not correct for the trend, over all hospitals, in the number of reported comorbidities over time. The second, most serious shortcoming is that it is unclear to what extent changes in the average number of comorbidities per hospital stay affect the estimated HSMR changes of that hospital. The reason is that this effect depends on the patient composition of the hospital. We therefore aim to develop a detection method that overcomes these current shortcomings.

3. Methodology

3.1. Basic Approach

Consider a population U_h of units i ($i = 1, \ldots, N_h$) that are reported in a decentralised administrative data source by data supplier h. Let $\mathcal{Y} = \{y_1, \ldots, y_m, \ldots, y_M\}$ be a set of M binary variables that are prone to under- or overreporting. Further, let the obtained values for the variables y_m for unit i of data supplier h be contained in the vector $\boldsymbol{y}_{hi} = (y_{1hi}, \ldots, y_{mhi}, \ldots, y_{Mhi})^T$. Also, let $\mathcal{Z} = \{z_1, \ldots, z_l, \ldots, z_L\}$ be a set of L covariates (continuous or categorical) for which it is reasonable to assume that they are error-free. Further, let $\boldsymbol{z}_{hi} = (z_{1hi}, \ldots, z_{lhi}, \ldots, z_{Lhi})^T$ be the corresponding vector with the obtained values for unit i of data supplier h. For instance, in the case study, the variables age, sex, socio-economic status and mortality for admissions i of hospital h were considered to be error-free. Further, let θ be the target parameter of interest. In our case study, we have the special situation that we publish a target parameter for each data supplier, denoted by θ_h, but with some minor adaptations our method can also be applied when there is one common target parameter. The target parameter is estimated as $\hat{\theta}_h$, which is a function of the variables y_m and z_l. Throughout the article a hat is used to indicate an estimate.

We aim to compute the effect of under- and overreporting on $\hat{\theta}_h$ for the variables y_m with $m = 1, \ldots, M$. We apply the following four steps to estimate the effect of under- and overreporting (the exact description is given in the next sections):

1. Select a group r of reference suppliers with similar reporting behaviour for the variables y_m. One might use multiple reference groups to analyse the sensitivity of the outcomes to the selected reference group;
2. For the units in the reference group, predict the probability that $y_{mhi} = 1$ given a set of covariates. Use the regression coefficients to predict the probabilities for the nonreference suppliers;

3. Use the predictions of 2) in a predictive mean matching imputation algorithm. If the observed y_{mhi} values of the nonreference suppliers differ significantly from the expected ones, replace them with the reference suppliers' values;
4. Compute the change in the target parameter between two periods for data supplier h as a function of the original $\mathbf{y}_{hi}$ and $\mathbf{z}_{hi}$ values and recompute this change using the imputed values. The difference between those two changes is a measure for the effect of the reporting behaviour of data supplier h on the outcomes.

The four steps are schematically represented in a flow chart, see Figure 1. The details of the steps, for instance the loop over units i in step 3, are explained in the next sections.

3.2. *Select a Reference Group*

We define a model for the variables y_m $(m = 1, \ldots, M)$ to describe the reporting behaviour of the data suppliers. This will be used to select data suppliers with a comparable reporting behaviour. When the intensity of reporting behaviour is expected to be the same for the set of variables y_m one can combine these variables into a single

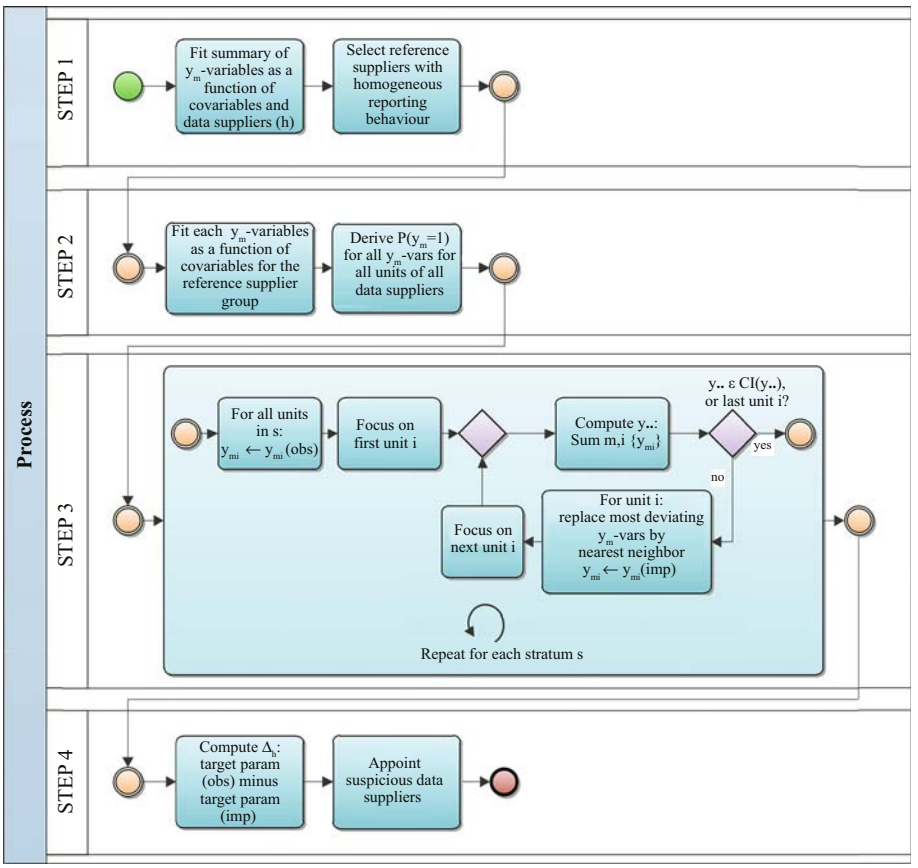

Fig. 1. Flow chart of the four steps of our methodology. The symbols in the chart are simplified compared to the main text; CI stands for confidence interval; the meaning of stratum s is explained in Subsection 3.4.

summary measure. We denote this summary measure as:

$$y_{\bullet hi} = g(\boldsymbol{y}_{hi})$$

where subscript "$\bullet$" denotes that it summarises over a set of variables. A summary variable can be

$$y_{\bullet hi}^{(1)} = 1 - \prod_{m=1}^{M} (1 - y_{mhi}). \tag{1}$$

Thus $y_{\bullet hi}^{(1)}$ equals 0 when $y_{mhi} = 0$ for all y_m variables and it equals 1 otherwise. Alternatively, one might use $y_{\bullet hi}^{(2)} = \sum_{m=1}^{M} y_{mhi}$, which stands for the number of variables with a score of 1. When the variables y_m are not related to each other or when the reporting intensity is expected to vary considerably among the y_m variables, it is better to analyse the effect of reporting behaviour for one variable at a time. In the remainder of this article, we limit ourselves to the analysis of reporting behaviour on a set of variables, because analysing one variable at a time is a special case of this.

In order to assess differences in reporting behaviour among data suppliers, one needs to correct for differences in the population on which they report. We use covariates to capture this population composition. These covariates may coincide with the error-free variables z_l ($l = 1, \ldots, L$) within the administrative data set, but they may also be amended with error-free variables that are not available in the administrative data set at hand. It is important that those variables are error-free to assure unbiased estimates for the supplier effects (the $\hat{\boldsymbol{\gamma}}$ in (2), see below). In the discussion we give some suggestions for the situation that the covariates contain measurement errors. We denote the set of covariates by $x = \{x_1, \ldots, x_k, \ldots, x_K\}$ and their obtained values for unit i of data supplier h are denoted by $\boldsymbol{x}_{hi} = (x_{1hi}, \ldots, x_{khi}, \ldots, x_{Khi})^T$. Let I_{hi} be an indicator variable that is 1 if unit i belongs to data supplier h ($h = 1, \ldots, H$) and 0 otherwise. Let $\boldsymbol{\delta}_{hi}$ be the vector $\boldsymbol{\delta}_{hi} = (I_{1i}, \ldots, I_{Hi})^T$. Further, let $P\left(y_{\bullet hi}^{(1)} = 1\right)$ denote the probability that $y_{\bullet hi}^{(1)} = 1$.

We estimate the data supplier effect on the reporting behaviour using the logistic model:

$$\text{logit}\left\{\hat{P}\left(y_{\bullet hi}^{(1)} = 1 | \boldsymbol{x}_{hi}, \boldsymbol{\delta}_{hi}\right)\right\} = (\boldsymbol{x}_{hi})^T \hat{\boldsymbol{\beta}} + (\boldsymbol{\delta}_{hi})^T \hat{\boldsymbol{\gamma}}, \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ is the vector of estimated regression coefficients concerning the covariates (including the intercept) and $\hat{\boldsymbol{\gamma}} = [\gamma_h]$ is the vector with the estimated data supplier effects. With Equation (2) we assume that there is an overall effect γ_h on a set of binary y_m variables ($m = 1, \ldots, M$) due to the administrative practice of the data supplier. As an alternative to (2) one might estimate the data supplier effect for summary variable $y_{\bullet hi}^{(2)}$ with a simple linear model. When the decentralised data also contains data suppliers that report for a smaller number of units, random effects models might give better estimates of γ_h (see discussion). Note that large γ_h values indicate high reporting levels, whereas small values stand for the opposite.

For (each) reference group r, we aim to select data suppliers with similar γ_h values. Since we are interested in changes of a target parameter ($\hat{\theta}_h$) as affected by shifts in reporting behaviour (between two subsequent periods), we estimate Equation (2) for two subsequent periods and select data suppliers with similar values over two periods. A directly related issue concerns the choice of the group size. This size should, on the one

hand, be small enough to reduce the variability in reporting behaviour within the set, but on the other hand it should be large enough to reliably predict the variables with reporting patterns. See Subsection 4.2 how we operationalised "similar γ_h values" and the group size for the case study.

Note that we regard the computed value for $\hat{P}\left(y^{(1)}_{\bullet hi} = 1 | \boldsymbol{x}_{hi}, \boldsymbol{\delta}_{hi}\right)$ in (2) to be an estimate, although it is derived from administrative data that covers the complete target population. The reason is that we are interested in the reporting behaviour of the data supplier concerning the y_m variables. We regard this reporting behaviour to be an unknown property; the obtained observations can be seen as "input" to monitor this reporting behaviour.

3.3. Predict the Variables with Data Supplier Effects

In the second step, we predict the scores for each of the variables y_m ($m = 1, \ldots, M$) for reference group r and judge how well the models fit. A good model fit leads to a better result for the next step: predictive mean matching. Let d be a domain, that is, a category of one variable or a category of a cross-classification of multiple categorical variables. Domains are used when the effect of the covariates on the error-prone y_m variables are expected to vary (over domains). Domains thus capture interactions between the population composition variables with respect to their effect on reporting intensity. Further, let y_{mhdi} be the score for unit i on variable y_m for data supplier h and domain d. Let U_{rd} be the set of units of reference group r within domain d. Denote by $p^{(r)}_{mhdi} = P\left(y^{(r)}_{mhdi} = 1 | \boldsymbol{x}_{hdi}\right)$ the probability that $y_{mhdi} = 1$ for reference group r given the values of a set of covariates. For the set of units $i \in U_{rd}$ we estimate $p^{(r)}_{mhdi}$ by:

$$\text{logit}\left\{\hat{p}^{(r)}_{mhdi}\right\} = (\boldsymbol{x}_{hdi})^T \hat{\boldsymbol{\beta}}^{(r)}_{md} \quad (i \in U_{rd},\, m = 1, \ldots, M) \tag{3}$$

where $\hat{\boldsymbol{\beta}}^{(r)}_{md}$ stands for the estimated regression coefficients that depend on reference group r, variable y_m and domain d. The periods, for instance years, can be included as dummy variables in $\boldsymbol{x}_{hdi}$, which means that the model captures that reporting behaviour for each of the variables y_m may vary with year. Next, also compute $\hat{p}^{(r)}_{mhdi}$ for the nonreference suppliers based on the same regression coefficients $\hat{\boldsymbol{\beta}}^{(r)}_{md}$ in (3). Note that Equation (3), in contrast to Equation (2), does not contain a data supplier effect ($\hat{\boldsymbol{\gamma}}$). The reason is that we wish to model how the comorbidity probabilities $\hat{p}^{(r)}_{mhdi}$ depend on a set of error-free background variables for a set of units $i \in U_{rd}$ that have a similar reporting behaviour.

We used the C-statistic as an evaluation criterion for the predictive validity of the logistic regressions. The C-statistic lies between 0.5 and 1. As a rule of thumb, values of 0.7 to 0.8 indicate an acceptable discrimination and values above 0.9 show an outstanding discrimination (Hosmer and Lemeshow, 2004).

3.4. Predictive Mean Matching

For ease of notation, in the remainder of the article, we will drop the super- and subscripts r, h and d from the notation of the variables, unless we need them to explain the equations. Thus, for instance $\hat{p}^{(r)}_{mhdi}$ will be abbreviated as $\hat{p}_{mi}$.

In order to analyse the effect of reporting behaviour on the target parameter, we did not directly replace the originally observed y_{mi} values by their $\hat{p}_{mi}$ values, for two reasons:

1. the $\hat{p}_{mi}$ values are estimated for each of the variables y_m separately without accounting for their covariances;
2. we wanted to replace the original data only when the existing values differed clearly from their expected values (see below).

Note that in the HSMR case study, the target variable θ is a nonlinear function of the binary variables y_m $(m = 1, \ldots, M)$ and z_l $(l = 1, \ldots, L)$ (see Section 7). Therefore, directly using the $\hat{p}_{mi}$ will not yield the same outcome as using the binary variables themselves.

We use a nearest neighbour hot deck imputation method, whereby the reference suppliers act as donors and the nonreference suppliers as recipients. We use predictive mean matching as our hot deck imputation method (De Waal et al. 2011). We do not impute all units of the nonreference suppliers: our baseline is that we keep the originally supplied data untouched as much as possible, unless there is a large difference between observed and expected values (similar to selective editing).

Before explaining the algorithm, we introduce some additional notation. Let $\mathcal{H}$ denote the full set of data suppliers and let $\mathcal{R}^{(r)}$ denote the set of reference suppliers for reference group r. Thus, $\mathcal{H} \backslash \mathcal{R}^{(r)}$ stands for the group of nonreference suppliers in case of reference group r. The imputation algorithm is repeated for each combination of reference group r, nonreference data supplier h, domain d and period t. We will refer to this combination by "stratum s" and the set of units in a stratum is denoted by U_s and its size by N_s. Within each stratum U_s we will impute the units one by one. After each imputation, we check the difference between observed and expected values to decide whether or not a new unit is to be imputed, see step three below. Let $\ell = 0, 1, \ldots, \mathcal{L}$ (with $\mathcal{L} \leq N_s$) be an index that counts the number of units that have been imputed (so far). Let $\check{y}_{mi}$ denote an imputed value (0 or 1) for variable y_m $(m = 1, \ldots, M)$ of unit i and let $\tilde{y}_{mi}^{(\ell)}$ denote the actual value of unit i when ℓ units have been imputed, that is

$$\tilde{y}_{mi}^{(\ell)} = \begin{cases} y_{mi} \text{ if not imputed, given that } \ell \text{ units have been imputed} \\ \check{y}_{mi} \text{ if imputed, given that } \ell \text{ units have been imputed} \end{cases} \tag{4}$$

The imputation algorithm consists of three steps:

1. For the units of all nonreference data suppliers ($h \in \mathcal{H} \backslash \mathcal{R}^{(r)}$) compute the sum $y_{\bullet i} = \sum_{m=1}^{M} y_{mi}$. Likewise, compute the expected value as $\hat{E}(y_{\bullet i}) = \sum_{m=1}^{M} \hat{p}_{mi}$. Denote its difference by $\hat{\omega}_{\bullet i} = y_{\bullet i} - \hat{E}(y_{\bullet i})$. Additionally, compute $\hat{E}(y_{\bullet\bullet}) = \sum_{i \in U_s} \hat{E}(y_{\bullet i})$, which is the expected total of $y_{\bullet i}$ within stratum s. Thus, in our case study, the total $y_{\bullet\bullet}$ stands for the number of registered comorbidities over all admissions i in nonreference hospital h and main diagnosis d and year t, and the expectation of $y_{\bullet\bullet}$ is determined for each reference group r.
   Let $V(y_{\bullet\bullet})$ denote the variance of $y_{\bullet\bullet}$. Further, let $\mathrm{L}(y_{\bullet\bullet})$ denote the lower and $\mathrm{U}(y_{\bullet\bullet})$ the upper bound of an (approximate) 95%-confidence interval for $\hat{E}(y_{\bullet\bullet})$.

When the stratum size N_s is large we can estimate these bounds by:

$$\hat{\mathrm{L}}(y_{\bullet\bullet}) = \hat{E}(y_{\bullet\bullet}) - 1.96\sqrt{\hat{V}(y_{\bullet\bullet})} \tag{5}$$

$$\hat{\mathrm{U}}(y_{\bullet\bullet}) = \hat{E}(y_{\bullet\bullet}) + 1.96\sqrt{\hat{V}(y_{\bullet\bullet})} \tag{6}$$

We now derive an expression for $V(y_{\bullet\bullet})$. y_{mi} follows a Bernoulli distribution with $E(y_{mi}) = p_{mi}$. Let $E(y_{mi}y_{ni}) = p_{mni}$. We then find $V(y_{\bullet i}) = \sum_{n=1}^{M}\sum_{m=1}^{M}(p_{mni} - p_{mi}p_{ni})$; for $m = n$ we obtain $p_{mmi} = p_{mi}$. Because the y_{mi} variables are independent across units, $V(y_{\bullet\bullet}) = \sum_{i=1}^{N_s}\sum_{n=1}^{M}\sum_{m=1}^{M}(p_{mni} - p_{mi}p_{ni})$. The values that are generated by (3), $\hat{p}_{mi}$, do not account for interactions between the variables, which implies that we use the approximation $\hat{p}_{mni} = \hat{p}_{mi}\hat{p}_{ni}$ for $m \neq n$. This leads to $\hat{V}(y_{\bullet\bullet}) = \sum_{i=1}^{N_s}\sum_{m=1}^{M}\left(\hat{p}_{mi} - \hat{p}_{mi}^2\right)$, which is an approximation of $V(y_{\bullet\bullet})$. When the y_{mi} variables are positively correlated, $V(y_{\bullet\bullet})$ is underestimated. When they are negatively correlated, $V(y_{\bullet\bullet})$ is overestimated.

2. Let u denote a recipient unit that belongs to the nonreference suppliers and let v be a donor unit that belongs to the reference suppliers. We seek a donor v for recipient u such that the sum of the observed values y_{mv} of the donor will be close to the expected sum of y_{mu} for the recipient. Since this expected sum, $\hat{E}(y_{\bullet u})$, follows from the corresponding probabilities, we select a donor by using the Euclidean distance between $\hat{p}_{mu}$ and $\hat{p}_{mv}$: $\sqrt{\sum_{m=1}^{M}(\hat{p}_{mv} - \hat{p}_{mu})^2}$.
3. Within each stratum s for the nonreference data suppliers:
   a. Set $\ell = 0$;
   b. Compute the totals $\tilde{y}_{\bullet\bullet}^{(\ell)} = \sum_{i\in Us}\sum_{m=1}^{M}\tilde{y}_{mi}^{(\ell)}$ of the actual scores (including imputations). If $\tilde{y}_{\bullet\bullet}^{(\ell)} \notin [\hat{\mathrm{L}}(y_{\bullet\bullet}), \hat{\mathrm{U}}(y_{\bullet\bullet})]$ continue with c), otherwise stop.
   c. If $\tilde{y}_{\bullet\bullet}^{(\ell)} > \hat{E}(y_{\bullet\bullet})$ then determine unit u, from the set of units in s that have not been imputed (so far), with the largest value of $|\hat{\omega}_{\bullet u}|$ given that $\hat{\omega}_{\bullet u} > 0$. Likewise, if $\tilde{y}_{\bullet\bullet}^{(\ell)} < \hat{E}(y_{\bullet\bullet})$ then determine unit u with the largest value of $|\hat{\omega}_{\bullet u}|$ given that $\hat{\omega}_{\bullet u} < 0$. Denote this unit by u_0. For recipient u_0, determine the closest donor v_0 according to the Euclidean distance and impute its values $\boldsymbol{y}_{v_0}$. So, the values for all y_m variables from donor v_0 are imputed.
   d. Let $\ell = \ell + 1$ and go to b.

3.5. Compute the Imputed Target Parameter

Let superscript 0 denote the parameters that are based on the original input values $\boldsymbol{z}_{hi}$ and $\boldsymbol{y}_{hi}$. Let $\hat{\theta}^{t,0}$ denote the target parameter of interest for period t, based on a function of the original input values $\boldsymbol{z}_{hi}$ and $\boldsymbol{y}_{hi}$ for all data suppliers $h \in \mathcal{H}$ and of the estimated model parameters $\hat{\boldsymbol{\beta}}^0$. Note that this function can be a simple sum of $\boldsymbol{y}_{hi}$ or a complex function, which is the case with the HSMR. We now evaluate the effect of the reporting behaviour of one specific data supplier, h_1, by replacing the original input values for that data supplier by its imputed values. Next, we reestimate the target parameter, denoted by $\hat{\theta}^{t,imp(h_1)}$ based on the imputed values for h_1 and the original values for all other data suppliers. Since we

aim to evaluate estimated changes, we compare the change based on the original input values, $\hat{\theta}^{t,0} - \hat{\theta}^{t-1,0}$, with the imputed version: $\hat{\theta}^{t,imp(h_1)} - \hat{\theta}^{t-1,imp(h_1)}$. We denote its difference by $\hat{\Delta}^{t,t-1(h_1)} = (\hat{\theta}^{t,0} - \hat{\theta}^{t-1,0}) - (\hat{\theta}^{t,imp(h_1)} - \hat{\theta}^{t-1,imp(h_1)})$. We analyse $\hat{\Delta}^{t,t-1(h)}$ for all nonreference data suppliers $h \in \mathcal{H}\backslash\mathcal{R}^{(r)}$. In the case study, we have the special situation that we have a target parameter $\hat{\theta}_h^0$, but the model parameters $\hat{\boldsymbol{\beta}}^0$ continue to be based on all $h \in \mathcal{H}$. This leads to a small modification, which is explained in Subsection 4.2.

Since our method aims to select nonreference data suppliers with extreme values of $\hat{\Delta}^{t,t-1(h)}$, we wish to have a practical rule to appoint which values we consider to be extreme. To that end, we assume that for data suppliers h that are free of under- or overreporting, the values of $\hat{\Delta}^{t,t-1(h)}$ are approximately normally distributed: $\hat{\Delta}^{t,t-1(h)} \sim N\left(0, \sigma^2_{\hat{\Delta}}\right)$. The expected value of $\hat{\Delta}^{t,t-1(h)}$ is taken to be 0, since we expect that the y_m values $(m = 1, \ldots, M)$ of data suppliers without under- or overreporting are not imputed. Further, $\sigma^2_{\hat{\Delta}}$ stands for variation in the outcomes of $\hat{\Delta}^{t,t-1(h)}$ that cannot be explained by the covariates used in the regression. Here, we limit the estimation of $\sigma^2_{\hat{\Delta}}$ to the situation that we have two reference groups r; it can easily be extended to a situation with more reference groups. The situation of a single reference group is treated in the discussion.

Denote the first reference group by A, its corresponding set of data suppliers by $\mathcal{R}^{(A)}$ and its size by $N_{\mathcal{R}^{(A)}}$. Likewise, we denote the second reference group by B with set $\mathcal{R}^{(B)}$ of size $N_{\mathcal{R}^{(B)}}$. When A is selected as the reference group, values of $\hat{\Delta}^{t,t-1(h)}$ are not available for $\mathcal{R}^{(A)}$, since only the values of the nonreference suppliers are imputed, but they are available for $\mathcal{R}^{(B)}$. We consider the variation in $\hat{\Delta}^{t,t-1(h)}$ for $\mathcal{R}^{(B)}$ when A is the reference group as an estimate of $\sigma^2_{\hat{\Delta}}$, since we have selected the set within a reference group to be (more or less) homogeneous in reporting behaviour. We define:

$$s^2_{\hat{\Delta}}(\mathcal{R}^{(B)}|r = A) = \frac{1}{N_{\mathcal{R}^{(B)}} - 1} \sum_{h \in \mathcal{R}^{(B)}} (\hat{\Delta}^{t,t-1(h)} - \hat{\bar{\Delta}}^{t,t-1(B)})^2 \tag{7}$$

with $\hat{\bar{\Delta}}^{t,t-1(B)} = \frac{1}{N_{\mathcal{R}^{(B)}}} \sum_{h \in \mathcal{R}^{(B)}} \hat{\Delta}^{t,t-1(h)}$. Note that in (7) we used the sample mean $\hat{\bar{\Delta}}^{t,t-1(B)}$ with $N_{\mathcal{R}^{(B)}} - 1$ degrees of freedom rather than using the expected value "0". Likewise, we define:

$$s^2_{\hat{\Delta}}(\mathcal{R}^{(A)}|r = B) = \frac{1}{N_{\mathcal{R}^{(A)}} - 1} \sum_{h \in \mathcal{R}^{(A)}} (\hat{\Delta}^{t,t-1(h)} - \hat{\bar{\Delta}}^{t,t-1(A)})^2 \tag{8}$$

We now estimate $\sigma^2_{\hat{\Delta}}$ as the pooled estimate of $s^2_{\hat{\Delta}}(\mathcal{R}^{(B)}|r = A)$ and $s^2_{\hat{\Delta}}(\mathcal{R}^{(A)}|r = B)$:

$$\hat{\sigma}^2_{\hat{\Delta}} = \frac{(N_{\mathcal{R}^{(A)}} - 1) s^2_{\hat{\Delta}}(\mathcal{R}^{(A)}|r = B) + (N_{\mathcal{R}^{(B)}} - 1) s^2_{\hat{\Delta}}(\mathcal{R}^{(B)}|r = A)}{N_{\mathcal{R}^{(A)}} + N_{\mathcal{R}^{(B)}} - 2} \tag{9}$$

Using $\hat{\sigma}^2_{\hat{\Delta}}$, we construct an (approximate) 95%-confidence interval for $\hat{\Delta}^{t,t-1(h)}$ as $0 \pm 1.96\sqrt{\hat{\sigma}^2_{\hat{\Delta}}}$. Data suppliers outside this confidence interval are considered to have a deviating reporting behaviour.

4. Application to the Case Study

The method described in the previous section was applied to the HSMR case study, which is calculated from the LBZ data. The HSMR is limited to hospital stays, also called inpatient admissions. Day admissions are excluded, since they are usually non-life-threatening. As was mentioned in the introduction, there are strong indications that some of the variables in this data set are affected by reporting differences between the hospitals, which affects the output. The next subsections describe the administrative data and how our method is applied to the HSMR case study. The method of calculating the HSMR is given in the Appendix. Section 5 describes the results.

4.1. LBZ Data

We used LBZ data for the consecutive years 2011 and 2012 with a total of 1,221,414 inpatient admissions. We wanted to use data near 2010, which was announced to be the first year when the HSMR would become publicly available. We found clear shifts in intensity of comorbidity reporting by some of the hospitals a few years before and after that period. Although the HSMR model (see Section 7, Appendix) is normally estimated over three years, we did not include 2013 or more recent years as most hospitals switched to a new coding system (ICD10) in 2013, which might affect the results (Van der Laan et al. 2015). In total, 83 hospitals provided LBZ data for both 2011 and 2012. Four of these hospitals submitted data of poor quality (an incomplete data set). One hospital was very specialised with only a few main diagnoses. We excluded those five hospitals and used a net population of 78 hospitals in the analysis.

4.2. Analysis of the Reporting Effects

The target parameter θ_h is in this case the HSMR of hospital h (see Appendix and Van der Laan et al. 2015 for how this parameter is calculated). The HSMR is the ratio between the observed mortality and the expected mortality. The expected mortality is calculated using a logistic regression model for mortality at patient level using background properties such as age, sex and comorbidities. We studied the effect of reporting intensity on the comorbidity variables. These comorbidity variables were grouped into 17 Charlson groups (see Section 7, Appendix). So in the case study, the variable y_m ($m = 1, \ldots, M$) stands for one of the 17 Charlson groups for admission i of hospital h. We now describe how we applied Subsections 3.2–3.5 to our case study.

4.2.1. First Step

We asked experts which hospitals they thought were expected to have correct comorbidity reporting, but they were unable to answer the question. Therefore, we decided to select *two* reference groups, to be able to determine the sensitivity of the outcomes for the choice of the reference group. We selected a "middle group" representing average levels of comorbidity reporting and a "top group" representing high levels of comorbidity reporting.

To summarise the variables y_m we used $y^{(1)}_{\bullet hi}$, defined in Equation (1), which has a value 1 when at least one comorbidity code is given to admission i and 0 otherwise. We used the

logistic regression model in (2) to model $y^{(1)}_{\bullet hi}$ as a function of the hospital effect γ_h and of a set of diagnoses- and patient-related variables that are given in Table 1 in the Appendix (column "select reference hospitals"). The logistic regression was applied to 2011 and 2012 separately. Within a year, it was applied to all admissions of all hospitals. This means that we did not let the hospital effect vary with main diagnosis, but it did vary with year. We did so, because we were interested in estimating the overall hospital effect on reporting behaviour and because previous experience showed that hospitals may vary their reporting behaviour from one year to the next. Note that in Equation (2) we used a fixed hospital effect γ_h. Prins (2016) has also computed outcomes where the hospital effect was included as a random effect within a multilevel model, which yielded near-identical results.

We wanted to select two reference groups that were homogeneous in their reporting behaviour for two subsequent years (2011, 2012). Thus, we wanted the γ_h values not to vary too much from one year to the next. We computed the difference $d_h^{t,t-1} = \gamma_h^t - \gamma_h^{t-1}$ of the hospital effects, with $t = 2012$ and its variance: $V\left(d_h^{t,t-1}\right) = \frac{1}{H-1}\sum_{h=1}^{H}\left(d_h^{t,t-1} - \bar{d}_h^{t,t-1}\right)^2$, with $\bar{d}_h^{t,t-1} = \frac{1}{H}\sum_{h=1}^{H} d_h^{t,t-1}$. Next, we computed an (approximate) 80% confidence interval according to $\bar{d}_h^{t,t-1} \pm 1.28\sqrt{V\left(d_h^{t,t-1}\right)}$, based on a Normal distribution. Hospitals with $d_h^{t,t-1}$ values outside this interval were excluded from the reference group. For the "middle reference group" we selected the hospitals within the 80% confidence interval whose absolute γ_h^{2012} values were closest to 0. For the "top reference group" we selected the hospitals within the 80% confidence interval with the largest γ_h^{2012} values. A reference group size of 15 (we investigated 10, 15 and 20) was found to be the smallest group size for which the models could be reasonably accurately

Table 1. Variables used in the various computations.

Variable (no of classes[1])	HSMR model	Select reference hospitals	Predict each comorbidity (15)
Age (5-year classes)	x	x	x (5 knot spline)
Comorbidity group (17)	x		
Hospital (78)		x	
Main diagnosis (50)	*	x	*
Medical specialty (44)			x
(bi-) Month of admission (6)	x	x	x
Re-admission (2)	x	x	x
Reason of admission (3)			x
Sex (2)	x	x	x
Severity main diagnosis (9)	x	x	x
Social-economic status (6)	x	x	x
Source of admission (3)	x	x	x
Type of hospital (2)			x
Urgency (2)	x	x	x
Year of discharge (2)	x	*	x

x: variable included as independent variable in the regression; *: class for which the regression is run separately.
[1] an explanation of the categories can be found in Israëls et al. (2012) and van der Laan (2013)

estimated (all categories reasonably filled; few categories in the recipient group that were not present in the donor group).

4.2.2. Second Step

The second step in the case study was to predict each of the comorbidity variables y_m as a logistic function of patient- and disease-related variables, according to Equation (3). Because the occurrence of comorbidities varied greatly with main diagnosis, the model was fitted separately for each main diagnosis. Thus, main diagnosis represents domain d in Equation (3). For instance, Charlson group 15 (HIV) occurs mainly with main diagnosis 38 (non-Hodgkins lymphoma), see Van der Laan et al. (2015). The patient- and disease-related variables used in this second step are given in the final column of Table 1 in the Appendix. In addition to the covariates that we used to select the reference hospitals, we added medical specialty, type of hospital and reason of admission. For the regressions, two sets of comorbidity groups were very similar and therefore combined: Charlson comorbidity group 17 (Severe liver disease) was combined with Charlson comorbidity group 9 (Liver disease) and Charlson comorbidity group 11 (Diabetes complications) was combined with Charlson comorbidity group 10 (Diabetes), leading to a total of 15 groups (see Table 1, final column). Charlson groups 17 and 11 occur only rarely, and a preliminary analysis showed that merging these comorbidity groups had only a minor effect on the HSMR outcomes (Israëls et al. 2012).

4.2.3. Third Step

The third step was to apply the imputation algorithm. To estimate $V(y_{\bullet\bullet})$ we assumed that the probabilities p_{mi} for most of the variables are very small. We assumed that $\hat{p}_{mi} - \hat{p}_{mi}^2 \approx \hat{p}_{mi}$. Recall from Subsection 3.4 that $\hat{V}(y_{\bullet\bullet}) = \sum_{i=1}^{N_s} \sum_{m=1}^{M} (\hat{p}_{mi} - \hat{p}_{mi}^2)$. We now approximate this variance by $\hat{V}(y_{\bullet\bullet}) \approx \sum_{i=1}^{N_s} \sum_{m=1}^{M} \hat{p}_{mi} = \hat{E}(y_{\bullet\bullet})$. Note that the same variance would have been obtained by assuming that the data are Poisson-distributed.

In the case study, we used the Euclidean distance between the probabilities $\hat{p}_{mu}$ of recipient u and $\hat{p}_{mv}$ of donor v. We explain why we use a distance function based on probabilities in Subsection 3.4 (step 2 of the imputation algorithm). In preliminary computations, we also performed the imputation algorithm using the Euclidean distance between the logit of the probabilities, which resulted in near identical results.

4.2.4. Fourth Step

The fourth and final step was to compute the HSMR for each of the hospitals, according to Equations (10) and (11) in the Appendix, based on the observed and the imputed comorbidity scores. Let $\hat{\theta}_h^{t,0}$ denote the HSMR based on the original input values for hospital h and year t and let $\hat{\theta}_h^{t,imp}$ be its imputed version. $\hat{\theta}_h^{t,0}$ is estimated according to the logistic regression for the HSMR given in (11). The imputed HSMR for a specific data supplier h_1, $\hat{\theta}_{h_1}^{t,imp}$, is defined in Subsection 3.5 as the outcome of (11), based on the values $\tilde{y}_{h_1 mi}$ and the original values for the other input variables, combined with the original values for all other data suppliers $h \neq h_1$. Since the data set is changed only slightly, we assume that we can ignore the change in the regression coefficients when we impute

only one data supplier. Therefore, we use the original regression coefficients when calculating $\hat{\theta}_h^{t,imp}$.

Finally, we compared the HSMR change based on the original comorbidity values, $\hat{\theta}_h^{2012,0} - \hat{\theta}_h^{2011,0}$, with the change based on the imputed version: $\hat{\theta}_h^{2012,imp} - \hat{\theta}_h^{2011,imp}$. We denote its difference by $\hat{\Delta}_h^{2012,2011} = \left(\hat{\theta}_h^{2012,0} - \hat{\theta}_h^{2011,0}\right) - \left(\hat{\theta}_h^{2012,imp} - \hat{\theta}_h^{2011,imp}\right)$.

5. Results

5.1. Selection of Reference Group

The hospital effects in 2012 $\left(\gamma_h^{2012}\right)$ and the differences in those hospital effects over the two years $\left(\gamma_h^{2012} - \gamma_h^{2011}\right)$ are given in Figure 2. The γ_h^{2012} values ranged from -2.57 to 1.29. These values are the logarithm of the log-odds of the probabilities of reporting at least one comorbidity per hospital admission. These probabilities can be found by $\hat{P}\left(y_{\bullet hi}^{(1)} = 1 | \boldsymbol{x}_{hi}, \boldsymbol{\delta}_{hi}\right) = \frac{1}{1+\exp\left\{-(\boldsymbol{x}_{hi})^T \hat{\boldsymbol{\beta}}_d - (\boldsymbol{\delta}_{hi})^T \hat{\boldsymbol{\gamma}}\right\}}$ from Equation (2). Values of the logistic regression are given as the difference to a reference category (in case of categorical variables). So, for a patient admission that matches the reference category, the probability of having at least one comorbidity in 2012 among the hospitals ranged from $\frac{1}{1+\exp\{2.57\}} = 0.22$ to $\frac{1}{1+\exp\{-1.29\}} = 0.93$. These results indicate that there was considerable variation

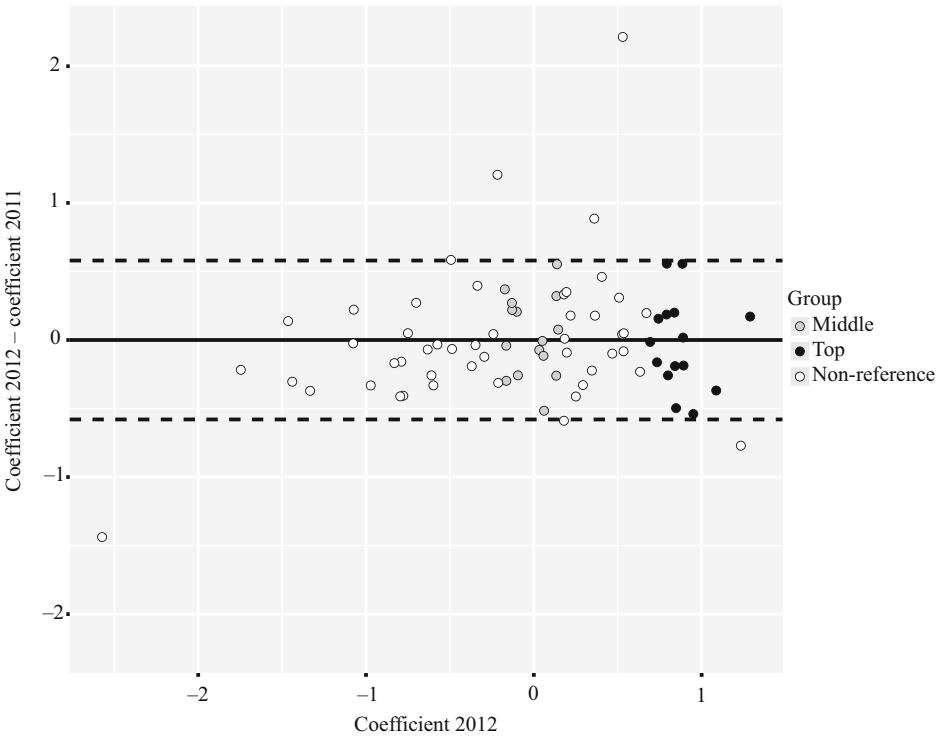

Fig. 2. The difference in the hospital effects (γ_h) of 2012 minus 2011 versus the hospital effects of 2012.

among the hospitals in intensity of comorbidity reporting, after correcting for differences in patient and diagnosis characteristics.

Given a group size of 15 units, the grey points in Figure 2 show the middle reference group and the black points show the top reference group. A standard deviation of 0.452 was found for $d_h^{2012,2011}$ leading to an 80% margin of $\pm$ 0.580. Using these margins as an additional selection criterion implied that one hospital was excluded from the top reference group (with $\gamma_h^{2011} = 2.00$ and $\gamma_h^{2012} = 1.23$) and no hospital was excluded from the middle group (the extreme value shown in Figure 2 with 0.179 in 2012 had order position 16).

5.2. *Prediction of the Incidence of the Charlson Groups*

The fit of the predicted probabilities $\left(\hat{p}_{mi};\ m = 1, \ldots, M\right)$ based on the admissions of the two reference groups according to (3) varied slightly between the different Charlson groups. The averages of the C-statistic for the middle and top groups were relatively small for Charlson group 6 (0.78, 0.71) and 10 (0.74, 0.72) whereas they were large for Charlson group 5 (0.89, 0.90), 8 (0.89, 0.89) and 9 (0.91, 0.87). Overall, the fraction of "main diagnosis $\times$ Charlson group" combinations with a C-statistic of at least 0.7 was 0.92 for the middle group and 0.86 for the top group. Since values of 0.7 and higher indicate an acceptable fit (see Subsection 3.3) we considered the results of the C-statistics to be sufficient to use the predicted probabilities ($\hat{p}_{mi}$) for predictive mean matching.

5.3. *Results of the Predictive Mean Matching*

Figure 3 displays for each hospital the distribution of the fraction-imputed records after applying the imputation algorithm across the 50 main diagnosis groups and the two years for both reference groups. We computed the average and third quantile per hospital of this distribution over diagnosis groups and years. The average per hospital was at most 0.16 (hospital 78) in case of the middle reference group and 0.24 (hospital 48) in case of the top reference group. Furthermore, the third quantile of this distribution was at most 0.24 (hospital 78) for the middle reference group and 0.34 (hospital 48) for the top reference group. These findings clearly show that only a limited number of records for each hospital were imputed, which is in line with the imputation approach that we intended (see Subsection 3.4). The minimum value of these averages over the set of recipient hospitals was 0.003 (middle reference group) and 0.015 (top reference group). This implies that for all recipient hospitals at least some records were imputed.

We also computed the average fraction of imputed records per main diagnosis group over the set of the recipient hospitals and the two years (not shown). For the middle reference group, the three smallest average fractions were 0.007, 0.010 and 0.015 and the three largest ones were 0.0977, 0.134 and 0.141. In the top reference group, the three smallest average fractions were 0.0208, 0.0255 and 0.0278 and the three largest ones were 0.181, 0.202 and 0.224. There were a few "main diagnosis $\times$ hospital" combinations where all admissions were imputed. This was not always for the same hospital or for the same diagnosis.

When a certain category of the covariates occurred in the recipient set that was absent in the donor set, the $\hat{p}_{mi}$ probabilities could not be predicted and those records were excluded

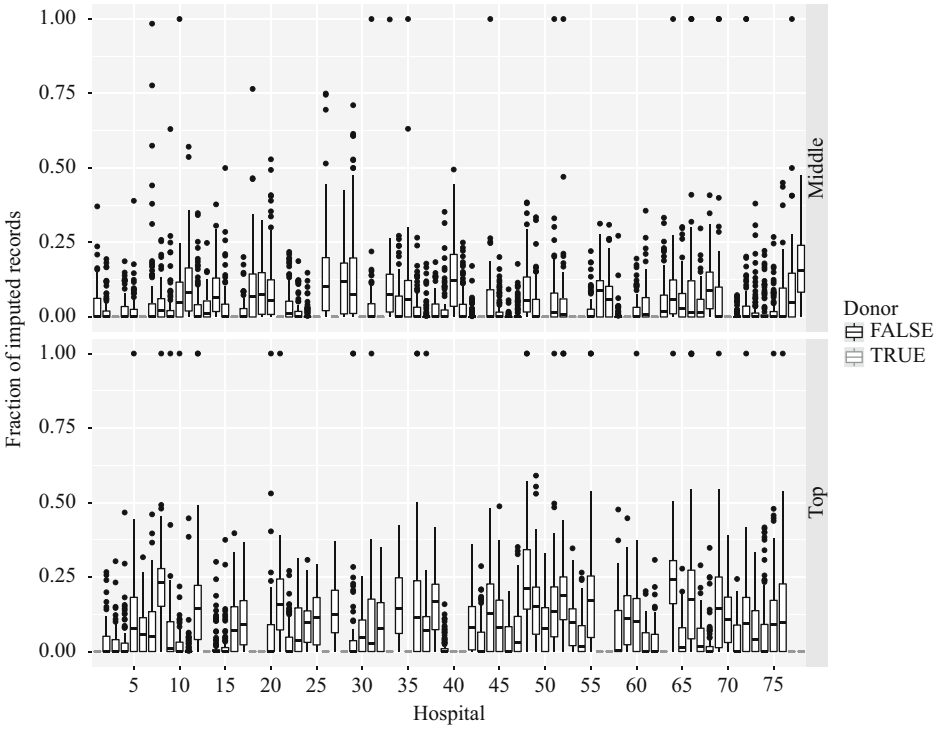

Fig. 3. Boxplot of the fraction of imputed records (distribution over 50 main diagnoses and both years) for each hospital with the middle (upper panel) or the top reference group (lower panel).

from the imputation algorithm. We computed the average, median, standard deviation, maximum and minimum fraction of units without a predicted comorbidity group score per main diagnosis. These values were 0.13, 0.070, 0.18, 0.93, 0.0058 for the middle and 0.080, 0.039, 0.14, 0.79, 0.0017 for the top reference group. In both reference groups, the median fraction of units without predicted scores was small, but in each reference group there were a few main diagnoses with a large fraction. The reason for this larger fraction was that those main diagnoses occurred mainly in certain categories of the patient- and diagnosis-related variables that were (by accident) absent in the reference group.

5.4. Computation of the Imputed HSMR

Production staff at CBS are interested in knowing to what extent the original HSMR development represents a change in the quality of hospital care or whether it results from a change in intensity of comorbidity reporting. The fraction of reported Charlson groups per admission for a given year is denoted by $\bar{y}_h$ and defined as $\bar{y}_h = \frac{1}{N_h M} \sum_{i \in Uh} \sum_{m=1}^{M} y_{mhi}$, where N_h stands for the number of admissions for hospital h. The development of $\bar{y}_h$ from 2011 to 2012 is denoted by $\bar{y}_h^{2012,2011}$, with $\bar{y}_h^{2012,2011} = \bar{y}_h^{2012} - \bar{y}_h^{2011}$. In Figure 4 we plotted $\hat{\Delta}_h^{2012,2011}$ against $\bar{y}_h^{2012,2011}$ and we fitted a simple linear regression through the data. We tested whether the slope differed from zero, under the assumption that the residuals are independent and identically distributed. We found a slope of -42.3 for the middle

reference group and of -55.7 for the top reference group at a p-value < 0.001. Recall that an increase in comorbidity reporting – everything else being the same – leads to a decrease in the HSMR. The latter represents an improvement in the quality of hospital care. The regression results imply that an increase in "the fraction of reported Charlson groups" $\left(\bar{y}_h^{2012,2011}\right)$ of 0.1 leads to an HSMR development which reduced by 4.2 points (middle group) or 5.5. point (top reference group). So, the hospitals that are plotted in the bottom-right of Figure 4 are hospitals with a large increase in comorbidity reporting from 2011 to 2012. It concerns hospitals where the original HSMR development $\left(\hat{\theta}_h^{2012,0} - \hat{\theta}_h^{2011,0}\right)$ is lower than the imputed one $\left(\hat{\theta}_h^{2012,imp} - \hat{\theta}_h^{2011,imp}\right)$, suggesting that their original improvement in the quality of hospital care was partly due to reporting effects.

Figure 5 shows that the $\hat{\Delta}_h^{2012,2011}$-values of the nonreference hospitals (circles) with the middle reference group were clearly related to those of the top group, with a correlation of 0.91. We found an estimated variance $s^2_{\hat{\Delta}}$ of 11.57 when the middle group was used as the reference group and of 15.24 when the top group was used as reference group. This resulted in a pooled variance of 13.41 and an estimated 95%-confidence interval of ± 7.17 index points. We thus found five hospitals with significant reporting effects. Three hospitals in the bottom-left of Figure 5 have a negative value for $\hat{\Delta}_h^{2012,2011}$ and in Figure 4

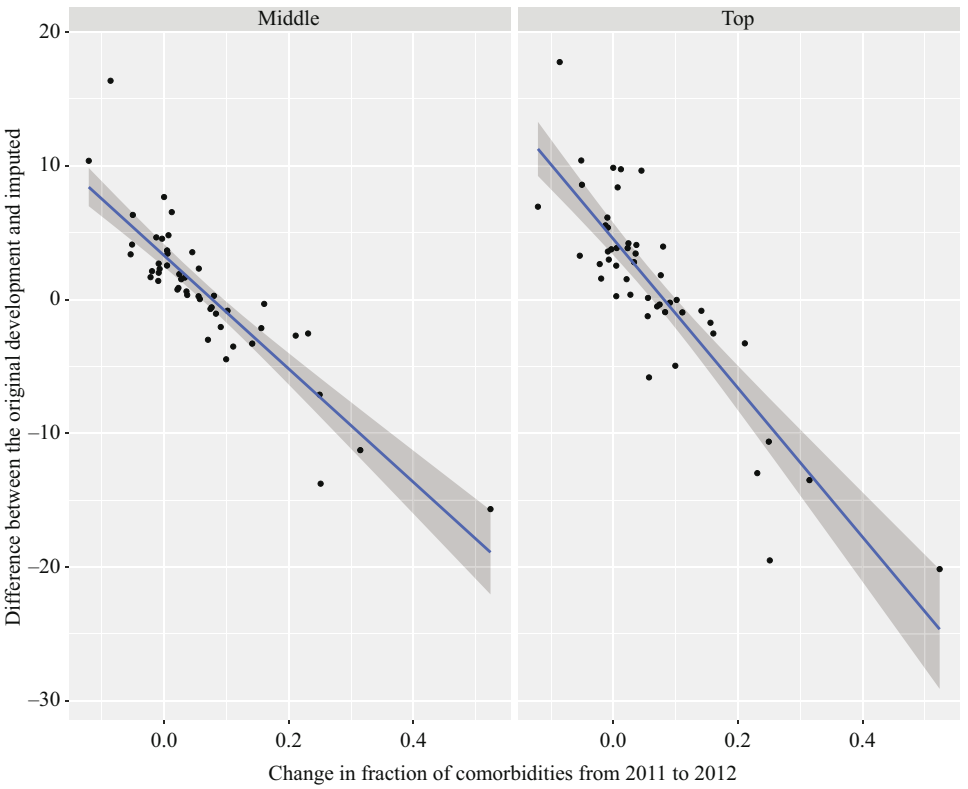

Fig. 4. Difference between the original and the imputed HSMR development as a function of the change in the fraction of comorbidities (2012 minus 2011) for both reference groups. Shaded areas represent the 95% confidence intervals of the linear regressions.

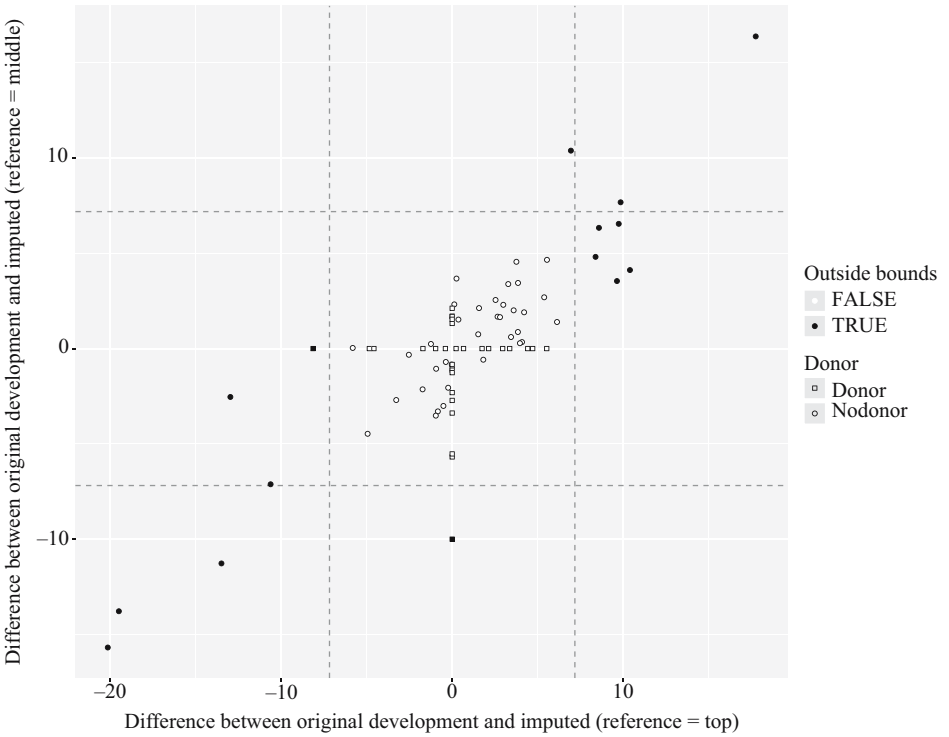

Fig. 5. Change (2012 – 2011) of the original minus that of the imputed HSMR $\left(\hat{\Delta}_h^{2012,2011}\right)$ *for the middle group versus the top group of reference hospitals. Broken lines represent confidence bounds based on* $\hat{\sigma}^2_{\hat{\Delta}}$.

we can see that it concerns three hospitals with a large increase in comorbidity reporting, suggesting that their original change in hospital care was "too positive". Two hospitals in the top-right have a positive value for $\hat{\Delta}_h^{2012,2011}$ suggesting that their original change in hospital care was "too negative". These two hospitals had a very low value of original comorbidity reporting in both years (not shown). These five hospitals are the suspicious hospitals to be contacted.

6. Discussion

We presented a method to detect under- and overreporting by data suppliers for decentralised administrative data in case of change estimates. With our approach, we estimated the impact of correlated measurement errors within a data supplier on the target outcomes. We successfully applied the method to administrative hospital data to detect hospitals that show large changes in reporting of the comorbidities of their patients. Previous studies have also found reporting differences among hospitals (Jarman 2008; Van der Laan 2013) but they were unable to estimate the impact of the reporting intensity on the outcomes. With the current method we expect to reduce the number of hours spent on manual data analysis. Moreover, we can contact the suspicious data suppliers, in order to improve the accuracy of future administrative data deliveries. The question remains how to proceed with the output based on the current data delivery. When reporting errors of a

variable that are widespread and severe, one might decide not to publish the outcomes that are based on this variable. When it concerns only a limiting number of data suppliers, then one might set those cases to "missing" and use a robust estimation, a weighting model or an imputation model to correct for it. In the special case of the HSMR, the output is at the level of the data supplier. When the quality of data delivered by a certain supplier is insufficient, one might publish a remark along with the outcomes, decide not to publish the output of that supplier, or exclude that data supplier from the data set, depending on the severity of the errors.

Our method is developed for a situation with decentralised administrative data, where it is possible to detect differences in reporting behaviour among data suppliers, but it is not possible to exactly pinpoint which units within the suspicious data supplier have measurement errors. This is opposed to the situation described in Van Delden and Scholtus (2017), where reporting patterns at unit level are detected with deterministic rules. That concerned turnover patterns derived from reported value added tax data. Our method requires that the total set of data suppliers $\mathcal{H}$ is large enough to set aside a group $\mathcal{R}$ that can act as reference suppliers.

A number of points are to be addressed before our method can be applied in statistical production and before it can be applied to forms of decentralised administrative data other than hospital data. First, tooling should be developed to enable analysts to perform the four steps in Subsection 3.1. Second, some practical guidance is needed in treating the decentralised data structure in the parameter estimates. Third, a practical application of our method would be enhanced by relaxing some of the assumptions and conditions of the currently reported method, because they may not hold in practice. Fourth, for application of the method to other decentralised data than the hospital data, it would be very useful to extend the method in terms of the types of variables it concerns, the types of errors treated and the forms of output. In the next four paragraphs we elaborate on the second, third and fourth point.

The decentralised, hierarchical, data structure needs to be accounted for in the estimation of the regression coefficients in step 1. In practical applications, one has to choose whether to treat the data supplier effects as random effects within a multilevel model or as fixed effects. In a data set where at least some of the data suppliers have a limited number of units per data supplier, we would prefer to model the data supplier effects as random effects, since one can then make use of the shrinkage factor (Efron and Morris 1975). In the HSMR case study, we had a large number of units for each data supplier. We found that treating the hospital effects as random or fixed effects yielded near-identical estimates, whereas the convergence of the latter model was much faster than that of the multi-level model, in line with Kim et al. (2013).

An example of a useful relaxation concerns the imputation algorithm (step 3). The current procedure does not account for the actually reported scores $y_{mi} = 1$. We propose the following refinement. Let $\mathcal{Y}_v$ be the set with $y_{mv} = 1$ (with $m = 1, \ldots, M$), for recipient v and let $\mathcal{Y}_u$ be the corresponding set for donor u. If the size of $\mathcal{Y}_v$ is smaller than expected, thus $y_{\bullet v} < \hat{E}(y_{\bullet v})$, which is an indication for underreporting, then it might be reasonable to assume that any reported values $y_{mv} = 1$ are correct and replacing them in the imputation algorithm by zeros should be avoided. In that case, one might limit the set of donors to those for which it holds that the observed set is a subset of the donor set:

$\mathcal{Y}_u \supset \mathcal{Y}_v$. Conversely, if the size of $\mathcal{Y}_v$ is larger than expected, thus $y_{\bullet v} > \hat{E}(y_{\bullet v})$, which is an indication of overreporting, a donor u could be selected such that the donor set is a subset of the observed set: $\mathcal{Y}_u \subset \mathcal{Y}_v$. This refinement is only feasible when the donor set is large enough.

Another example of a relaxing a condition of the reported method concerns the estimation of the residual variance $\sigma^2_{\hat{\Delta}}$. In our article the estimation of $\sigma^2_{\hat{\Delta}}$ is based on multiple reference groups, but the question remains how this variance can be estimated with a single reference group. The latter situation might occur when a reference group is appointed by experts. This residual variance stems from four error sources. The first is that not all covariates explaining the y_m variables ($m = 1, \ldots, M$) in Equation (3) might in fact be available, so there is unexplained variance. A second, related, cause is uncertainty in the imputation procedure due to uncertainty in the regression coefficients and in appointing the nearest neighbour. A third error source is the presence of random reporting errors in the y_m variables among the data suppliers in the reference group. The extent of this error source might be investigated by letting two or more administrators independently register the same cases. The final issue is that we are interested in capturing the reporting behaviour of data suppliers, whereas data from a single data supplier can be seen as just one realisation of an (unknown) distribution. Using multiple years of data from the same supplier might help to analyse the extent of this error source. When all four error sources are quantified, one might apply a repeated sampling procedure to estimate their effect on the variance $\sigma^2_{\hat{\Delta}}$. Possibly, a multiple imputation approach, originating from Rubin (1978, 1987), is useful in this context. Using that approach, we then aim to draw multiple versions of the regression coefficients of Equation (3) that capture the combined effect of the four error sources. Next, multiple versions of the matching algorithm and of $\hat{\Delta}^{t,t-1(h)}$ are obtained. It needs to be tested whether this approach yields good results.

Before our method can be applied to forms of decentralised administrative data other than the hospital data, research is needed on adaptation and extension of the method to other types of variables, errors and output forms. First, we will give two examples of potential other applications and then we will go into those adaptations and extensions. CBS has municipalities' administrative data on inhabitants' receipt of social benefits. It not only concerns social benefit data, but also additional information such as fraud occurrence, estimated fraud values, and training activities to find a job. Different municipalities have different reporting intensities, especially concerning the additional information. Suppose our aim is to detect changes in the intensity of fraud activities. We can then use covariates such as received social benefit, age, profession, social economic status, current duration of unemployment and so on (we have a social statistical database with many potentially useful variables). We could compute the (expected) changes in fraud intensity per municipality after applying steps 1–3, compare this with the original changes and select the suspicious municipalities. Likewise, we could detect underreporting of the occurrence of environmental damage reported by fire brigades using covariates like type of building, type of surrounding, presence of chemicals and so on.

It would be useful to apply the method to new examples to find out whether it works, and which adaptations or extensions are needed. We have foreseen some of those

adaptations and extensions already. A first small adaptation to the method can be done when one applies the approach to a *single* binary variable (representing reporting behaviour) at a time, rather than to a *set* of variables. Then, one could replace the predictive mean matching step by drawing a binary value from a Bernoulli distribution for each unit in the data set using the estimated probabilities. Second, the method could be extended by handling *continuous* variables with reporting errors in a selective group of data suppliers. That requires a robust way of estimating the data supplier effects (γ_h) especially in the case of large measurement errors (Rousseeuw and Leroy 1987). Third, it would be useful to develop an analysis procedure that combines the detection of under- and overreporting in classification variables with that of misclassifications. A fourth extension would be to increase the level of detail: in addition to analysing effects at data supplier level, one could analyse effects in underlying domains. Those underlying domains could, in fact, represent administrative agencies underlying the formal data suppliers, for instance clinics within large hospitals or establishments within large schools. When the reference group is selected at the more detailed domain level one may have to find a procedure to deal with a limited number of units per domain. A fifth extension is to address the effect of reporting behaviour on *level* estimates. This requires a subset of data suppliers for which we are certain that they are reporting correctly. One way to do this is to use expert knowledge to obtain such a set. It is a point of future research whether there are other possibilities for such an analysis.

7. Appendix: Computation of the HSMR

The target parameter θ_h, denoting the HSMR for hospital h, is computed as follows. Let O_{hd} be the observed mortality for main diagnosis d of hospital h and let E_{hd} be the corresponding expected mortality based on the patient population. Further, let θ_{hd} be standardised mortality ratio (SMR) for the set of units U_{hd} within main diagnosis d of hospital h. The SMR is an indicator for the quality of hospital care per main diagnosis. θ_{hd} is given by

$$\theta_{hd} = 100\frac{O_{hd}}{E_{hd}} \tag{10}$$

with $O_{hd} = \sum_{i \in U_{hd}} D_{hdi}$ and $E_{hd} = \sum_{i \in U_{hd}} E_{hdi}$, where D_{hdi} is a variable that equals 1 when the patient died during hospital admission i and 0 otherwise. The expected mortality E_{hdi} for admission i is estimated from a logistic regression with patient- and diagnosis-related variables as covariates. Hospital-related variables are left out of the model, such as the number of doctors per bed, because these are directly related to the quality of hospital care that the HSMR tries to measure. This logistic regression is fitted for each main diagnosis separately. Recall that the patient- and diagnosis-related covariates are split up into an error-free and an error-prone part. Let $\boldsymbol{z}_{hdi} = \left(z_{1hdi}, \ldots, z_{lhdi}, \ldots, z_{Lhdi}\right)^T$ denote the L-vector of error-free covariates (including the intercept), $\boldsymbol{y}_{hdi} = \left(y_{1hdi}, \ldots, y_{mhdi}, \ldots, y_{Mhdi}\right)^T$ denote the M-vector of error-prone covariates and $\boldsymbol{\beta}_d$ denote the vector of regression coefficients for the joint covariates

vector. $\hat{E}_{hdi}$ is given by:

$$\hat{E}_{hdi} = \hat{P}(D_{hdi} = 1|\boldsymbol{x}_{hdi}) = \frac{1}{1 + \exp\{-[(\boldsymbol{z}_{hdi})^T, (\boldsymbol{y}_{hdi})^T]\hat{\boldsymbol{\beta}}_d\}} \quad (11)$$

The patient- and diagnosis-related variables are given in Table 1.

Let $\mathcal{D}$ be the set of main diagnoses that are included in the computation of the HSMR. We included 50 out of 200 main diagnoses in the HSMR computation. These 50 diagnoses comprised about 80 per cent of all hospital admissions (Israëls et al. 2012). Further, let θ_h be the HSMR of hospital h, which is computed by $\theta_h = \sum_{d\in\mathcal{D}}\theta_{hd}$.

The SMR in (10) and HSMR θ_h are estimated using the observed mortality relative to the estimated expected mortality according to (11). The comorbidities are not directly used in Equation (11) as covariates. Instead, they are transformed into 17 binary variables. Each binary variable stands for a group of related diseases according to the classification of the so called Charlson Index (Charlson et al., 1987). This binary variable is 1 when one or more comorbidities are registered for that specific class of the Charlson index and 0 otherwise.

When the number of admissions for a certain class within the patient- or diagnosis-related variables was smaller than 50, classes were merged. This was done to avoid that the standard errors of the regressions became too large. The procedure for merging classes can be found in Van der Laan et al. (2015).

8. References

Backor, K., S. Golde, and N. Nie. 2007. "Estimating Survey Fatigue in Time Use Study." Paper presented at the 29th Annual Conference of the International Association of Time Use Research, 17–19 October 2007, Washington, DC, U.S.A. Available at http://www.atususers.umd.edu/wip2/papers_i2007/Backor.pdf (accessed October 2018).

Bakker, B.F.M. and P.J.H. Daas. 2012. *Methodological Challenges of Register-based Research*. Statistica Neerlandica, 66: 2–7. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00505.x.

Berenschot. 2012. Inventarisatie informatiebehoefte brandweerstatistiek. Eindrapport (in Dutch). Available at https://www.wodc.nl/binaries/2131-volledige-tekst_tcm28-72208.pdf (accessed February 2018).

Bottle, A., B. Jarman, and P. Aylin. 2011. "Hospital Standardized Mortality Ratios: Sensitivity Analyses on the Impact of Coding." *Health Services Research* 46: 1741–1761. Doi: http://dx.doi.org/10.1111/j.1475-6773.2011.01295.x.

Brackstone, G.J. 1987. "Issues in the Use of Administrative Records for Statistical Purposes." *Survey Methodology* 13: 29–43. Available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1987001/article/14467-eng.pdf?st=DEZo9O3B (accessed October 2018).

Charlson, M.E., P. Pompei, K.L. Ales, and R. MacKenzie. 1987. "A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation." *Journal of Chronic Diseased* 40: 373–383. Doi: http://dx.doi.org/10.1016/0021-9681(87)90171-8.

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. "Handbook of Statistical Data Editing and Imputation." New York: John Wiley and Sons.

Efron, B. and C.N. Morris. 1975. "Data Analysis using Stein's Estimator and Its Generalizations." *Journal of the American Statistical Association* 74: 311–319. Doi: http://dx.doi.org/10.1080/01621459.1975.10479864.

Elixhauser, A., C. Steiner, D.R. Harris, and R.M. Coffey. 1998. "Comorbidity Measures for Use with Administrative Data." *Medical Care* 36: 8–27. Doi: http://dx.doi.org/10.1097/00005650-199801000-00004.

Groen, J.A. 2012. "Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures." *Journal of Official Statistics* 28: 173–198. Available at http://www.sverigeisiffror.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/sources-of-error-in-survey-and-administrative-data-the-importance-of-reporting-procedures.pdf (accessed October 2018).

Harteloh, P., K. de Bruin, and J. Kardaun. 2010. "The Reliability of Cause-of-death Coding in The Netherlands." *The European Journal of Epidemiology* 25: 531–538. Doi: http://dx.doi.org/10.1007/s10654-010-9445-5.

Hosmer, D.W. and S. Lemeshow. 2004. Applied Logistic Regression. New York: John Wiley and Sons.

Israëls, A., J. van der Laan, J. van der Akker-Ploemacher, and A. de Bruin. 2012. HSMR 2011: Methodological report. Technical report, Statistics Netherlands. Available at https://www.cbs.nl/NR/rdonlyres/E7EC3032-B244-4566-947D-543B8AAE6E4A/0/2012hsmr2011methoderapport.pdf (accessed February 2018).

Jarman, B. 2008. "In Defence of the Hospital Standardised Mortality Ratio." *Healthcare Papers* 8: 37–41. Doi: http://dx.doi.org/10.12927/hcpap.2008.19974.

Jarman, B., S. Gault, B. Alves, A. Hider, S. Dolan, A. Cook, B. Hurwitz, and L.I. Iezzoni. 1999. "Explaining Differences in English Hospital Death Rates Using Routinely Collected Data." *Biomedicial Journal (BMJ)* 318: 1515–1520. Doi: http://dx.doi.org/10.1136/bmj.318.7197.1515.

Kim, Y., Y-K. Choi, and S. Emery. 2013. "Logistic Regression with Multiple Random Effects: A Simulation Study of Estimation Methods and Statistical Packages." *The American Statistician* 67: 171–182. Doi: http://dx.doi.org/10.1080/00031305.2013.817357.

Oberski, D.L., A. Kirchner, S. Eckman, and F. Kreuter. 2017. "Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models." *Journal of the American Statistical Association*. Available at http://dx.doi.org/10.1080/01621459.2017.1302338 (accessed February 2018).

Pieter, D., R.B. Kool, and G.P. Westert. 2010. Nederlands Tijdschrift voor Geneeskunde, 154, A2186 [in Dutch]. Available at https://www.ntvg.nl/artikelen/beperkte-invloed-gegevensregistratie-op-gestandaardiseerd-ziekenhuissterftecijfer-hsmr/volledig (accessed February 2018).

Pitches, D.W., M.A. Mohammed, and R.J. Lilford. 2007. "What Is the Empirical Evidence That Hospitals with Higher-Risk Adjusted Mortality Rates Provide Poorer Quality Care? A Systematic Review of the Literature." *BMC Health Services Research* 7: 91–98. Doi: http://dx.doi.org/10.1186/1472-6963-7-91.

Prins, M.J. 2016. *The Effect of Coding Practice on the Hospital Standardised Mortality Ratio, Master Thesis*. Utrecht University. (available upon request).

Quan, H., B. Li, L.D. Saunders, G.A. Parsons, C.I. Nilsson, A. Alibhai, and W.A. Ghali. 2008. "Assessing Validity of ICD-9-CM and ICD-10 Administrative Data in Recording Clinical Conditions in a Unique Dually Coded Database." *Health Services Research* 43: 1424–1441. Available at http://onlinelibrary.wiley.com/doi/10.1111/j.1475-6773.2007.00822.x/full (accessed February 2018).

Rousseeuw, P.J. and A.M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: John Wiley and Sons.

Rubin, D.B. 1978. *Multiple Imputations in Sample Surveys – a Phenomenological Bayesian Approach to Nonresponse. Proceedings of the Section on Survey Research Methods*. American Statistical Association. Available at http://ww2.amstat.org/sections/srms/proceedings/papers/1978_004.pdf (accessed February 2018).

Rubin, D.B. 1987. *Multiple Imputation for Non-response in Surveys*. New York: John Wiley and Sons.

Shields, J. and N. To. 2005. "Learning to Say No: Conditioned Underreporting in an Expenditure Survey." Paper presented at the American Association for Public Opinion Research Annual Conference, 12–15 May 2005, Miami Beach, U.S.A. Available at http://ww2.amstat.org/sections/srms/proceedings/y2005/Files/JSM2005-000432.pdf (accessed October 2018).

Silberstein, A.R. and C.A. Jacobs. 1989. Symptoms of Repeated Interview Effects in the Consumer Expenditure Survey. In *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh, 289–303. New York: John Wiley and Sons.

Tourangeau, R., R.M. Groves, and C. Redline. 2010. "Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error." *Public Opinion Quarterly* 74(3): 413–432. Doi: http://dx.doi.org/10.1093/poq/nfq004.

Tourangeau, R. and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133(5): 859–883. Doi: http://dx.doi.org/10.1037/0033-2909.133.5.859.

United Nations Economic Commission for Europe. 2011. *Using Administrative and Secondary Sources for Official Statistics: a Handbook of Principles and Practices*. New York and Geneva: United Nations. Available at http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf (accessed February 2018).

Van Delden, A. and S. Scholtus. 2017. "Correspondence between survey and admin data on quarterly turnover." CBS Discussion Paper 2017-03. Available at https://www.cbs.nl/en-gb/background/2017/07/correspondence-between-survey-and-admin-data-on-quarterly-turnover (accessed February 2018).

Van den Bosch, W.F., J. Silberbusch, K.J. Roozendaal, and C. Wagner. 2010. Variatie in codering Patiëntengegevens beïnvloedt gestandaardiseerd ziekenhuissterftecijfer (HSMR). Nederlands Tijdschrift voor Geneeskunde, 154 A1189 [in Dutch]. Available at https://www.ntvg.nl/artikelen/variatie-codering-patiëntengegevens-beïnvloedt-gestandaardiseerd-ziekenhuissterftecijfer/volledig (accessed February 2018).

Van der Laan, J. 2013. *Quality of the Dutch Medical Registration (LMR) for the calculation of the Hospital Standardised Mortality Ratio*. Discussion Paper. Statistics

Netherlands. Available at https://www.cbs.nl/NR/rdonlyres/6290A0A8-4CC9-4DBF-AF0B-A3C6742EEA89/0/201308x10pub.pdf (accessed February 2018).

Van der Laan, J., A. de Bruin, J. van den Akker-Ploemacher, C. Penning, and F. Pijpers. 2015. *HSMR 2014: Methodological Report. Technical Report*. Statistics Netherlands. Available at https://www.cbs.nl/NR/rdonlyres/2AFF4E96-C02F-4BB0-97A9-035D17DF1104/0/2015hsmrmethodologicalreport2014.pdf (accessed February 2018).

Wallgren, A. and B. Wallgren. 2014. *Register-based Statistics. Statistical Methods for Administrative Data* (2nd edition). New York: John Wiley and Sons.

West, B.T. and A.G. Blom. 2017. "Explaining Interviewing Effects: A Research Synthesis." *Journal of Survey Statistics and Methodology* 5: 175–211. Doi: http://dx.doi.org/10.1093/jssam/smw024.

Received June 2017
Revised April 2018
Accepted May 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 889–908, http://dx.doi.org/10.2478/JOS-2018-0044

Population Size Estimation and Linkage Errors: the Multiple Lists Case

Loredana Di Consiglio[1] *and Tiziana Tuoto*[1]

Data integration is now common practice in official statistics and involves an increasing number of sources. When using multiple sources, an objective is to assess the unknown size of the population. To this aim, capture-recapture methods are applied. Standard capture-recapture methods are based on a number of strong assumptions, including the absence of errors in the integration procedures. However, in particular when the integrated sources were not originally collected for statistical purposes, this assumption is unlikely and linkage errors (false links and missing links) may occur. In this article, the problem of adjusting population estimates in the presence of linkage errors in multiple lists is tackled; under homogeneous linkage error probabilities assumption, a solution is proposed in a realistic and practical scenario of multiple lists linkage procedure.

Key words: Probabilistically linked data; capture-recapture model; multiple system estimation; log-linear model.

1. Introduction

The integration and combination of external sources with traditional statistical survey data is a pressing challenge for National Statistical Institutes. Micro-level integration of different sources is standard practice, generally performed by means of record linkage techniques. However, the linkage process is not completely error-free and statisticians must take linkage errors into account in subsequent analyses performed on integrated data (Chambers 2009). Linkage errors appear particularly relevant when the goal is to measure the size of a population (partially) enumerated in different lists, as shown in Di Consiglio and Tuoto (2015). A widespread method for population size estimation in the presence of two lists is the capture-recapture model (see Petersen 1896; Lincoln 1930; Pollock et al. 1990; Wolter 1986).

The capture-recapture method is subject to the following assumptions:

1. Perfect matching among lists,
2. Independence of lists,
3. Homogeneity of capture probabilities,
4. Closure of population, and
5. No out-of-scope units in the lists.

[1] Italian National Institute of Statistics (Istat), via Balbo 16, 00184 Roma, Italy. Emails: diconsig@istat.it and tuoto@istat.it

When more than two lists are considered, say k, the observations from multiple captures can be organized into a 2^k table, with the presence/absence on the ith list defining the category for the ith dimension. The cell count corresponding to no capture for all the k lists is unknown. Therefore, the goal of estimating the number of units in the population corresponds to the estimation of the unknown count of the missing cell in the 2^k incomplete contingency table.

Several procedures using log-linear models have been proposed (Fienberg 1972; Cormack 1989). When more than two lists are considered, the use of log-linear models enables the independence assumption to be weakened, even if higher order interactions are still subject to restrictions due to model identification. The original log-linear models proposed in Fienberg (1972) rely on the other assumptions: perfect linkage, homogeneity of capture probabilities, closed population, absence of over-coverage. Extensions to the basic log-linear models are provided. Just to mention a few examples, Cormack (1989) discusses the use of log-linear models for dependence and the detection of the presence of heterogeneity in capture probabilities; Darroch et al. (1993) and Agresti (1994) introduce models in the generalised class of Rasch models to explain the heterogeneity in capture probabilities; Coull and Agresti (1999) introduce generalised mixture models. Evans et al. (1994) suggest applying log-linear models when the heterogeneity effects can be explained by the observable covariates. IWGDMF (1995) reviews these approaches, see Chao (2001) for an overview. Zwane and van der Heijden (2005) propose conditional multinomial logit models allowing the inclusion of covariates in the models; Bartolucci and Forcina (2006) introduce latent class models that can be viewed as an extension of conditional multinomial logit models. These models permit accounting for both the observed heterogeneity using covariates and the unobserved heterogeneity, by assuming units to belong to distinct latent classes. Finally, a Bayesian approach can be found in Farcomeni and Tardella (2009).

When more than two lists are available, Di Cecco et al. (2017) discuss the use of a generalisation of the Latent Class models that can be expressed as log-linear models with a latent variable to deal with the problem of out-of-scope units.

Few contributions (Ding and Fienberg 1994, Lee et al. 2001; Di Consiglio and Tuoto 2015) have addressed the issue of matching errors in the population size estimation with two lists. This article explores adjustments for linkage errors in population size estimators, when $k > 2$ lists are considered. Extending the previous works of Di Consiglio and Tuoto (2015) and Fienberg and Ding (1996), this article takes into account both erroneous links and missing links in a realistic linkage error generation model.

The article is organised as follows: Section 2 briefly describes the linkage model and the errors when more than two lists are integrated, Section 3 presents the effects of linkage errors on the observed 2^k incomplete contingency table, as well as a formulation that relates the observed table with the true one, via the linkage errors. In Section 4, the procedure to estimate the population size using the log-linear model is defined. Section 5 discusses the definition of linkage errors used in this framework and reviews a few proposals for their estimation. In Section 6, the application of the proposed method is illustrated in the context of census and administrative data, whereas simulated data are used to analyse its statistical performance and to carry out a sensitivity analysis on the misspecification of linkage errors. Finally, Section 7 provides some concluding remarks and open issues to be tackled by future research.

2. Multiple Lists and Record Linkage

Record linkage is the activity of recognising the same real word entity, even if differently represented in the several data sources. When a common unique identifier is not available, the record linkage techniques exploit common attributes, potentially affected by errors and missing values, to identify the same unit. Therefore, at the end of a linkage procedure, records referring to the same real world entity may emerge unlinked (false negative). In a similar way, false matches may occur when the integration procedure links a pair of units that do not actually relate to the same real-world entity (false positive).

To exemplify, let us consider the two-list case, as in the seminal article of Fellegi and Sunter (1969), say L_1 and L_2, of size N_1 and N_2. Let $\Omega = \{(a, b), a \in L_1 \text{ and } b \in L_2\}$ be the Cartesian product of all possible pairs, of size $|\Omega| = N_1 \times N_2$. The record linkage between L_1 and L_2 is viewed as a classification problem, where the pairs in Ω have to be assigned to two subsets M and U, independent and mutually exclusive, such that:

M is the link set $(a = b)$

U is the non-link set $(a \neq b)$.

Common identifiers (linking variables) are chosen and, for each pair, a comparison vector, denoted by γ, is obtained. Let r be the ratio between the conditional probability of γ given that the pair belongs to the set M and the conditional probability of γ given that the pair belongs to the set U. The ratio r is the likelihood ratio test statistic for testing the null hypothesis H_0: $(a, b) \in M$ against the alternative hypothesis H_1:$(a, b) \in U$, that is

$$r = \frac{P(\gamma|(a,b) \in M)}{P(\gamma|(a,b) \in U)} = \frac{m(\gamma)}{u(\gamma)}. \tag{1}$$

The pairs for which r is greater than an upper threshold value T_m are assigned to the set of linked pairs, M^*; the pairs for which r is smaller than a lower threshold value T_u are assigned to the set of unlinked pairs, U^*; if r falls in the range (T_u, T_m), a no-decision is made automatically and the pair is classified by a clerical review.

The previous thresholds are chosen to minimise the false link probability, denoted by β, and the false non-link probability, denoted by $1 - \alpha$, which are defined as follows:

$$\beta = \sum_{\gamma \in \Gamma} u(\gamma) P(M^*|\gamma) = \sum_{\gamma \in \Gamma_{M^*}} u(\gamma) \quad \text{where} \quad \Gamma_{M^*} = \{\gamma : T_m \leq m(\gamma)/u(\gamma)\} \tag{2}$$

$$1 - \alpha = \sum_{\gamma \in \Gamma} m(\gamma) P(U^*|\gamma) = \sum_{\gamma \in \Gamma_{U^*}} m(\gamma) \quad \text{where} \quad \Gamma_{U^*} = \{\gamma : T_u \geq m(\gamma)/u(\gamma)\}. \tag{3}$$

In applications, the probabilities m and u can be estimated by treating the true link status as a latent variable, and using the EM algorithm (Jaro 1989). Alternatively, Larsen (1996) applies a Bayesian latent class and Bayesian log-linear models to fit the mixture models (Larsen and Rubin 2001).

When more than two lists have to be linked, for instance, multiple administrative data sets, there are different ways to proceed. Indeed, the standard record-linkage methodologies in use at National Statistical Institutes deal mainly with pairs of lists.

Some proposals for simultaneously linking more than two lists are given by Sadinle et al. (2011); Sadinle and Fienberg (2013); Steorts et al. (2014); Ventura et al. (2014), and Fienberg and Manrique-Vallier (2009). However, currently these methods still need to be "industrialised", so they are not yet suitable for applications in the official statistics production systems due to their computational complexity (Fienberg 2015).

Alternatively, one can match all lists in pairs. A drawback of pairwise linkages is the risk of discrepancies in the linkage decisions. For instance, considering three lists, one can link the record of the individual a in list 1 and the record of an individual b in list 2 from a bipartite record linkage. Then, from a second bipartite record linkage, one links the record of b to the record of an individual c in list 3. Based on these two linkages, one might conclude that a, b, and c are the same individual. However, one also links the first and third lists, but the records a and c may emerge unmatched. If the records a, b, and c truly correspond to the same individual (entity), a nonmatch may occur due to measurement error or incomplete record information. On the other hand, if the records of a, b, and c do not refer to the same individual, we have four possibilities: a and b refer to the same individual but c refers to another one, a and c refer to the same individual but b refers to another one, b and c refer to the same individual but a refers to another one, or a, b, and c all refer to different individuals. By using bipartite record linkage for each pair of files, one cannot resolve the matching pattern. While there are various *ad hoc* approaches to resolve the results of multiple bipartite matchings, no formal methodology has appeared in the statistical literature (Herzog et al. 2007).

To solve multiple linkage, a widespread practice in the National Statistical Institutes is to consider a list as a master frame, and then to link each list sequentially into the master frame. In this case, the linkage procedure involving three lists consists of linking firstly list 1 and list 2, and then the resulting frame with list 3. This procedure has the advantages of needing only two linking operations, while the corresponding pairwise links involve three linkage operations; in addition it does not require solving potential discrepancies.

In the following, we consider the latter multiple-list linkage scenario. In the next session, we describe the linkage errors generated by these linkage operations and how they affect the capture-recapture model.

3. Capture-Recapture Model and Transition Matrix

3.1. Capture-Recapture Model

To focus on the effect of linkage errors in the multiple-capture framework, we consider the case of three captures (lists). In the absence of linkage errors, the capture-recapture data can be classified in the following incomplete 2^3 table (Fienberg 1972):

where n_{ijk} is the cell count of the presence/absence in the lists, with $i,j,k = 1,0$. Let π_{ijk} denote the corresponding cell probability. The table is incomplete, due to the fact that the count n_{000} is unobservable.

The linkage errors modify the counts in Table 1 in two ways: the number of observations may increase in some cells and decrease in others; and the total number of different individuals observed in the three lists may change, provided that the total number of observations in each list, $n_{1++}n_{+1+}$, n_{++1}, remains unchanged.

Table 1. True table for cell counts, without linkage errors.

	List 1			
	Present		*Absent*	
	List 3		List 3	
List 2	*Present*	*Absent*	*Present*	*Absent*
Present	n_{111}	n_{110}	n_{011}	n_{010}
Absent	n_{101}	n_{100}	n_{001}	n_{000}

Table 2 reports the observed counts, subject to linkage errors:
where $n^* = \{n^*_{ijk},\ i,j,k = 1,0\}$ denotes the observed counts after the linkage. Let $\pi^* = \{\pi^*_{ijk}, i,j,k = 1,0\}$ denote the corresponding probabilities. Finally, let $n^*_{\cup L}$ *be the sum of observed distinct units.*

3.2. Error Model with Missing and False Links

Fienberg and Ding (1996) propose a correction of the log-linear model that considers the possible transitions from the true configuration n to the observed one n^*, taking into account only the missing links. They assume that: (i) there are no erroneous matches in the linkage process; (ii) a transition can only go downwards by at most one level, and (iii) the probability of remaining at the original state (no missing error) equals α and the probability of a transition to any of the possible states is equal to $(1 - \alpha)/(m - 1)$, where m is the number of all possible states to which transitions are possible and allowed. For example, an individual truly recorded in all the three lists (111) can produce the following patterns {(110), (001)} or {(101), (010)} or {(110), (001)} with equal probability $(1 - \alpha)/3$.

In this article, we suppose that the transition probabilities are related to both the probability of missing a true match and the probability of a false link. Moreover, we apply a more realistic error model that mimics more closely a real three-list linkage process as described in Section 2, that is, we first assume a linkage step of list 1 and 2 and then a linkage to list 3, taking into account different linkage errors in the two linkage steps.

To this purpose, let $1 - \alpha_1$ be the probability of missing a match in the first linkage and $1 - \alpha_2$ be the probability of missing a match in the second linkage; moreover let β_1 be the probability of a false link in step 1 and let β_2 be the probability of a false link in step 2.

Table 2. Observed table for cell counts.

	List 1			
	Present		*Absent*	
	List 3		List 3	
List 2	*Present*	*Absent*	*Present*	*Absent*
Present	n^*_{111}	n^*_{110}	n^*_{011}	n^*_{010}
Absent	n^*_{101}	n^*_{100}	n^*_{001}	n^*_{000}

We study the effect of linkage errors, introducing first the probability of missing a true match. However, differently from Fienberg and Ding (1996), we aim at taking into account the realistic linkage process in two phases. Then, if one only considers the probability of missing a match, the possible alternative "decompositions" generated by a real unit observed in all the three lists (111), counted in n_{111}, result in the observed ones $(ijk)*$ counted in multiple cells, n^*_{ijk}, as follows:

a. $(111)^*$ with probability $\alpha_1\alpha_2$,
b. $(110)^*$ and $(001)^*$ with probability $\alpha_1(1-\alpha_2)$
c. $(101)^*$ and $(010)^*$ with probability $\frac{(1-\alpha_1)(\alpha_2)}{2}$
d. $(011)^*$ and $(100)^*$ with probability $\frac{(1-\alpha_1)(\alpha_2)}{2}$ and finally
e. $(100)^*$ and $(010)^*$ and $(001)^*$ with probability $(1-\alpha_1)(1-\alpha_2)$.

The five events above are complementary and mutually exclusive.

On the second line, for example, when we correctly link the first two lists but we miss the link with the third one, the event b results in the "decomposition" of (111) in $(110)*$ and $(001)*$ with probability $\alpha_1(1-\alpha_2)$. On the contrary, when an error occurs at the first linkage step, the individual is decomposed in two different units, then the third list is correctly linked to either the first or the second one with the same probability $\frac{(\alpha_2)}{2}$ (event c or d).

For convenience, following the terminology of Fienberg and Ding (1996), we call such a decomposition (or combination in case of false matches, discussed below) a "transition".

A similar reasoning for the decomposition of the other true individual patterns allows for a transition matrix M_1 to be obtained. Table 3 reports the matrix with the transition probabilities resulting from the different events that generate the observed patterns after linkage. For instance, the $(001)*$ is generated from (111) when either the event b or the event e of the above example occur. The probability of the transition from (111) to $(001)*$ is then $\alpha_1(1-\alpha_2)+(1-\alpha_1)(1-\alpha_2)=1-\alpha_2$, as in Table 3.

It is worth noting that the columns of the transition matrix M_1 do not necessarily add up to one. The probabilities of the alternative events (missingness/unmissingness of matches in one/two steps) obviously add up to one. However, when a linkage error occurs (e.g., a true match is missed) it affects more than one row of the matrix, generating decomposition/combination of the true unit of the population. This property is consistent with the observation that the sum of the distinct individuals enlisted in Table 1 differs from the sum of the observed distinct units in Table 2.

Table 3. Transition matrix M_1 from real to observed cells when only missing links occur.

	111	110	101	100	011	010	001
$(111)*$	$\alpha_1\alpha_2$	–	–	–	–	–	–
$(110)*$	$\alpha_1(1-\alpha_2)$	α_1	–	–	–	–	–
$(101)*$	$\frac{(1-\alpha_1)(\alpha_2)}{2}$	–	α_2	–	–	–	–
$(100)*$	$\frac{(1-\alpha_1)(2-\alpha_2)}{2}$	$1-\alpha_1$	$1-\alpha_2$	1	–	–	–
$(011)*$	$\frac{(1-\alpha_1)(\alpha_2)}{2}$	–	–	–	α_2	–	–
$(010)*$	$\frac{(1-\alpha_1)(2-\alpha_2)}{2}$	$1-\alpha_1$	–	–	$1-\alpha_2$	1	–
$(001)*$	$1-\alpha_2$	–	$1-\alpha_2$	–	$1-\alpha_2$	–	1

The transition matrix M_1 can be further extended to include the false linkage errors. As before, different linkage errors are assumed for the first and the second phase. In addition, we assume that whenever a true match is missed, the related records cannot be involved in false matches in the same phase, because this event happens when at least two errors occur: the records are incorrectly linked and the correct match is missed. Then, we assume it has a negligible probability of occurrence, as in Ding and Fienberg (1994) and in Di Consiglio and Tuoto (2015). Under the above assumptions, and, at the same time, treating the transitions caused by false and missing linkage errors, we obtain the transition matrix M_2 in Table 4. It is worth noting that the matrix M_2 can contain negative values due to algebra on the probabilities of composition/decomposition generated by the false links.

4. Estimation of Population Size

The true counts in Table 1 can be estimated by a linear combination of the observed counts via the inverse of the transition matrix:

$$n = M^{-1} n^* \tag{4}$$

The transition matrix M can be either M_1 or M_2 (see Tables 3 or 4) according to the adopted error model. Similarly, the cell probabilities can be estimated by $\pi = M^{-1}\pi^*$.

Once the true cell counts are obtained by (4), in order to estimate the population size N, one needs to estimate the unknown count of the missing cell in the 2^k incomplete contingency table, for example applying a suitable log-linear model. For instance, when dealing with three lists, one can use the log-linear saturated model

$$\log\left(E(n_{ijk})\right) = \lambda + \lambda_i^{L_1} + \lambda_j^{L_2} + \lambda_k^{L_3} + \lambda_{ij}^{L_1 L_2} + \lambda_{ik}^{L_1 L_3} + \lambda_{jk}^{L_2 L_3} \tag{5}$$

where the sum of any λ over any subscript is zero. The fitted count $\tilde{n}_{000}$ from the log-linear model is finally used to estimate the population size:

$$\tilde{N} = n + \tilde{n}_{000}. \tag{6}$$

Under the assumption of independence of each pair of lists, we have

$$\tilde{n}_{000} = \frac{n_{111} n_{001} n_{100} n_{010}}{n_{101} n_{011} n_{101}} \tag{7}$$

The assumption of independence of each pair of lists is equivalent to setting $\lambda^{L_u L_v} = 0$ for each u and v. The use of log-linear model, however, enables list pair dependency and its extensions to also take account of the heterogeneity of capture probabilities (see Section 1).

In general, to obtain an estimation of the population size N, we first compute the Maximum Likelihood (ML) estimates of the parameters from the conditional likelihood associated with observed cell count n^* given $n^*_{\cup L}$, as suggested in Fienberg and Ding (1996). Sanathanan (1972) shows that, under suitable regularity conditions, the conditional maximum likelihood estimates and the unconditional ones are both consistent and have the same asymptotic normal distribution. Once the conditional maximum likelihood estimates of the log-linear model are obtained, we use the log-linear model specified for the not-observed real values to compute the conditional maximum likelihood

Table 4. Transition matrix M_2 from real to observed cells with missing and false links.

	111	110	101	100	011	010	001
(111)*	$\alpha_1\alpha_2$	$\alpha_1\beta_2$	$\alpha_2\beta_1$	$\beta_1\beta_2$	$\alpha_2\beta_1$	–	–
(110)*	$\alpha_1(1-\alpha_2)$	$(\alpha_1)(1-\beta_2)$	$\beta_1(1-\alpha_2)$	$(\beta_1)(1-\beta_2)$	$\beta_1(1-\alpha_2)$	–	–
(101)*	$\frac{(1-\alpha_1)(\alpha_2)}{2}$	$(1-\alpha_1)\beta_2/2$	$(1-\beta_1)(\alpha_2)$	$(1-\beta_1)(\beta_2)$	–	–	–
(100)*	$\frac{(1-\alpha_1)(2-\alpha_2)}{2}$	$(1-\alpha_1)\left(1-\frac{\beta_2}{2}\right)$	$(1-\beta_1)(1-\alpha_2)$	$(1-\beta_1)(1-\beta_2)$	$-\beta_1$	–	–
(011)*	$\frac{(1-\alpha_1)(\alpha_2)}{2}$	$(1-\alpha_1)\beta_2/2$	–	–	$(1-\beta_1)\alpha_2$	–	–
(010)*	$\frac{(1-\alpha_1)(2-\alpha_2)}{2}$	$(1-\alpha_1)\left(1-\frac{\beta_2}{2}\right)$	$-\beta_1$	$-\beta_1$	$(1-\beta_1)(1-\alpha_2)$	1	–
(001)*	$1-\alpha_2$	$-\beta_2$	$1-\alpha_2$	$-\beta_2$	$1-\alpha_2$	–	1

estimates of the expected cell counts $\tilde{n}_{ijk}$, including the one of the missing cell. Thus, $\tilde{N} = \sum_{ijk} \tilde{n}_{ijk}$.

5. Focus on Linkage Errors

The linkage errors defined in Formulas (2) and (3) are based on the Fellegi and Sunter (1969) theory for record linkage that is very effective for the link identification. Note that, conceptually, in (2) and (3) the probabilities β and α are defined for each element of the product space $\Omega = L_1 \times L_2$. However, as it is well known in practice, the Fellegi and Sunter (1969) linkage procedure is not reliable for estimating the linkage errors. Tuoto (2016) proposes a supervised learning method to predict both types of linkage errors, without relying on strong distribution assumptions, as in Belin and Rubin (1995). Alternatively, Chipperfield and Chambers (2015) apply a bootstrap method to the actual linkage procedure to evaluate the mismatch probabilities.

On the other hand, in the population size estimation context, it may be necessary to adopt alternative definitions of the linkage errors than (2) and (3). For instance, let us consider the multiple capture counts in Table 2 and the two linkage steps that produced it. At any of the linkage stages, if the true linkage status was known, the errors rates could be defined comparing the links made with the true ones. At the first stage, the results of this comparison could be reported as in Table 5.

Then to assess the quality of the linkage process, the following ratios could be defined:

$$\text{False nonmatched (missed match) rate: } 1 - \alpha = \frac{c}{a+c} = \frac{n_{11} \cap n_{11}^*}{n_{11}}; \tag{8}$$

$$\text{False match rate: } \beta = \frac{b}{b+d} = \frac{n_{11}^* - n_{11} \cap n_{11}^*}{(N_1 - n_{11}) + (N_2 - n_{11})}. \tag{9}$$

Clearly, the definition of false match error β in (9) is more pragmatic than in (2), because the set of all the unlinked pairs $U = (N_1 - n_{11}) \times (N_2 - n_{11})$ is a much larger set than $(N_1 - n_{11}) + (N_2 - n_{11})$, since the false matches in (9) are related to the unlinked cases of both the lists, rather than to the cross-product of the lists, as in (2), where the unlinked pairs set U is considered. Moreover, it is worth noting that one can expect the number of false links involving the actually linked records to be much lower than the number of false links between unlinked records, because the former implies two linkage errors simultaneously, that is, missing the true match and erroneously linking the matched record to a different record.

Finally, it should be pointed out that the false match rate defined by (9) is a different quantity to the false match rate used for adjusting regression analysis (e.g., in Chambers, 2009), where the latter is defined in relation to the number of actual links n_{11}^*. While both

Table 5. Comparison of true matches and assigned links.

	True matches	True non-matches
Links	a – true positives	b – false positives or false links
No links	c – false negatives or missing links	d – true negatives

quantities target the same number of false links among the links made, the two rates are not the same measure, because they have different denominators.

6. Applications

In this section, we present some applications of multiple capture estimation method in the presence of linkage errors. Firstly, in Subsection 6.1, the adjusted estimator derived applying transformation (4) with M_2 is applied in a real-life context, the census, post-enumeration survey and administrative data example already considered by Fienberg and Ding (1996). In Subsection 6.2, we propose a simulation study to analyse the empirical statistical properties of the suggested estimators; in Subsection 6.3, the simulation study provides a sensitivity analysis to show the robustness of the population size estimates with respect to the linkage error evaluation.

6.1. Example from Census, PES and Administrative Data

First, let us consider the data from the three lists previously used by Fienberg and Ding (1996): the 1990 U.S. Census, the corresponding post-census survey (PES), and the administrative list supplement (ALS). Data for sampling strata PES 11 at St. Louis are given in Table 6.

For the evaluation of the matching errors, Fienberg and Ding (1996) use the Matching Error study (see Mulry et al. 1989) to assess both the probability of missing a link in the linking procedure between the Census and the PES, and the probability of missing a link in the linkage involving the ALS, under the assumption of no errors in the rematch. The results of the Matching Error Study for 1990 U.S. Census in St. Louis stratum are reported in Table 7 (see Table 4, 562 in Fienberg and Ding, 1996).

Ignoring the unresolved cases, Fienberg and Ding (1996) estimate the probability of missing a true link as $(1 - \hat{\alpha}_1) = (1 - \hat{\alpha}_2) = 9/(2{,}667 + 9) = 0.3363\%$. Following the same reasoning, we evaluate the probability of a false link as $\hat{\beta}_1 = \hat{\beta}_2 = 7/(7 + 427) = 1.6129\%$. It is worth noting that the false linkage error is much greater than the missing linkage error, suggesting the need to correct also for false links.

For the estimation of the unknown size of the population, Fienberg and Ding (1996) examine various log-linear models with different dependency structures in order to better fit the data in Table 6. The model [CensusPes][PesALS] results to fit the data

Table 6. Three-sample data for stratum 11, St. Louis, 1990 U.S. Census.

	Census = List 1			
	Present		Absent	
	PES = List 2		PES = List 2	
ALS = List 3	Present	Absent	Present	Absent
Present	300	51	53	180
Absent	187	166	76	–

Table 7. St. Louis rematch study.

Original match classification	Rematch classification			
	Matched	Not matched	Unresolved	Total
Matched	2,667	7	8	2,682
Not matched	9	427	30	466
Unresolved	0	7	20	27
Total	2,676	441	58	3,175

much better. The corresponding naïve estimate is $\hat{N} = 1,599$. Applying their correction for missing links, Fienberg and Ding (1996) estimate $\tilde{N}_{DF} = 1,585$. Instead, including the false linkage errors as well, with the error matrix M_2 specified in Table 4, we get $\tilde{N}_{MDF} = 1,680$. This value is within the confidence interval of both of the previous estimates.

6.2. *Results on Simulated Data*

This section describes the results of a simulation on fictitious data. To simulate the linkage process in a realistic way, we use person identifiers from the fictitious population census data (McLeod et al. 2011) created for the ESSnet DI, which was a European project on data integration (Record Linkage, Statistical Matching, Micro integration Processing) running from 2009 to 2011.

The ESSnet DI provides three entirely fictitious data sources, which are supposed to have captured details of persons at the same reference time. The first data set consists of observations from the Patient Register Data of the National Health Service (PRD, in the following); the second data set contains observations from the Customer Information System (CIS), which combines administrative data from the tax and social security systems; the third data set reports observations from a decennial Census (CEN). In these data sets, which comprise over 26,000 records each, linking variables (names, dates of birth, addresses) for individual identification may be distorted by missing values and typos, to imitate real-life situations. These synthetic data reproduce the real data and the actual observed errors that make the linkage procedure difficult. For details on the generation of synthetic data and the perturbation of the key variables, see McLeod et al. (2011). The simulation setting lets us know the true match status to benchmark the linkage results. In the simulation, 500 populations of the size 1,000 were generated, sampling the data independently and randomly without replacement.

For each replicate, the three lists were randomly drawn by the PRD, CIS and CEN on the basis of the following capture probabilities: $\pi_{1++} = 0.65$, $\pi_{+1+} = 0.53$ and $\pi_{++1} = 0.57$, respectively.

At each replicate, the linkage was made as illustrated in Section 2: in the first phase, the PRD and CIS lists were linked; in the second phase, the linked and un-linked records of the first phase were linked to the third list (CEN).

Table 8. Distribution of the linkage error rates over the 500 replicates.

Linkage errors%	Min	Median	Mean	Max
First step				
$1\text{-}\alpha_1$	0.00	2.39	2.51	7.33
β_1	0.00	4.58	4.31	7.63
Second step				
$1 - \alpha_2$	0.90	2.86	2.91	5.93
β_2	0.20	4.53	4.10	8.56

In both steps, the linkage variables were Name, Surname, Day, Month and Year of Birth, and the probabilistic record linkage model (Fellegi and Sunter 1969, Jaro 1989) was implemented by the batch version of the software RELAIS (RELAIS, 2015).

Table 8 summarises the results of the linkage procedure in terms of realised linkage error rates, reporting the probability of missing a true match $1\text{-}\alpha$ and the probability of a false match β for both steps, as defined in Section 5 (see Formulas 8 and 9 for step 1). The realised $1\text{-}\alpha$ and β can be evaluated in light of the known true linkage status.

At each replicate, we compute the naïve log-linear estimator and the adjusted estimators, applying the transformation (4) with M_1 or M_2 as described in Section 4. Having generated the three lists independently, the log-linear model assumes the independency of the lists. The adjusted estimator was computed using the true values of the probability of nonmissing true matches α and the probability of false match β obtained in each replicate. The use of the true values of α and β allows us to compare the estimators without the effect of the linkage error estimation, hence focusing on the performance of the adjusted estimator.

Figure 1 shows the distributions over the 500 replicates of the several estimators: the naïve estimator, the adjusted estimator taking account of missing links only (DF as Ding and Fienberg) according to the matrix M_1 and the adjusted estimator taking account of the two types of linkage errors (MDF, modified DF) according to the matrix M_2 in Table 4. For comparison, the figure shows the estimates that can be obtained with the true counts unaffected by linkage errors.

The relative percentage errors of the estimators are summarised in Table 9. The table shows the minimum value, the first quartile, the median, the average, the third quartile

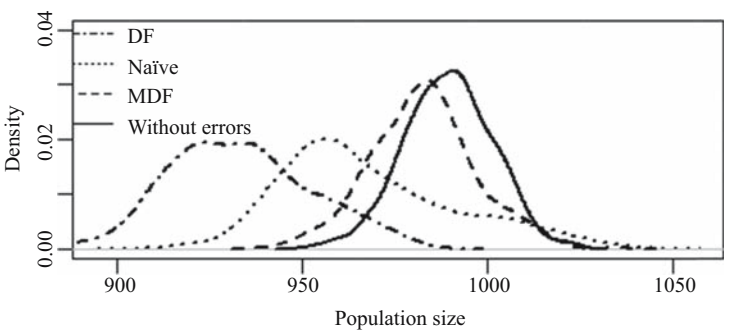

Fig. 1. Empirical density of the alternative estimates of the population size over the replicates (true N = 1,000).

Table 9. Distribution of percentage relative error.

	Percentage relative error					
Estimator	Min	Q1	Median	Mean	Q3	Max
Naïve	− 8.70	− 4.90	− 3.70	− 3.19	− 1.70	3.90
DF	− 11.70	− 8.11	− 6.70	− 6.67	− 5.40	− 1.60
MDF	− 5.90	− 2.70	− 1.75	− 1.74	− 0.90	3.50
True values	− 4.80	− 1.82	− 1.00	− 1.05	− 0.20	2.30

and the maximum value of the relative percentage of error calculated over the 500 replicates.

The results in Figure 1 and Table 9 show that the proposed adjustment reduces the bias of the naïve estimator without side effects on the variability of the estimator, even if the bias is not entirely removed due to the non-linear nature of the population size estimator. Likewise, the residual bias may be due to the misspecification of the linkage error model: it is observed in this simulation, as well as in other real applications (Tuoto et al. 2017) that the probability of double errors (i.e., missing a true link and false link of the records at the same time) may be not negligible, as assumed in the proposed transition matrix M_2.

6.3. A Sensitivity Analysis

The simulation setting is exploited for a sensitivity analysis of the proposed estimator with respect to the misspecification of the linkage errors. In the previous subsection, the MDF estimator was calculated under optimal conditions, that is, knowing the values of the linkage errors made. In this section, several values of α_1, α_2, β_1 and β_2 are tested to evaluate the statistical properties of the MDF estimator in different nonoptimal scenarios. First, we apply the MDF estimator with the four average linkage errors over the 500 replicates – we denote the estimator as MDF_{mmmm} in the following. Moreover, the variability of the linkage errors is accounted for in MDF estimates by evaluating the matrix M_2 with several combinations of the lower and upper bounds of the confidence intervals over the 500 replicates. We denote $\mathbf{MDF}_{\boldsymbol{\alpha_1\beta_1\alpha_2\beta_2}}$ where the subscripts take values in {o, m, l, u}, standing for "observed", "mean", "lower bound of the confidence interval", "upper bound of the confidence interval" respectively.

Figure 2 compares the true values, the naïve estimates and the adjusted estimators $\text{MDF}_{\alpha_1\beta_1\alpha_2\beta_2}$.

As expected, the MDF estimator with true observed linkage errors outperforms the MDF estimators with different values (m, l, u) of the linkage errors, both in terms of bias and variability. However, when we compare the naïve estimator and the MDF estimators with inaccurate values of the linkage errors, the results are diverse. Figure 2 shows that the MDF estimate still improves the naïve one, at a cost of a slight increase in variability, when using the linkage error averages. As expected, when using the lower bound of the confidence intervals of the errors, the MDF estimates tend to the naïve one. On the contrary, when applying the upper bound of the confidence intervals (i.e., on average

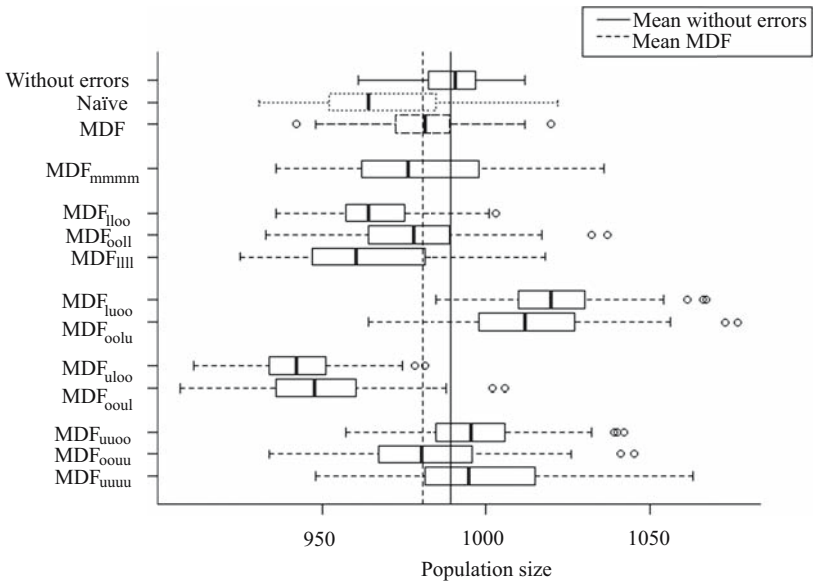

Fig. 2. Simulated alternative estimates of the population size (true N = 1,000) with different values of linkage errors.

applying an over-correction), there is a tendency to overestimate. Finally, the MDF correction is ineffective when the missing linkage errors are overestimated and the false linkage errors are underestimated, or viceversa.

This analysis also shows that the adjustment with an inaccurate evaluation of the second step linkage errors causes an increase in the variability but produces less bias in the estimates compared to the bias caused by an inaccurate evaluation of the first step errors, that is, once the linkage errors at the first step are misspecified we cannot adjust only with the second step error probabilities. On the other hand, this sensitivity analysis indicates that the independence assumption on linkage errors may not hold, as anticipated at the end of the previous section: in fact, the MDF_{uuoo} and MDF_{uuuu} are on average closer to the true values than MDF estimates.

7. Discussion and Concluding Remarks

This article proposes an extension of the Fienberg and Ding (1996) approach in order to take account of linkage errors in the evaluation of the population size when more than two lists are considered in a multiple system estimation framework.

However, the proposed estimator presents some open issues that need further investigation, partially inherited from the general context of multiple captures. Some reflections are briefly discussed in the next subsections.

7.1. A Note on Variance Estimation

In the estimation of the population size, it is assumed that the counts are distributed according to a log-linear model; using the delta method, Darroch (1958) derives an estimator of the variance of the population size estimator. For instance, when the three lists

are independent, the estimator proposed by Darroch (1958) is as follows

$$\widehat{Var}(\tilde{N}) = \tilde{N}\tilde{n}_{000}\left(\sum_{\{ijk\}\in S}\tilde{n}_{ijk}\right)^{-1} \tag{10}$$

where S contains all cells corresponding to individuals caught more than once.

However, in our context, a straightforward application of Formula (10) on the estimated counts would omit the additional source of variability introduced by the linkage errors process. In fact, when encountering linkage errors, the observations are subject to the multinomial process generating the true captures plus the additional probabilistic process of linking the lists. Then the variance evaluation needs to consider this additional probabilistic process generating the linkage errors. Simply replacing the counts in Formula (10) with their estimates obtained via the transformation (4) would not take into account the latter source of variation. Moreover, in practice, the linkage errors are not known and their estimation will introduce an additional source of error that should be considered.

As an alternative to analytical variance analysis, one can explore a bootstrap approach. The variance estimation of the adjusted estimator is an open issue for future research.

7.2. *Scalability*

This article explicitly evaluates a general adjustment for linkage errors when the population size is based on three sources. The method is readily applicable to the multiple list case; however, a generalisation to $k > 3$ lists requires the evaluation of the transition matrix M and the knowledge of the multiple step linkage mechanism. Considering only the missing link error α, the transition matrix for $k = 5$ is implemented in Link et al. (2010) – see below in Subsection 7.3 for more details. Obviously, when the false link errors are introduced into the analysis, the evaluation of the transition matrix is not straightforward.

It is worth noting that the trade-off between the risk of potential linkage errors and the advantages of increasing the number of lists for the population size estimation should be further investigated by means of case studies.

7.3. *A Bayesian Perspective on the Population Size Estimation*

Alternative approaches to record linkage are based on Bayesian methods. For instance, in Fortini et al. (2001) and Liseo and Tancredi (2011), the interest is focused on a matrix-valued parameter $\boldsymbol{C}$, which represents the true pattern of matches between the two lists. The sum of the elements of $\boldsymbol{C}$ is an estimate of the number of true matches between the two lists, given the following constraints on the parameter space of C that avoid multiple matches:

$$C_{ij} = \{0, 1\}, \qquad \sum_{L_1} C_{ij} \leq 1, \qquad \sum_{L_2} C_{ij} \leq 1.$$

The Bayesian approach enables the propagation of the uncertainty of the linkage process to subsequent analysis of the linkage data in a natural way. According to the knowledge of the authors, this method is only described in the two-list case, but similarly to the Fellegi-Sunter approach, it could be applied by incremental steps that consider an

augmented number of lists. A practical difficulty with the Bayesian approach is the lack of scalability to large data sets, which is the case of the population size estimation in official statistics.

Steorts et al. (2015) propose an alternative Bayesian approach that allows linking records from multiple lists simultaneously while de-duplicating the lists. Similarly to Parag and Domingos (2004), the linkage is considered as a process of recognising latent "entities" with a graphical representation, that is, each record in the lists can be linked to a latent unit from 1 to $N_{\max}$, where $N_{\max}$ is the total number of units in all the lists, if no unit is present in more than one. A uniform prior is assumed on the linkage structure, that is, any observed unit is equally assigned to any of the latent individual. A hybrid MCMC algorithm is used to improve the computing performances. However, Steorts et al. (2015) do not utilise their model for the estimation of the population size in the presence of undercoverage. Further research is needed to apply their method in such setting.

Finally, we mention the linkage errors adjustment proposed by Link et al. (2010). They assume only missing matches and no erroneous links; they model the capture-recapture history with a vector where the components are indicators of:

(i) Presence in the given capture and correct identification of the individual,
(ii) Presence in the given capture but missing identification, and
(iii) Absence in the given capture.

However, this model is still subject to the specification of a matrix M. They define the recorded frequency vector n^* as a linear combination of true history n, which is considered as a latent variable. So, the application of the method still requires the actual specification of the M matrix that connects the observed values to the true one, similarly to what is described in this article.

7.4. Concluding Remarks

To summarise, this article first defines a realistic and widely used linkage setting for multiple sources, then the errors caused by both missing and erroneous links are included in the contingency table of the presence/absence of the units in the various sources. The originality of the proposal consists in adjusting for false matches in addition to missing matches, extending the previous works of Fienberg and Ding (1996) and Link et al. (2010). Indeed, the false matches are frequent, as well as missing matches; this fact is also observed in the Matching Error Study (Mulry et al. 1989) on the linkage between 1990 U.S. Census and PES, which is used to apply the proposed adjustment.

The suggested estimator allows reducing the bias of the naïve estimator without relevant effects on variability, even if the bias is not entirely cancelled out due to the nonlinear nature of the estimator. It is worth recalling the assumptions underlying the estimator (6): a. the linkage procedure acts in sequential steps, for instance, two steps in the description of the three-list case provided in Subsection 3.2; b. linkage errors are independent in different steps; c. at each step, the probability of missing a true match and erroneously linking the related records in false matches is negligible, as in Fienberg and Ding (1996); d. the linkage errors are either known or accurately estimated; and e. the linkage errors are homogeneous, at least in sub-groups.

The independence assumption b should be verified as, linkage errors are caused by errors in the matching variables, one can, given the occurrence of these errors, assume that linkage errors in different step are independent. However, the linkage mechanism can be such that if a link is missed (or a false link is introduced) in the first step, this may increase the probability of a linkage error in the second step. In our simulation setup, we tested the adjustment with known linkage errors, evaluating them by means of the known actual matches. The sensitivity analysis shows that the adjusted estimator outperforms the naïve one in several cases, even if the linkage errors are unknown. However, when the missing linkage errors are overestimated and the false linkage errors are underestimated, and viceversa, both at the first and the second step, the MDF correction is ineffective. The simulation and the sensitivity analyses are restricted to one population framework (i.e., Census and administrative data) and one linkage scenario. Other applications or simulation settings can provide further insights and prove the generalisability of the observed results. Moreover, the evaluation of linkage errors and the effect of these errors on the variability of the population size estimates are still open issues.

The proposed estimator is developed assuming constant linkage errors across the entire population. This may not always hold in practice; in those cases, the adjustment can still be applied considering strata in which homogeneous linkage errors occur. As linkage errors depend on errors in the key variables, then homogeneous groups can be built on the basis of them. The gain of the adjusted estimator in the presence of homogeneous strata compared to the use of average values of the errors over the entire population could be examined; this is an aspect to be tackled in future research. However, the sensitivity analysis already provides the insight that the adjustment can still be valuable compared to the naïve estimator, even with error values not corresponding to the true ones.

Finally, additional case studies should be carried out to analyse the statistical properties of the suggested adjustment when considering extensions to basic log-linear models.

8. References

Agresti, A. 1994. "Simple Capture-Recapture Models Permitting Unequal Catchability and Variable Sampling Effort." *Biometrics* 50: 494–500. Doi: http://dx.doi.org/10.2307/2533391.

Bartolucci, F. and A. Forcina. 2006. "A Class of Latent Marginal Models for Capture-Recapture Data with Continuous Covariates." *Journal of the American Statistical Association* 101: 786–794. Doi: http://dx.doi.org/10.1198/073500105000000243.

Belin, T.R. and D.B. Rubin. 1995. "A Method for Calibrating False-Match Rates in Record Linkage." *Journal of the American Statistical Association* 90: 694–707. Doi: http://dx.doi.org/10.1080/01621459.1995.10476563.

Chambers, R. 2009. "Regression Analysis of Probability-Linked Data." *Official Statistics Research Series* 4. Available at http://www3.stats.govt.nz/statisphere/Official_Statistics_Research_Series/Regression_Analysis_of_Probability-Linked_Data.pdf (accessed November 2018).

Chao, A. 2001. "An overview of closed Capture-Recapture Models." *Journal of Agricultural, Biological, and Environmental Statistics* 6: 158–175. Doi: http://dx.doi.org/10.1198/108571101750524670.

Chipperfield, J. and R. Chambers. 2015. "Using the Bootstrap to Account for Linkage Errors when Analysing Probabilistically Linked Categorical Data." *Journal of Official Statistics* 31(3): 397–414. Doi: http://dx.doi.org/10.1515/jos-2015-0024.

Cormack, R.M. 1989. "Log-Linear Models for Capture-Recapture." *Biometrics* 45: 395–413. Doi: http://dx.doi.org/10.2307/2531485.

Coull, B.A. and A. Agresti. 1999. "The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-Recapture Studies." *Biometrics* 55: 294–301. Doi: http://dx.doi.org/10.1111/j.0006-341X.1999.00294.x.

Darroch, J.N. 1958. "The Multiple-Recapture Census: I. Estimation of a closed population." *Biometrika* 45: 343–359. Doi: http://dx.doi.org/10.2307/2333183.

Darroch, J.N., S.E. Fienberg, G.F.V. Glonek, and B.W. Junker. 1993. "A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability." *Journal of the American Statistical Association* 88: 1137–1148. Doi: http://dx.doi.org/10.2307/2290811.

Di Cecco, D., M. Di Zio, D. Filipponi, and I. Rocchetti. 2016. "Estimating Population Size from Multisource Data with Coverage and Unit Errors." In Proceeding of the ICES-V, Geneva, Switzerland, June 20–23, 2016. Available at http://ww2.amstat.org/meetings/ices/2016/proceedings/165_ices15Final00072.pdf (accessed November 2018).

Di Consiglio, L. and T. Tuoto. 2015. "Coverage Evaluation on Probabilistically Linked Data." *Journal of Official Statistics* 31(3): 415–429. Doi: http://dx.doi.org/10.1515/JOS-2015-0025.

Ding, Y. and S.E. Fienberg. 1994. "Dual System Estimation of Census Undercount in the Presence of Matching Error." *Survey Methodology* 20: 149–158. Available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1994002/article/14422-eng.pdf?st=YtHflfaV (accessed November 2018).

Evans, M.A., D.G. Bonett, and L.L. McDonald. 1994. "A General Theory for Modeling Capture-Recapture Data from a Closed Population." *Biometrics* 50(2): 396–405. Doi: http://dx.doi.org/10.2307/2533383.

Farcomeni, A. and L. Tardella. 2009. "Reference Bayesian Methods for Recapture Models with Heterogeneity." *Test, May* 2010, 19(1): 187–208. Doi: http://dx.doi.org/10.1007/s11749-009-0147-9.

Fellegi, I. and A. Sunter. 1969. "A Theory of Record Linkage." *Journal of the American Statistical Association* 64: 1183–2010. Doi: http://dx.doi.org/10.1080/01621459.1969.10501049.

Fienberg, S.E. 1972. "The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables." *Biometrika* 59: 409–439. Doi: http://dx.doi.org/10.1093/biomet/59.3.591.

Fienberg, S.E. 2015. "Discussion." *Journal of Official Statistics* 31(3): 527–535. Doi: http://dx.doi.org/10.1515/JOS-2015-0032.

Fienberg, S.E. and Y. Ding. 1996. "Multiple Sample Estimation of Population and Census Undercount in the Presence of Matching Error." In Proceedings of 1994 Annual research conference and CASIC technologies Interchange, Bureau of Census, United States. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1996001/article/14385-eng.pdf?st=8LhKz2Tt (accessed November 2018).

Fienberg, S.E. and D. Manrique-Vallier. 2009. "Integrated Methodology for Multiple Systems Estimation and Record Linkage Using a Missing Data Formulation." *Advances in Statistical Analysis* 93: 49–60. Doi: http://dx.doi.org/10.1007/s10182-008-0084-z.

Fortini, M., B. Liseo, A. Nuccitelli, and M. Scanu. 2001. "On Bayesian Record Linkage." *Research in Official Statistics* 4(1): 185–198.

Herzog, T., F. Scheuren, and W. Winkler. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer-Verlag. Doi: http://dx.doi.org/10.1007/0-387-69505-2.

IWGDMF – International Working Group for Disease Monitoring and Forecasting. 1995. "Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development." *American Journal of Epidemiology* 142: 1047–1058. Doi: http://dx.doi.org/10.1093/oxfordjournals.aje.a117558.

Jaro, M. 1989. "Advances in Record Linkage Methodology as Applied to Matching the 1985 Test Census of Tampa, Florida." *Journal of American Statistical Association* 84: 414–420. Doi: http://dx.doi.org/10.1080/01621459.1989.10478785.

Larsen, M.D. 1996. *Bayesian Approaches to Finite Mixture Models*, Ph.D. Thesis, Harvard University.

Larsen, M.D. and D.B. Rubin. 2001. "Iterative Automated Record Linkage Using Mixture Models." *Journal of the American Statistical Association* 96: 32–41. Doi: http://dx.doi.org/10.1198/016214501750332956.

Lee, A.J., G.A.F. Seber, J.K. Holden, and J.T. Huakau. 2001. "Capture-Recapture, Epidemiology, and List Mismatches: Several Lists." *Biometrics* 57: 707–713. Doi: http://dx.doi.org/10.1111/j.0006-341X.2001.00707.x.

Lincoln, F.C. 1930. *Calculating Waterfowl Abundance on the Basis of Banding Returns*. United States Department of Agriculture Circular, 118, 1–4.

Link, W.A., J. Yoshizaki, L.L. Bailey, and K.H. Pollok. 2010. "Uncovering a Latent Multinomial: Analysis of Mark-Recapture Data with Misidentification." *Biometrics* 66: 178–185. Doi: http://dx.doi.org/10.1111/j.1541-0420.2009.01244.x.

Liseo, B. and A. Tancredi. 2011. "Bayesian Estimation of Population Size Via Linkage of Multivariate Normal Data Sets." *Journal of Official Statistics* 27(3): 491–505. Available at: https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/bayesian-estimation-of-population-size-via-linkage-of-multivariate-normal-data-sets.pdf (accessed November 2018).

McLeod, P., D. Heasman, and I. Forbes. 2011. Simulated data for the on the job training. Essnet DI. Available at http://www.cros-portal.eu/content/job-training.

Mulry, M.H., A. Dajani, and P. Biemer. 1989. "The Matching Error Study for the 1988 Dress Rehearsal." In Proceedings of the Section on Survey Research Methods, ASA, 704–709. Available for instance at researchgate: https://www.researchgate.net/publication/267379153_THE_MATCHING_ERROR_STUDY_FOR_THE_1988_DRESS_REHEARSAL/download.

Parag and P. Domingos. 2004. "Multi-Relational Record Linkage." In Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining. Available at: https://homes.cs.washington.edu/~pedrod/papers/mrdm04.pdf (accessed November 2018).

Petersen, C.G.J. 1896. *The Yearly Immigration of Young Plaice into the Limfiord from the German Sea*. Report of the Danish Biological Station 6: 5–84.

Pollock, K.H., J.D. Nichols, C. Brownie, and J.E. Hines. 1990. "Statistical Inference for Capture-Recapture Experiments." *Wildlife monographs* 107.

RELAIS. 2015. *User's Guide Version 3.0*. Available at http://www.istat.it/en/tools/methods-and-it-tools/processing-tools/relais.

Sadinle, M., R. Hall, and S.E. Fienberg. 2011. "Approaches to Multiple Record Linkage." In Proceedings of the ISI World Statistical Congress, 21–26 August 2011, Dublin: 1064–1071. Available at: http://2011.isiproceedings.org/papers/450092.pdf (accessed November 2018).

Sadinle, M. and S.E. Fienberg. 2013. "A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems." *Journal of the American Statistical Association* 108: 385–397. Doi: http://dx.doi.org/10.1080/01621459.2012.757231.

Sanathanan, L. 1972. "Estimating the Size of a Multinomial Population." *Annals of Mathematical Statistics* 43: 142–152. Available at: https://projecteuclid.org/download/pdf_1/euclid.aoms/1177692709 (accessed November 2018).

Steorts, R., R. Hall, and S.E. Fienberg. 2014. "SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication." *Journal of Machine Learning Research* 33: 922–930. Available at: http://proceedings.mlr.press/v33/steorts14.pdf (accessed November 2018).

Steorts, R., R. Hall, and S.E. Fienberg. 2015. "A Bayesian Approach to Graphical Record Linkage and De-duplication." *Journal of the American Statistical Association*. Available at: URL http://arxiv.org/abs/1312.4645.

Tuoto, T. 2016. "New Proposal for Linkage Error Estimation." *Statistical Journal of the IAOS* 32(2): 413–420. Doi: http://dx.doi.org/10.3233/SJI-160995.

Tuoto, T., B.F.M. Bakker, L. Di Consiglio, D.J. van der Laan, P.-P. de Wolf, and D. Zult. 2017. "Two Improvements of the Method for Population Size Estimation." in Proceedings of the 61st World Statistics Congress 16–21 July 2017, Marrakech.

Ventura, S. and R. Nugent. 2014. "Hierarchical Clustering with Distributions of Distances for Large-Scale Record Linkage." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer, 283–298. Berlin: Springer Link. Lecture Notes in Computer Science 8744.

Wolter, K.M. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81: 338–346. Doi: http://dx.doi.org/10.1080/01621459.1986.10478277.

Zwane, E. and P.G.M. van der Heijden. 2005. "Population Estimation using the Multiple System Estimator in the Presence of Continuous Covariates." *Statistical Modelling* 5: 39–52. Doi: http://dx.doi.org/10.1191/1471082X05st086oa.

Received June 2017
Revised April 2018
Accepted August 2018

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 909–933, http://dx.doi.org/10.2478/JOS-2018-0045

Statistical Matching as a Supplement to Record Linkage: A Valuable Method to Tackle Nonconsent Bias?

Jonathan Gessendorfer[1], *Jonas Beste*[2], *Jörg Drechsler*[2], *and Joseph W. Sakshaug*[2]

Record linkage has become an important tool for increasing research opportunities in the social sciences. Surveys that perform record linkage to administrative records are often required to obtain informed consent from respondents prior to linkage. A major concern is that nonconsent could introduce biases in analyses based on the linked data. One straightforward strategy to overcome the missing data problem created by nonconsent is to match nonconsenters with statistically similar units in the target administrative database. To assess the effectiveness of statistical matching in this context, we use data from two German panel surveys that have been linked to an administrative database of the German Federal Employment Agency. We evaluate the statistical matching procedure under various artificial nonconsent scenarios and show that the method can be effective in reducing nonconsent biases in marginal distributions, but that biases in multivariate estimates can sometimes be worsened. We discuss the implications of these findings for survey practice and elaborate on some of the practical challenges of implementing the statistical matching procedure in the context of linkage nonconsent. The developed simulation design can act as a roadmap for other statistical agencies considering the proposed approach for their data.

Key words: Data fusion; survey data; administrative data; linkage nonconsent.

1. Introduction

Many survey organizations link their surveys to large-scale administrative databases in order to increase research opportunities, minimize data collection costs, and enhance data utility (Calderwood and Lessof 2009). To give only a few examples, the Avon Longitudinal Study of Parents and Children (Ness 2004) and the UK Millennium Cohort Study (Mostafa 2016) link interview data to various health and social administrative records. Statistics Netherlands conducts linkages of surveys and various administrative registers to conduct the Dutch census (Schulte Nordholt et al. 2014). In Germany, several

[1] Email: jonathan.gessendorfer@gmail.com
[2] Institute for Employment Research, Regensburger Str. 100, 90478 Nuremberg, Germany. Emails: jonas.beste@iab.de, joerg.drechsler@iab.de, and joe.sakshaug@iab.de

Acknowledgments: We thank the anonymous reviewers, the associate editor and the guest editor, whose insightful comments and suggestions helped improve the manuscript considerably. This article uses data from the National Educational Panel Study (NEPS): Starting Cohort Adults, doi:10.5157/NEPS:SC6:7.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. This study also uses the factually anonymous data of the Panel Study 'Labour Market and Social Security' (PASS) and the factually anonymous Sample of Integrated Labour Market Biographies (SIAB). For both, data access was provided via a Scientific Use File supplied by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).

surveys, including the study "Working and Learning in a Changing World" (ALWA; Antoni and Seth (2011)) and the Socio-Economic Panel (SOEP) Migration Sample (Brücker et al. 2014), link to the Integrated Employment Biographies (IEB) – an administrative database of the German Federal Employment Agency (BA) that covers nearly the entire German population of employable age (Jacobebbinghaus and Seth 2010).

Due to data protection regulations, surveys in many countries are required to obtain informed consent from respondents prior to record linkage. In the European Union, for example, this requirement is part of the General Data Protection Regulation (GDPR 2016). Nonconsent and other reasons for record linkage failure lead to incomplete data and therefore a reduction in statistical power and precision of statistical estimates. Reviews of the linkage consent literature show that the amount of incomplete data can be quite severe with linkage consent rates below 50% in several studies (Sakshaug and Kreuter, 2012; da Silva et al. 2012). Even more alarming is the fact that linkage consent rates have been declining over time (Fulton 2012). Given this declining trend and low observed consent rates, there is increasing concern that nonconsenters could be systematically different from consenters, introducing bias in subsequent analyses based on the linked data. Numerous studies have demonstrated the biasing effects of nonconsent in actual linkage applications (Jenkins et al. 2006; Sakshaug and Huber 2016; Sakshaug et al. 2012; Sala et al. 2012; Mostafa 2016). Some of the most common variables affected by nonconsent bias include socio-demographics (for example, age, sex, ethnicity), economic variables (for example, income, income assistance benefits), and socio-environmental variables (for example, urbanicity, regional variation). While the majority of such biases have been found in survey variables, biases in the linked administrative variables have also been identified (Sakshaug and Kreuter 2012; Sakshaug and Vicari 2017; Sakshaug et al. 2017), suggesting that neither data source is immune to nonconsent bias.

Nonconsent generates a very specific missing data situation. In many ways, it is similar to the situation created by unit nonresponse if auxiliary information is available for both respondents and nonrespondents. However, an important difference is that the amount of information available for both consenters and nonconsenters — the data obtained from the survey — typically far exceeds the amount of information available for both respondents and nonrespondents. It is not obvious whether best practice methods developed to reduce nonresponse biases (Brick and Kalton 1996) would perform similarly for the missing data situation generated by nonconsent to record linkage. While extensive research has been done on optimizing linkage consent rates at the design stage – for example, by improving the wording or placement of the consent question in the survey (Kreuter et al. 2016) – no general guidelines have been proposed to reduce linkage nonconsent bias post-survey data collection.

One strategy to overcome the missing data problem induced by linkage nonconsent is to use statistical matching (Rässler 2002; D'Orazio et al. 2006b). Statistical matching methods merge individual records from two (or more) data sources based on their similarity on variables observed in all data sources. The main goal of the research presented here is to investigate the idea of performing statistical matching on the nonconsenting cases in order to 1) make the administrative data available for all survey participants, including linkage nonconsenters; and 2) reduce linkage nonconsent biases in estimates derived from the linked survey and administrative data. We evaluate this

strategy through a case study involving two major surveys in Germany that link to the IEB database: the "National Educational Panel Study" (NEPS) and the Panel "Labour Market and Social Security" (PASS).

The remainder of this article is organized as follows. In Section 2 we review record linkage and statistical matching as tools for combining information from different sources. In Section 3 we illustrate how statistical matching may be used as a supplement to record linkage for nonconsent bias reduction. Section 4 discusses problems in practice including why extensions to the classical statistical matching approaches, although promising from a methodological perspective, cannot be used in this context. In Section 5 we describe the two surveys and the administrative data source used to evaluate the proposed methodology.

In Section 6 we describe the study design and evaluation procedures. The results of the evaluation are presented in Section 7. The article concludes with a discussion of the case study results, their implications for survey practice, and practical issues associated with implementing the proposed methodology.

2. Record Linkage and Statistical Matching

To facilitate our review of record linkage and statistical matching we introduce the following notation. Let A and B be two data sets to be merged where vectors of random variables (X, Y) are observed in data set A and vectors of random variables (X, Z) are observed in data set B. For brevity, we limit our discussion to the most common scenario of merging two data sets. The goal of both record linkage and statistical matching is to use the information from both data sets in order to estimate the joint density $f(x, y, z)$ of the combined vector (X, Y, Z) in the population.

2.1. Record Linkage and Reasons for Unsuccessful Linkage

Record linkage techniques aim to identify and merge records of different data sources that refer to the same entity (Herzog et al. 2007). In the context of survey and administrative data linkage, the goal is to identify administrative data records that belong to the same survey respondents. Assuming the survey respondents represent a random subset of the population and linkage is successful for every respondent, the merged vectors consist of random realizations of (X, Y, Z) and thus inference regarding $f(x, y, z)$ is straightforward.

However, there are many reasons why record linkage may be unsuccessful for some survey respondents. One of those reasons is that record linkage techniques sometimes fail to identify true links. This happens particularly if only imperfect linkage identifiers – such as name and address information, instead of unique identifiers like national identification numbers – are used to merge both data sets. Fellegi and Sunter (1969) provide a mathematical framework for this situation. Imperfect identifiers can produce nonlinks or false positive links – merging of records that do not belong to the same unit – which can lead to attenuated associations between Y and Z. Another reason for unsuccessful record linkage is that some survey respondents might not have records in the administrative data set. If individuals with specific traits are missing systematically from the administrative data, this can also lead to biased inferences.

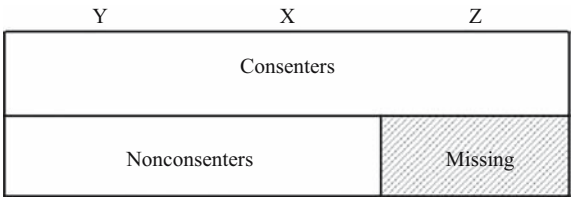

Fig. 1. The missing data situation in the combined data set.

A further reason for unsuccessful linkage – and the focus of this article – is due to the fact that data protection regulations often require that informed consent be obtained from survey respondents prior to data linkage. This creates a missing data situation in the combined data set as depicted in Figure 1. As noted in the introduction, contemporary research shows that linkage nonconsent is prevalent in surveys and can introduce bias in linked survey and administrative variables if only the completely observed parts of the data are used for analyses. Thus, methods to mitigate nonconsent bias are needed to obtain valid inferences from linked data sets.

Several methods have been developed and evaluated that deal with a very similar problem: unit nonresponse bias in surveys. The two most prominent methods are weighting adjustments and (multiple) imputation. Both are applicable to the linkage nonconsent scenario but have significant drawbacks in this context. The main drawback of weighting for linkage nonconsent is that analyses can only be performed on the consenting cases. That is, after constructing the weights, the survey information for all nonconsenters is completely ignored, making the approach inefficient especially for high nonconsent rates.

If the goal is to obtain a complete, rectangular data set on (X, Y, Z), multiple imputation can be considered to fill in the missing linked data. However, most imputation routines need good parametric models for the missing Z variables. The modeling step can become highly complex in the context of merging survey and administrative data, as the structure of administrative data is often not suitable for parametric specification. For example, administrative variables from the IEB database are measured in terms of spells with varying beginning and endpoints (Jacobebbinghaus and Seth 2010). Creating good parametric models for such variables is a very difficult and labor-intensive task.

Nearest neighbor hot-deck imputation is a possible alternative to parametric imputation (Chen and Shao, 2000; Andridge and Little 2010). This method may be used to identify a consenting respondent who is similar in (X, Y) to a nonconsenting respondent.

The consenting respondent then donates the observed Z to the observed X and Y data of the nonconsenting respondent. However, the feasibility of nearest neighbor hot-deck imputation depends heavily on the consent rate and on the sample size of the survey, since hot-deck runs into problems if the donor pool is sparse (Andridge and Little 2010).

2.2. *Review of Statistical Matching*

Statistical matching, sometimes known as data fusion, aims to integrate multiple data sources to draw inference on $f(X, Y, Z)$. Micro approaches to statistical matching create a synthetic data set, where X, Y and Z are available as if they were jointly observed, whereas

macro approaches attempt to draw inference on parameters that are nonestimable using only the separate data sets. We only consider micro approaches here, as macro approaches are less suited for the application we are considering in this article. For an overview of macro approaches, see D'Orazio et al. (2006b). This reference is also recommended for further information on the micro approaches we discuss below.

In standard statistical matching applications, data sets A and B are both random samples drawn from a much larger population. In this scenario, record linkage would be infeasible, as there is unlikely to be any overlap between the two data sources. Traditional approaches to statistical matching use a set of common variables X to combine A and B. For example, nearest neighbor matching techniques merge data of units that are similar in X. Specifically, for each unit in A – the recipients – a unit in B that is similar in X donates its Z information to the observed (X, Y) vector of the recipient (see Figure 2 for a visualization). Besides nearest neighbor, there are various other traditional statistical matching techniques. Beyond nonparametric methods, like nearest neighbor, fully parametric models or mixtures of parametric models and nonparametric matching techniques have been suggested in the literature (Rässler 2002; D'Orazio et al. 2006b).

Given that only the information in X is used in all traditional matching procedures, the distribution of (X, Y, Z) after statistical matching $\tilde{f}(x, y, z)$ will necessarily have a very specific characteristic: conditional on X, Z and Y will be independent:

$$\tilde{f}(y|x,z) = \tilde{f}(y|x) \quad \wedge \quad \tilde{f}(z|x,y) = \tilde{f}(z|x) \tag{1}$$

Therefore, if the aim is to draw inference regarding the relationship of Y and Z, one must implicitly assume that the two variables are independent conditional on X in the population. This is referred to as the conditional independence assumption (CIA). If the assumption is not met, the joint distribution of (X, Y, Z) after statistical matching will differ from the true distribution. Potentially, this can lead to biased inferences from the statistically matched data set (Sims 1972; Rodgers 1984; D'Orazio et al. 2006b). For example, correlations between Y and Z variables will typically be biased towards zero, as only the part of the correlation that can be explained by the X variables will be preserved in the statistically matched file. Similar to the effect of omitted variables, regression coefficients in models using Y and Z variables can either be over- or underestimated. Another consequence of the CIA is that statistically matched files are only suited for analyses of unconditional associations between Y and Z and associations conditional on only a subset of all possible confounding variables, that is, on only a subset of X variables. By design, Z and Y will be independent conditional on all X variables in the matched file.

Lacking additional information on $f(X, Y, Z)$, one approach to avoid the assumption of conditional independence is to perform sensitivity analyses (consider among others Kadane (1978); Moriarity and Scheuren (2001); D'Orazio et al. (2006a); D'Orazio et al.

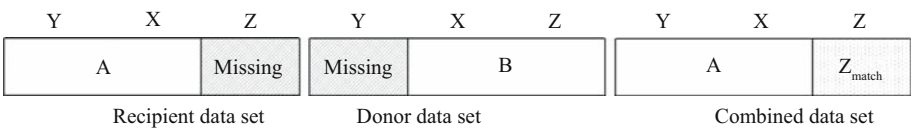

Fig. 2. Goal of micro approaches to statistical matching.

(2009); Conti et al. (2012, 2016); Rubin (1986); Rässler (2002, 2003); Rässler and Kiesl (2009)). These approaches typically utilize logical constraints to reduce uncertainty, for example, on the unknown correlation of Y and Z, ρ_{YZ}. In this case, the constraints follow from the necessity of (X, Y, Z)'s correlation matrix to be positive semidefinite and the fact that, apart from ρ_{YZ}, the correlation matrix can be estimated from A and B alone. Depending on the strength of the correlation between X and Y, and X and Z, the range of possible ρ_{YZ}s can be very small or not restricted at all.

An alternative approach for avoiding the assumption of conditional independence is to make use of additional available information. Singh et al. (1993), for example, provides a nonparametric micro approach (based on ideas in Paass (1985)) that can utilize auxiliary information in the form of a data set C in which X, Y and Z are jointly observed by first finding a nearest neighbor with respect to (X,Y) for each unit from A in C and donating their Z information to obtain (X, Y, Z_C). In a second step, Z_C is replaced by Z_B by finding a nearest neighbor with respect to (X, Z) in B. Other approaches include Bayesian methods, parametric, nonparametric and mixed approaches (for example, Kadane (1978); Paass (1985); Rässler (2003); Moriarity and Scheuren (2001, 2003); Filippello et al. (2004); Gilula et al. (2006); Gilula and McCulloch (2013); Fosdick et al. (2016)). Some utilize information on parameters regarding the distribution of Y and Z, others use C to estimate the conditional distribution of Y and Z given X. Again, we refer to D'Orazio et al. (2006b) for an overview.

3. Statistical Matching as a Supplement to Record Linkage

The goal of using statistical matching as a supplement to record linkage is to handle the missing data situation explained in Subsection 2.1 and depicted again using the statistical matching notation in the left-most panel of Figure 3. For consenting units, record linkage is performed in the usual way, while statistical matching is performed for all units that did not provide linkage consent. Note that in the statistical matching literature, A always denotes the data recipients, B denotes the donors, and C denotes the auxiliary data set in which all variables are jointly observed. Thus, to be consistent with this notation, A only comprises the survey data of the nonconsenters, B is still the donor data set, and C contains the combined data of the consenters in our context.

Besides conditional independence, there is another implicit assumption if traditional statistical matching is to be used as a supplement to record linkage. This assumption (described below) is necessary because the missing data situation generated by nonconsent to record linkage is different to the situation that statistical matching techniques are designed for. In the standard statistical matching scenario, the units in data set A and B are disjoint and X, Y, and Z are never jointly observed (D'Orazio et al. 2006b). The missingness of Z in A and of Y in B is therefore missing by design. If A and B are both

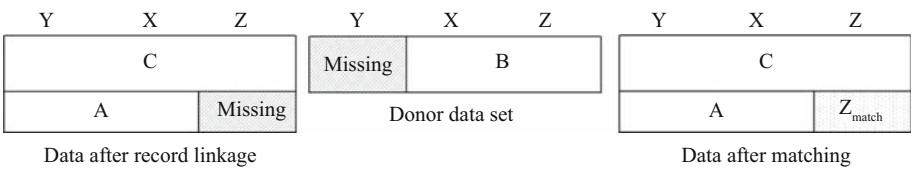

Fig. 3. Statistical matching as a supplement to record linkage

independent random samples of the population, the missing information in both files is missing completely at random (MCAR; Rubin (1976)).

The missingness generated by nonconsent to record linkage, on the other hand, is likely to not be MCAR. This implies that nonconsenters are not a random sample from the population. The partially observed realizations of the nonconsenters are random vectors of $f(x, y, z|\text{nonconsent})$ (with Z unobserved) which is not necessarily identical to $f(x, y, z)$. However, statistical matching does assume that both data sets A and B are random samples from the same distribution. If this is not true, it suffices to assume (at least for traditional statistical matching) that the conditional distribution of Z given X is the same in A and B (D'Orazio et al. 2006b). In our situation – assuming no selectivity in B – this translates to:

$$f(z|x, \text{nonconsent}) = f(z|x) \tag{2}$$

The distribution of Z conditional on X of the nonconsenters $f(z|x,\text{nonconsent})$ has to be the same as in B. This essentially means that one must assume the Z information for all nonconsenters is missing at random given X (MAR; Rubin (1976)). However, it is important to keep in mind that for Equation 2 to hold it is not sufficient that the probability to consent only depends on X. The selection mechanism for the complete process that leads to the final data of the consenters must only depend on X, that is, any selectivity introduced at the sampling stage, or because of nonconsent, must be fully explainable by X. This assumption can be seen as critical if only a small number of variables exist in X.

4. Problems in Practice

For statistical matching to be successful, the variables contained in X need to be measured similarly across the data sources (D'Orazio et al. 2006b; Meinfelder 2013). The main idea of statistical matching is to utilize the common variables X, and structural differences in the measurement of X in A, the recipients, and B, the donor data set, can therefore be highly problematic, for example, if the matching variables are measured with different levels of precision in the two data sources. Most importantly, the measurements should be free from bias, or, if bias exists, both sources need to be affected similarly. For example, with traditional statistical matching techniques, if X is biased differently in the recipient data set than in the donor data set, then the imputation will be based on the wrong value of X.

The assumptions regarding the bias behavior are especially problematic in the context of matching survey data with administrative records. Surveys are prone to measurement error since interviewers, question wording, memory of respondents, and various other factors can have effects on both accuracy and precision – bias and variance – of the measurement (Biemer et al. 2011). While administrative data can have different measurement problems (Oberski et al. 2017), the errors on the survey side alone can have detrimental effects on statistical matching even if the measurement in the administrative data is perfectly accurate and precise. For this reason, it is essential to identify all potential measurement differences in the two files and adjust the matching procedure accordingly.

Besides these very general remarks that apply to all statistical matching procedures, there are difficulties specific to only a subset of the available statistical matching techniques. Without going into extensive detail, we note that good parametric models are necessary to express the relationship between Z and X for all traditional statistical

matching methods, with the exception of nonparametric techniques like nearest neighbor. Similar to parametric imputation, parametric statistical matching is thus infeasible in the context of highly complex administrative data structures (see also our discussion at the end of Subsection 2.1).

Statistical matching techniques that utilize logical constraints are almost never used in practical statistical matching applications (Meinfelder 2013). The main reason is that they are only feasible if the number of variables within each of the vectors X, Y, and Z is relatively small. In addition, some methods make assumptions regarding the distribution of X, Y, and Z – the most prominent being multivariate normality. Given the complexity of variables X, Y and Z used in the context of merging survey and complex administrative data sets, with bounds, skip patterns, and logical constraints between the variables, such approaches are not feasible in applications similar to our setting. Besides, in many surveys most of the variables are discrete in nature or are measured on a discrete scale. Thus, the assumption of multivariate normality in particular, is often unrealistic. Furthermore, the uncertainty evaluation becomes much more complex if MCAR does not hold. The uncertainty is then a combination of the uncertainty of the missingness model and of the model parameter uncertainty (D'Orazio et al. 2006b).

In the supplement to record linkage scenario, there is auxiliary information available in the form of the successfully linked data of all consenting survey respondents. This could potentially be used as an auxiliary data set, C, for which X, Y and Z are jointly observed. Excluding parametric techniques for the same reason as above, to our knowledge the only nonparametric method proposed in the literature for incorporating C is the method by Singh et al. (1993) explained in Subsection 2.2. However, in settings like ours, it is very similar and offers essentially no benefit compared to nearest neighbor hot-deck imputation (cf. Subsection 2.1), which is essentially the first step of the method. When merging survey and administrative data, the donor pool is the complete population, which typically means that we will be able to find donors that match (almost) exactly on all the variables in X and Z_C. In this case, the second step of Singh et al. (1993) will not lead to any improvements, since Z_B will be equal to Z_C for all units. Therefore, the true donor pool will remain to be the records contained in C and the large pool in B cannot be utilized.

Given that methods that quantify the uncertainty from matching and methods that use auxiliary information cannot be exploited for our application for the reasons given above, we focus on traditional nearest neighbor techniques for the remainder of this article. Nearest neighbor methods are especially attractive in our case as they are nonparametric and thus are unaffected by the complexity of Z in administrative data sets.

We note that nearest neighbor hot-deck imputation (as explained in Subsection 2.1) has some similarities with statistical matching. The major difference is, data sets A and C are matched using both X and Y as matching variables instead of data sets A and B using only X. This means that with imputation we would not need to assume conditional independence. In addition, the missing at random assumption would be weakened to:

$$f(z|x, y, \text{nonconsent}) = f(z|x, y, \text{consent}) \tag{3}$$

However, as stated above, hot-deck methods are heavily dependent on the size of the donor pool and the donor-to-recipient ratio. While statistical matching can utilize the vast donor pool in the administrative data set B, imputation can only use the donors in C. This

means that if statistical matching can be used beneficially, it is more generally applicable than nearest neighbor hot-deck imputation, as it is independent of the consent rate and the sample size.

We conclude this section by noting that we do not believe that the conditional independence assumption and the missing-at-random assumption will ever be fully met in practice. However, we know that if we only use those cases that consented to the linkage of the data sources, we generally need to assume *consenting completely at random* if we want to get unbiased results. Arguably, this is also a rather strong assumption. Thus, the empirical question to answer is: would we be better off using only the data of the consenters, or could statistical matching be used to reduce the bias from assuming *consenting completely at random*? We do not expect to get completely unbiased results through statistical matching, but if the impacts of violating the statistical matching assumptions are minor, we might still be able to improve over the results based on using only the data of the consenters.

This reasoning is the motivation for the simulation studies described in the next sections.

5. Data Sources Used in the Evaluation Study

To evaluate whether statistical matching can be a viable supplement to record linkage, we use two large (and independent) panel surveys in Germany: the National Educational Panel Study and the Panel Study "Labour Market and Social Security". Both are linked to individual administrative process data from the German Federal Employment Agency. In our application, this administrative data set – the Integrated Employment Biographies – is used as the donor file B. The recipient file A consists of the nonconsenters of the National Educational Panel Study and the nonconsenters of the Panel Study "Labour Market and Social Security", respectively. We perform separate evaluation studies on both panel surveys. Before we discuss the design of these evaluation studies in more detail, this section provides a brief overview of the survey and administrative data sources.

5.1. Integrated Employment Biographies

The Integrated Employment Biographies (IEB) consists of administrative data obtained from social security notifications and different business processes of the German Federal Employment Agency. The different data sources are integrated for and by the Institute for Employment Research.

Figure 4 provides an overview of the business processes that generate IEB data. BeH information is provided for every employee covered by social security. Exclusions include individuals who did not enter the labor market and individuals who were self employed, since these groups are not subject to mandatory social security contributions. LeH and (X)LHG data are generated for individuals who received benefits in accordance with the Social Code Books (SGB) II (Sozialgesetzbuch 2003) and III (Sozialgesetzbuch 1997) (SGB II regulates welfare benefits for employable jobseekers in need and SGB III regulates employment promotion, in particular unemployment insurance). MTH and (X)ASU data are generated for individuals who were registered as jobseekers with the Federal Employment Agency or who participated in an employment or training program.

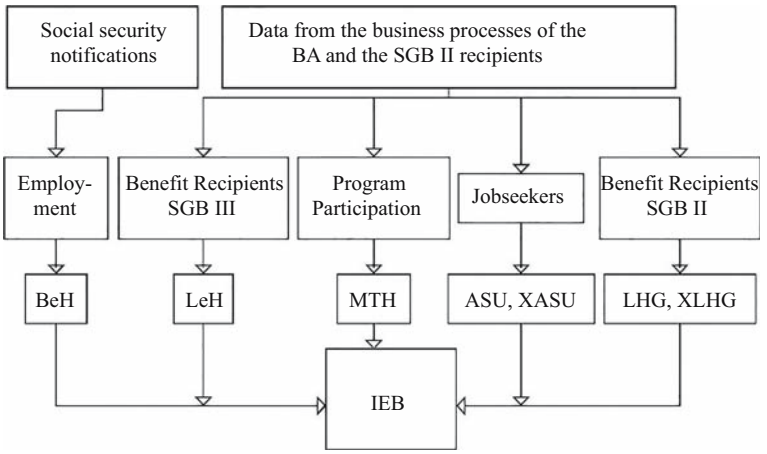

Fig. 4. Process data of the German Federal Employment Agency.

We refer to Jacobebbinghaus and Seth (2010) for a detailed description of the different data sources and of the IEB.

The IEB consists of a very large proportion of German residents, but not all. Thus, recipients of the statistical matching procedure should be limited to survey respondents who are also part of this subset of German residents. Note that in our evaluation study, this is guaranteed by design, as we only use the successfully linked cases.

Due to computational demands of the statistical matching procedure, it is mandatory to restrict the number of observations in B. Furthermore, researchers at the IAB cannot access the full IEB directly, since the size of the data set containing several billion records makes data handling difficult. For this reason the SIAB – a two-percent random sample from the IEB – is provided as a scientific use file that is easily accessible for all researchers at the IAB. Thus, we use the SIAB as the donor data set B for statistical matching. It still provides a very large donor pool of more than 1.7 million individuals and therefore guarantees a non-problematic donor to recipient ratio.

Availability and quality of IEB and SIAB data depend on various factors, including the data generating processes. It is out of scope of this article to go into further details (more information can be found in Antoni et al. (2016)). However, note that data on residents of federal states in the former German Democratic Republic are only available from 1993. Thus, to avoid biases, statistical matching is used on information generated after 1993. The information exclusive to the IEB, that is, the Z variables, mainly refer to individual employment history.

5.2. *National Educational Panel Study*

The National Educational Panel Study (NEPS) is carried out by the Leibniz Institute for Educational Trajectories at the University of Bamberg. The NEPS collects longitudinal data on competency development, educational processes, educational decisions and returns to education in Germany. Panel surveys on different age cohorts are conducted that provide data throughout the life course. The NEPS Starting Cohort 6 collects data on the

adult cohort. After a longer period between the first and second waves, which were carried out in 2007/2008 and 2009/2010, respectively, surveys for the adult cohort have been conducted yearly since 2011. The sample is drawn from municipality registration records of residents using a two-stage cluster sampling design with communities defining the primary sampling units and simple random sampling without replacement of individuals at the second stage. The target population of the adult cohort comprises residents in Germany who were born between 1944 and 1986, regardless of their nationality (Blossfeld et al. 2011). Variables that are exclusive to the NEPS data (that is, unavailable in the IEB/SIAB) are numerous. A unique characteristic of the NEPS compared to other surveys is the detail in information regarding the educational history of the respondents that form the Y variables of interest in the NEPS evaluation study.

Record linkage of NEPS and IEB data was carried out based on the nonunique identifiers, first and last name, date of birth, sex, and address information – postal code, city, street name, and house number. The consent rate in the NEPS adult cohort at the time of the linkage was 82% – yielding 14,065 consenters. Among the units that consented, 83.7%, that is, 11,778 units, could be linked deterministically and 7.5% (1,053 units) probabilistically. In the NEPS linkage, a link is called deterministic if the identifiers either match exactly or differ only in such a way that the probability for false positive links is still extremely low. For our evaluation study, we need a data set for which it is prudent to assume that all records are linked correctly. Therefore, we only keep those cases for which a deterministic linkage was possible. After additionally excluding every survey respondent whose latest linked IEB information is older than 1993, we arrive at a final data set consisting of 11,550 individuals. This data set is denoted as D^N_{det}.

5.3. *Panel Study "Labour Market and Social Security"*

The Panel Study "Labour Market and Social Security" (PASS) is an ongoing, nationally representative German household panel study, started in 2006 by the Institute for Employment Research. The aim of this study is to provide a database that enables an analysis of the dynamics of welfare benefits receipt after the introduction of the Unemployment Benefit II scheme in Germany in 2005. Information on labor market outcomes, household income, and unemployment benefit receipt are collected from more than 12,000 households annually. In addition to household interviews with the heads of the households, about 15,000 interviews with individual household members aged 15 and older are carried out.

The original PASS sample is composed of two subsamples: 1) a sample of households receiving unemployment benefit II (UB II Sample), which is drawn from recipient registers at the Federal Employment Agency; and 2) a sample of households from the general German population with an oversample of households with low economic status. The UB II Sample is refreshed each year to include new entries into the UB II population. PASS also introduced a replenishment sample for the general population sample in its fifth wave (for further information, see Trappmann et al. (2013)). As in the NEPS, there are many variables in the PASS that are not included in the administrative data of the IEB/SIAB. In particular, information on behaviors, attitudes, and subjective perceptions on the topics of social welfare benefits and labor market integration are available, which are the Y

variables of interest in the PASS evaluation study. The linkage consent rate to the IEB administrative data after the first five waves of PASS was at 79% (24,599 consenters). 87% (21,363 units) of the consenters could be successfully linked to the administrative data. 86% of the linkages were deterministic (18,425) using first and last name, date of birth, sex, and address information. Using our exclusion restriction that records need to be linked deterministically and that IEB spells need to be available after 1993, we end up with 18,202 individuals who are included in D^P_{det}.

6. Design of the Evaluation Study

We create synthetic nonconsent in the subset of deterministically linked respondents and check if, and to what extent, differences in estimates compared to before-deletion estimates can be reduced by using statistical matching as a supplement to record linkage. The design of our evaluation study comprises three steps. In the first step, we identify the deterministically linked cases in both data sources. In the second step, we model the probability of nonconsent based on the full survey data and use the predicted consent probabilities to introduce synthetic nonconsent among the true consenters. In the third step, we use statistical matching to find suitable administrative data donors for the generated nonconsenters and evaluate whether statistical matching reduces these differences. The three steps are visualized in Figure 5. Since statistical matching is only performed for the synthetic nonconsenters, they are denoted by A (the data recipients) in the figure, while the synthetic consenters are denoted by C, as record linkage is possible and thus all variables are available for them.

We note that although both surveys use complex sampling designs, we do not need to take any extra steps during the matching to account for the design, since both surveys are matched to a simple random sample of the IEB and the sampling design of the surveys is not relevant for nearest neighbor matching in this case. For statistical matching methods dealing with matching survey data sets with differing sampling designs, we refer the interested reader to Rubin (1986), Renssen (1998), Wu (2004) and Conti et al. (2016) for more details.

6.1. Generating Synthetic Nonconsent

In the data of all deterministically linked respondents D_{det}, the variable vector (X,Y,Z) is completely observed and the empirical distribution $f_{det}(x, y, z)$ of (X, Y, Z) is known (note that we always drop the superscripts N and P when we are not referring to a specific data source). These fully observed data will serve as the benchmark to evaluate whether nonconsent bias can be reduced by the proposed methodology. To introduce synthetic nonconsent among the true consenters based on realistic assumptions we use the full

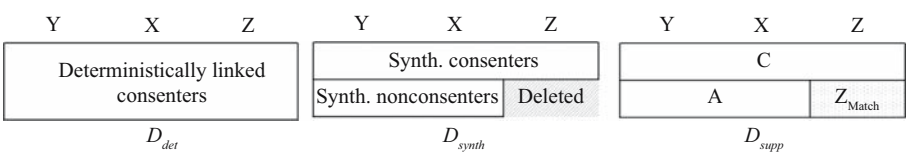

Fig. 5. The three steps of the evaluation study (left to right).

survey data to set up a model for the consent propensity. Specifically, we estimate a flexible, nonparametric logistic spline regression model with consent/nonconsent as the outcome variable and all X variables, as well as additional variables from Y for all survey respondents as covariates. All continuous covariates are included as B-splines. Since we need an estimated response propensity of every individual in D_{det}, survey variables used in the consent model that are subject to missingness need to be imputed. We use the software package mice in R (Van Buuren and Groothuis-Oudshoorn 2011) to generate m=5 imputations (based on eight iterations) using predictive mean matching and classification trees for metric and categorical variables, respectively. Hence, we need to make the implicit assumption that the missingness mechanism for all the imputed variables is missing at random.

Following the multiple imputation framework, the predicted consent propensities are obtained by averaging the predictions of the consent model from each of the imputed data sets. Respondents in D_{det} are stochastically chosen to be synthetic nonconsenters with a probability equal to their estimated consent propensity. Their Z variables are deleted to form D_{synth}.

In the case of the NEPS and the PASS, the resulting synthetic nonconsent rates are roughly 15 and 13%, respectively. To evaluate the effects of nonconsent and the performance of statistical matching under various assumptions, we also created data sets with synthetic nonconsent rates of 40 and 60% by adjusting the nonconsent probabilities accordingly. The observed empirical distribution of (X,Y,Z) for the remaining consenters is denoted as $f_{synth}(x, y, z)$.

Note that we implicitly make the assumption that the consent mechanism is *consenting at random* with respect to X and Y, that is, the probability to consent only depends on the survey variables. This assumption is necessary in our evaluation setup, since the true Z values are not observed for the nonconsenters by definition. The nonconsent model can therefore only include survey variables. Thus, the findings from our study will only be generalizable to situations where this assumption regarding the consent mechanism holds. However, it is prudent to assume that if statistical matching performs poorly in our evaluation study, this will also be the case if the assignment mechanism also depends on Z.

Since the data of the synthetic consenters are a random subsample of D_{det}, we cannot determine directly whether differences between the estimate using the synthetic consenters only and the estimate based on D_{det} are systematic or due to chance. A first indicator of potential bias in the estimates using only the synthetic consenters would be if any of the coefficients in the consent propensity model are significant. If there are significant parameters in the model, which is true in our case, then the assumption that the consent mechanism is consenting completely at random (CCAR) likely does not hold. As an additional evaluation whether, and to what degree, the observed differences in our study are systematic, we provide 95% confidence intervals for the estimates of interest under the null hypothesis that the consent mechanism is CCAR. Thus, the confidence intervals computed under the null hypothesis will be a measure of how much additional uncertainty we might expect due to the reduced sample size because of synthetic nonconsent. We can use these confidence intervals for classical hypothesis testing. If the confidence interval does not include the estimate of interest obtained using the remaining consenters based on the model described above, the null hypothesis that the consenting process is CCAR can be rejected.

The confidence intervals are obtained using Monte Carlo simulations. To obtain the confidence interval for an estimate of interest given a specific synthetic nonconsent rate r, we randomly delete $r \times 100\%$ of the data in D_{det} and compute the estimate of interest, based on the remaining cases. This is a realization of the estimand under the null hypothesis. By repeating this process 5,000 times, we make certain that the resulting empirical distribution is a good approximation of the true distribution under the null hypothesis. 95% confidence intervals are obtained by searching for the 2.5% and 97.5% quantiles of this distribution.

Note that we use Monte Carlo simulations only to create the confidence intervals – D_{synth} is created only once. This is a limitation of this evaluation study, since the creation of D_{synth} is subject to randomness and thus the results could differ over repeated simulation runs. However, due to the numerical expensiveness of the matching procedure, we are limited to a single run for the actual matching.

6.2. *Statistical Matching for all Synthetic Nonconsenters*

With the aim of reducing the nonconsent bias, statistical matching is performed for all synthetic nonconsenters. The specific matching method used here is called random distance hot-deck matching (D'Orazio et al. 2006b). For every synthetic nonconsenter, the method finds those k individuals from the administrative database who have the lowest distance regarding X, and from these k records, selects one at random and uses it as a donor. The main idea is that the empirical distribution of the k nearest neighbors' Z values approximates the posterior predictive distribution of missing data in Z given the survey respondent's realized X value, and the approach takes a random sample of size one from this conditional distribution. Similarly to stochastic versus deterministic imputation, it is preferable to draw from the posterior predictive distribution instead of simply using the expected value (Little and Rubin 2002). One could pick more than one donor in the spirit of multiple imputation to fully reflect the uncertainty that comes from matching randomly among the k closest donors (Rubin 1978, 1987). However, this approach is computationally intensive and is unlikely to affect the differences in point estimates since parameter estimates after multiple imputation are just averages over the parameter estimates in all imputed data sets. Nonetheless, as a sensitivity check we evaluated whether our findings change if we used $m = 5$ donors for each record. Since we did not find any differences in the results, the results reported below are based on picking only one donor.

We use the standardized Euclidean distance as a distance measure and set k to 20. While the Mahalanobis distance should do a better job for most statistical matching purposes, it is computationally more expensive. In addition, due to the large donor pool contained in the SIAB, the benefits of the Mahalanobis compared to the Euclidean distance should be negligible, since matching will be almost exact for most survey respondents.

The following variables are used as matching variables X: an indicator of whether the individual was ever married, age, an indicator for having children, salary in 2010, occupation, place of residence (formerly West or East Germany), and three variables on whether BeH, LeH, and LHG information is available. Some of these matching variables are only available for specific individuals due to the different data-generating processes in the IEB. Sex and nationality (German yes/no) are used as blocking variables, that is, IEB

units are excluded as potential donors for survey respondents if they do not have identical values in these variables. All of these variables are used in the matching procedure of both PASS and NEPS data to allow a comparison of the results in the two case studies. More information on the variables used can be found in Section 9, Appendix.

After statistical matching, we add the Z information of the identified matches for all synthetic nonconsenters to D_{synth}, and thus obtain a data set D_{supp}, for which (X, Y, Z) is again available for all deterministically linked survey respondents. The resulting empirical distribution is denoted as $f_{supp}(x, y, z)$. We can then evaluate whether differences in $f_{synth}(x, y, z)$ compared to $f_{det}(x, y, z)$ are reduced in $f_{supp}(x, y, z)$. Specifically, we look at marginal distributions in Z variables, correlations between Y and Z variables, and coefficients of regression models that use both Y and Z variables. For ease of reading, we use the terms *reference* or *benchmark* estimate for the estimates of interest using D_{det}.

7. Results

Using the predicted consent probabilities directly induces almost no differences in estimates in D_{synth} compared to the reference. After increasing the nonconsent rate to 40%, large differences can be observed for some estimands. Increasing the nonconsent rate to 60% increases these differences further. However, the general findings regarding the bias and the success of the statistical matching approach are similar for both consent rates. Therefore, we will only present the results using the smaller and more realistic nonconsent rate of 40% in this section.

7.1. Marginal Distributions and Means

In principle, marginal distributions for administrative data variables are available for the population in the complete administrative data. This means that data linkage would not be necessary to begin with. Thus, one could argue that biases in these marginal distributions should not be of any concern. However, this argument is only valid if the population of interest and the population of the administrative data are actually the same. If the survey population is a subset of the population of the administrative data, as for example in the case of the PASS subsample of unemployment benefit II recipients, it will still be important to evaluate whether the proposed method helps to correct for nonconsent bias in marginal distributions. However, note that the assumption that the conditional distribution $f(z|x)$ is the same for units in A and B is stronger if the survey population and administrative data population are different.

In our evaluation study, we examine the marginal distributions of some key measures in the IEB. An important characteristic of the IEB is its accurate information regarding the employment history of each individual (Jacobebbinghaus and Seth 2010). Thus, we evaluate whether statistical matching can reduce nonconsent bias in marginal distributions of these Z variables. Figure 6 presents results regarding the means of three important Z variables from the IEB: time in employment, complete gross salary earned in 2011, and the complete duration of Unemployment Benefit I receipt. Unemployment Benefit I is a specific social welfare payment in Germany that is paid during the first 6 to 18 months of unemployment. All values depicted here are ratios of the respective means to their benchmark values, that is, to the means in the complete linked data set D_{det}^{N}. As a reference,

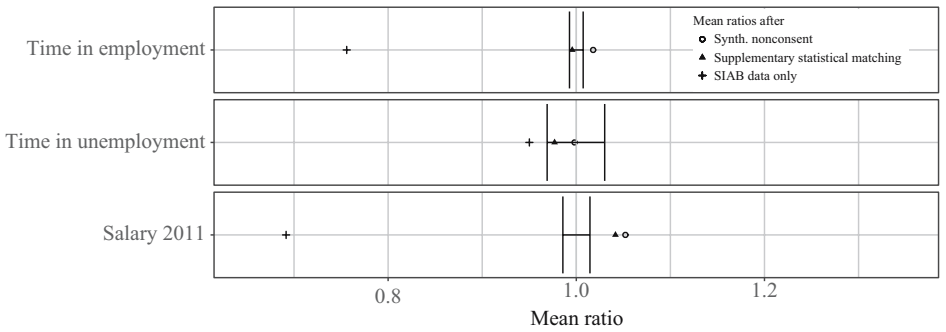

Fig. 6. Estimated means of three IEB variables divided by their NEPS benchmark estimate. The ratios are computed 1) after inducing 40% synthetic nonconsent and 2) after subsequent matching (ratios for the SIAB are included as a reference). The bars indicate the bounds of the 95% confidence intervals for the respective mean under the assumption of consenting completely at random.

Figure 6 also contains this ratio for the variables in B (the SIAB). Furthermore, a 95% confidence interval for the estimates assuming *consenting completely at random* is provided.

In the case of the NEPS, the synthetic nonconsent generates systematic differences in the variables time in employment and salary in 2011. In both cases, the confidence interval for the respective mean under the assumption of *consenting completely at random* does not cover the mean after inducing synthetic nonconsent. The difference in both variables can be reduced by using subsequent statistical matching for all synthetic nonconsenters. In contrast, no differences are created by the synthetic nonconsent process for the total duration of Unemployment Benefit I receipt and subsequent statistical matching slightly worsens the estimate from a bias perspective. The estimates based on the matched data are always close to the benchmark value despite the fact that estimates using the SIAB data would be substantially different, especially for time in employment and salary in 2011.

Looking at the mean ratios for the PASS (Figure 7), the findings are similar for the employment-related variables. For the variable salary 2011, the difference is slightly larger after supplementing the data with statistical matches.

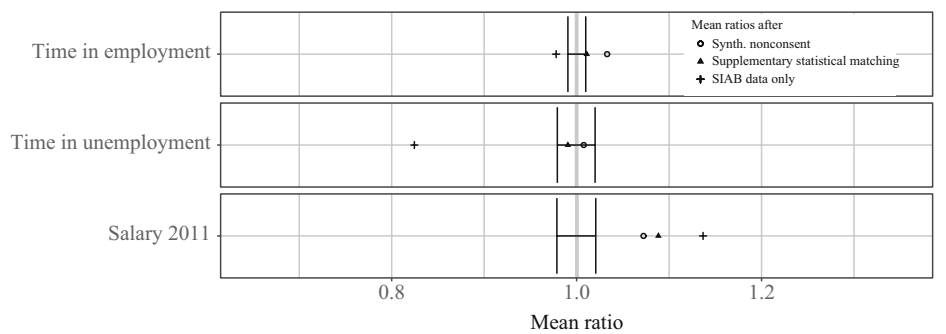

Fig. 7. Estimated means of three IEB variables divided by their PASS benchmark estimate. The ratios are computed 1) after inducing 40% synthetic nonconsent and 2) after subsequent matching (ratios for the SIAB are included as a reference). The bars indicate the bounds of the 95% confidence intervals for the respective mean under the assumption of consenting completely at random.

7.2. Correlations and Regression Model Parameters

The goal for record linkage is to be able to analyze the combined data. Univariate analyses can be performed on the administrative data without linkage (though, not always with respect to the specific survey population). Thus, it is essential to evaluate the methodology for bivariate and multivariate estimands that utilize both Y and Z variables. We only present results for the NEPS data in this section. Results obtained from the PASS data showed similar patterns and thus we exclude them for brevity.

In Figure 8, we present the effects of statistical matching on correlations of the three aggregate administrative data variables introduced in the previous section with three Y variables, that is, variables that are only available in the survey: years of schooling, age at first employment, and length of first employment. The before-deletion correlation – the empirical correlation in D^N_{det} – is plotted against the observed correlations in the data sets after creating synthetic nonconsent and using statistical matching for all synthetic nonconsenters.

We observe that there are more or less no differences in estimates using D_{synth} compared to the benchmark. However, almost all estimates are shrunk towards zero if statistical matching is applied. Instead of correcting for any differences created by nonconsent, statistical matching actually increases these differences. As explained in Subsection 2.2, the shrinkage towards zero is an indication that the conditional independence assumption is invalid.

We also estimate two regression models to evaluate to what extent multivariate relationships can be preserved after statistical matching; first, a Cox proportional hazards model (Cox and Oakes 1984) of the (log) length of the first unemployment episode as the dependent variable, and second, a linear regression model of the (log) gross salary in 2011

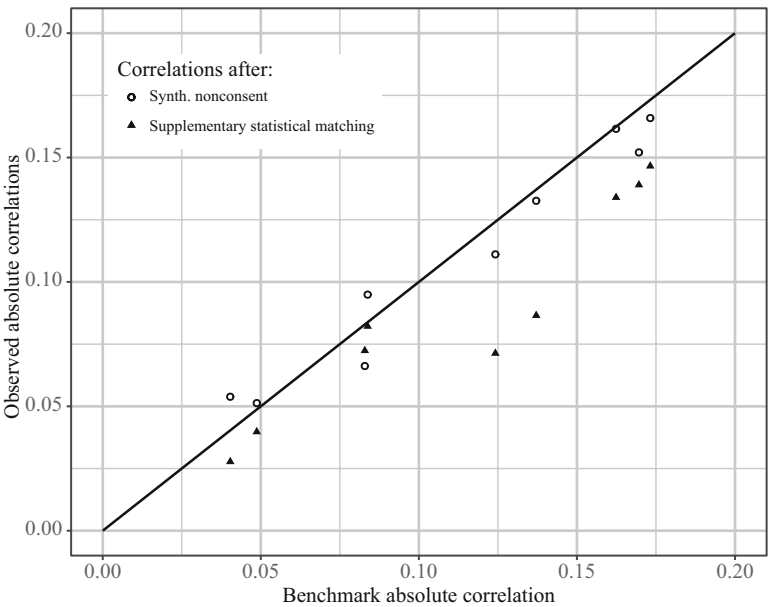

Fig. 8. Bivariate correlations for a subset of survey and administrative variables after 40% synthetic nonconsent and subsequent matching based on the NEPS data compared to the benchmark correlations.

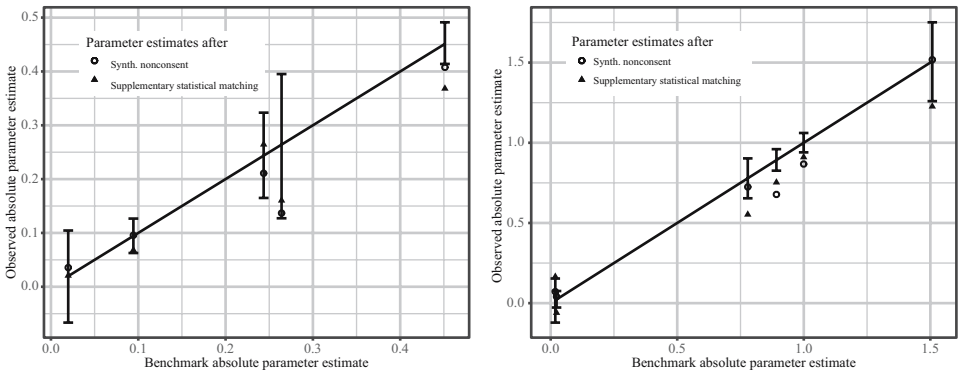

Fig. 9. Regression parameter estimates after 40% synthetic nonconsent and subsequent matching based on the NEPS data compared to the benchmark estimates (the bars indicate the bounds of the 95% confidence intervals for the respective parameter estimate under the assumption of consenting completely at random).

as the dependent variable. Both models use sex, age (and a quadratic term of age in the case of the salary model), occupational training, number of school years, and whether the mother's mother tongue is German as independent variables.

Similar to Figure 8, Figure 9 plots the before-deletion parameter estimates against the parameter estimates after inducing synthetic nonconsent and subsequent supplementation with statistical matching. The results of the parameter estimates are less cohesive. There are only a couple of coefficients for which differences in estimates after synthetic nonconsent compared to the reference are substantial. Again, the confidence intervals provide a test of whether or not the synthetic nonconsent process is significantly different from consenting completely at random. Point estimates are in some cases closer to the reference value after statistical matching, but in other cases the differences are larger. Also, absolute values of parameter estimates after supplementation are sometimes lower and sometimes higher than the true values. The unconditional relationship between Y and Z variables that can be observed in the matched data set is only the part of the relationship that can be explained by X. As explained in Subsection 2.2, omitting important confounding variables in X can lead to overestimation, as well as underestimation, of the unconditional effect after statistical matching. Also, even if all important confounding variables are included as matching variables, the lack of a potentially existing effect – conditional on every confounding variable – in the matched data can also lead to underestimation of the absolute value of the regression coefficients after matching. Both of these problems are only relevant if the conditional independence assumption is violated, but as our results concerning the correlations suggest, this is the case for almost all pairs of Y and Z variables that we examined.

8. Conclusion

Supplementing record linkage of survey and administrative data with nearest neighbor statistical matching is a straightforward idea when trying to reduce nonconsent biases. Since good parametric models are not necessary for nearest neighbor techniques and donor

sparseness is never an issue if large administrative data sets are used as the donor pool, it is the most widely applicable method available. However, the assumptions that are implicit when traditional statistical matching is used as a supplement to record linkage are very strong and will most likely never hold completely. The goal of the simulation study presented in this article was to evaluate empirically how well nearest neighbor matching performs, despite these assumptions. Our results suggest that biases in marginal distributions of administrative data variables can be corrected quite well, depending on the predictive power of the matching variables for the variable of interest. This is a particularly useful finding for situations where marginal distributions of administrative variables are desired for the survey population under study. However, the method is less suited for more complex analyses. The implications of the violation of the conditional independence assumption were substantial for both bivariate and multivariate analyses on the supplemented data sets.

Another downside of the approach is that the seemingly simple matching problem turns into a tedious task in practice, since preparing multiple data sources for statistical matching is a time-consuming and resource intensive process. In this study, significant efforts were undertaken to implement a high quality statistical matching procedure and analysis. Multiple issues, most of them related to measurement differences in the two data sets had to be dealt with in advance. These differences are likely to be present in any application of statistical matching of survey and administrative data.

The results from our simulations suggest that – with the exception of marginal distributions – the problems created by statistically matching nonconsenting units are worse than ignoring the nonconsent problem. Thus, even though the results are not easily generalizable to other applications, we advise caution when using nearest neighbor statistical matching to reduce linkage nonconsent bias for more complex estimates.

If other statistical agencies are considering statistical matching as a supplement to record linkage, our simulation design can be seen as a roadmap to empirically evaluate whether biases from nonconsent can be reduced for the specific application at hand. We emphasize that it is generally impossible to derive analytically which assumptions are stronger: the consenting completely at random assumption implied when analyzing only the data of the consenters or the assumptions required for statistical matching as discussed in Subsection 2.2 and Section 3. Both assumptions will never be fully met in practice, but the impact of the violation of the assumptions will depend on the available data, the nonconsent process, and the analysis of interest. Thus, statistical agencies might follow the simulation setup laid out in Section 6 to decide whether statistical matching could be a useful tool for their analysis goals, especially if a rich pool of jointly observed variables is available. We note that our evaluation study focused only on biases. Further research could extend our approach by including appropriate procedures based on the multiple imputation framework for enabling valid variance estimates after matching.

As discussed in Section 4, the missing at random assumption necessary for nearest neighbor imputation is weaker than for supplemental statistical matching. Therefore, one area of future research could focus on nearest neighbor imputation as a method to reduce nonconsent bias in a similar evaluation setting. Additional research questions related to how analyses on the imputed data set will be influenced by poor donor to recipient ratios due to low consent rates, should also be explored in this context.

Table 1. Variables used in the Matching Procedure of NEPS/PASS and IEB Data.

Variable	Type	Prereq.	NEPS Mean (SD)	PASS Mean (SD)	IEB Mean (SD)
Sex (female = 1)	Blocking	-	.504	.527	.480
Nationality (ger = 1)	Blocking	-	.955	.906	.868
BEH Spell	Asymm. Block.	-	.915	.579	.889
LEH Spell	Asymm. Block.	-	.208	.287	.404
No LEH Spell	Asymm. Block.	-	.431	.281	.596
Fulltime Employed	Asymm. Block.	BEH	.855	.489	.761
Single (no = 1)	Matching	LEH	.602	.444	.452
Has Children	Matching	LEH	.747	.536	.385
Place of residence	Matching	-	.156	.229	.156
Year of birth	Matching	-	1,963.7 (11.3)	1,965.6 (16.5)	1,965.2 (17)
Income (adjusted)	Matching	BEH	2,630 (3,248)	1,839 (1,682)	2,009 (2,568)
Occupation 1	Matching	BEH	.020	.030	.025
Occupation 2	Matching	BEH	.0004	.0006	.0008
Occupation 3	Matching	BEH	.170	.281	.229
Occupation 4	Matching	BEH	.072	.041	.053
Occupation 5	Matching	BEH	.731	.627	.686
Occupation 6	Matching	BEH	.006	.020	.006

Means and standard deviations for continuous variables; proportions for dichotomous matching variables.
Income values are inflation adjusted to the level of 2010.
Occupation is a nominally scaled variable with six categories.

9. Appendix

Table 1 shows all variables that are used in the matching procedure. We categorize every variable into blocking, asymmetric blocking, and matching variables. If a variable is used as a blocking variable, only individuals in the administrative data who have identical values in this variable are allowed to be used as donors. Matching variables are used to compute the Euclidean distance. Asymmetric blocking variables are used if blocking is only possible for specific respondents.

To illustrate, in our application, asymmetric blocking is required for the following reason: the IEB combines different sources of data that are generated from different BA business processes (see Figure 4) and all data sources provide different information. The variables from the different sources can only be used to find a donor for a survey respondent if it is certain that this information should be available in this respondent's (and therefore every similar individual's) administrative data. Therefore, we have to find proof – or at least strong indicators – in the survey data that this BA process should have been initiated by the respondent. However, not finding these indicators does not necessarily mean that the respondent's administrative data does not include this information. Therefore, these indicators are used in the matching process as asymmetric blocking variables.

10. References

Andridge, R.R. and R.J. Little. 2010. "A Review of Hot Deck Imputation for Survey Non-response." *International Statistical Review* 78(1): 40–64.

Antoni, M., A. Ganzer, and P. vom Berge. 2016. *Sample of Integrated Labour Market Biographies (SIAB) 1975–2014*. FDZ-Datenreport 4, Institute for Employment Research, Nuremberg, Germany. Avaiable at: http://doku.iab.de/fdz/reporte/2016/DR_04-16_EN.pdf.

Antoni, M. and S. Seth. 2011. *ALWA-ADIAB – linked individual survey and administrative data for substantive and methodological research*. FDZ-Methodenreport 12, Institute for Employment Research, Nuremberg, Germany. Avaiable at: http://doku.iab.de/fdz/reporte/2011/DR_05-11.pdf.

Biemer, P.P., R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman. 2011. *Measurement Errors in Surveys*. John Wiley & Sons.

Blossfeld, H.-P., H-G. Roßbach, and J. Von Maurice. 2011. "Education as a Lifelong Process." *Zeitschrift für Erziehungswissenschaft Sonderheft* 14. ISBN: 978-3-531-17785-4.

Brick, J.M. and G. Kalton. 1996. "Handling Missing Data in Survey Research." *Statistical Methods in Medical Research* 5(3): 215–238. Doi: http://dx.doi.org/10.1177/096228029600500302.

Brücker, H., M. Kroh, S. Bartsch, J. Goebel, S. Kühne, E. Liebau, P. Trübswetter, I. Tucci and J. Schupp. 2014. "The New IAB-SOEP Migration Sample: An Introduction into the Methodology and the Contents." *SOEP Survey Papers 216*. Avaiable at: http://hdl.handle.net/10419/103964.

Calderwood, L. and C. Lessof. 2009. "Enhancing Longitudinal Surveys By Linking to Administrative Data." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 55–72. New York: Wiley. ISBN: 978-0-470-01871-2.

Chen, J. and J. Shao. 2000. "Nearest Neighbor Imputation for Survey Data." *Journal of Official Statistics* 16(2): 113–131. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/nearest-neighbor-imputation-for-survey-data.pdf.

Conti, P.L., D. Marella and M. Scanu. 2012. "Uncertainty Analysis in Statistical Matching." *Journal of Official Statistics* 28(1): 69–88. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/uncertainty-analysis-in-statistical-matching.pdf.

Conti, P.L., D. Marella and M. Scanu. 2016. "Statistical Matching Analysis for Complex Survey Data with Applications." *Journal of the American Statistical Association* 111(516): 1715–1725. Doi: http://dx.doi.org/01621459.2015.1112803.

Cox, D.R. and D. Oakes. 1984. *Analysis of Survival Data*. CRC Press.

da Silva, M.E.M., C.M. Coeli, M. Ventura, M. Palacios, M.M.F. Magnanini, T.M.C.R. Camargo and K.R. Camargo. 2012. "Informed Consent for Record Linkage: A Systematic Review." *Journal of Medical Ethics* 38(10): 639–642. Doi: http://dx.doi.org/10.1136/medethics-2011-100208.

D'Orazio, M., M. Di Zio and M. Scanu. 2006a. "Statistical Matching for Categorical Data: Displaying Uncertainty using Logical Constraints." *Journal of Official Statistics* 28(1): 137–157. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-matching-for-categorical-data-displaying-uncertainty-and-using-logical-constraints.pdf.

D'Orazio, M., M. Di Zio and M. Scanu. 2006b. *Statistical Matching: Theory and Practice*. John Wiley & Sons.

D'Orazio, M., M. Di Zio and M. Scanu. 2009. "Uncertainty Intervals for Nonidentifiable Parameters in Statistical Matching." Proceedings of the 57th session of the International Statistical Institute, August 16–22, 2009, Durban, South Africa.

Fellegi, I.P. and A.B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64(328): 1183–1210. Doi: http://dx.doi.org/10.1080/01621459.1969.10501049.

Filippello, R., U. Guarnera and G. Jonas Lasinio. 2004. "Use of auxiliary information in statistical matching." Proceedings of the XLII Conference of the Italian Statistical 9–11 June 2014, Bari, Italy: 37–40.

Fosdick, B.K., M. DeYoreo and J.P. Reiter. 2016. "Categorical Data Fusion using Auxiliary Information." *The Annals of Applied Statistics* 10(4): 1907–1929. Doi: http://dx.doi.org/10.1214/16-AOAS925.

Fulton, J.A. 2012. *Respondent Consent to Use Administrative Data*, Ph. D. thesis, University of Maryland.

GDPR. 2016. "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)." *Official Journal of the European Union* L119: 1–88. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Gilula, Z. and R. McCulloch. 2013. "Multi Level Categorical Data Fusion using Partially Fused Data." *Quantitative Marketing and Economics* 11(3): 353–377. Doi: http://dx.doi.org/10.1007/s11129-013-9136-0.

Gilula, Z., R.E. McCulloch and P.E. Rossi. 2006. "A Direct Approach to Data Fusion." *Journal of Marketing Research* 43(1): 73–83. Doi: http://dx.doi.org/10.1509/jmkr.43.1.73.

Herzog, T.N., F.J. Scheuren and W.E. Winkler. 2007. *Data Quality and Record Linkage Techniques*. Springer Science & Business Media.

Jacobebbinghaus, P. and S. Seth. 2010. *Linked-Employer-Employee-Daten des IAB: LIAB – Querschnittmodell 2, 1993–2008*. FDZ-Datenreport, Institute for Employment Research, Nuremberg, Germany.

Jenkins, S.P., L. Cappellari, P. Lynn, A. Jäckle and E. Sala. 2006. "Patterns of Consent: Evidence from a General Household Survey." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169(4): 701–722. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00417.x.

Kadane, J.B. 1978. "Some Statistical Problems in Merging Data Files." *Compendium of Tax Research*, 159–179, Reprint in Journal of Official Statistics 17(3): 423–433. Avaiable at: https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/some-statistical-problems-in-merging-data-files.pdf.

Kreuter, F., J.W. Sakshaug and R. Tourangeau. 2016. "The Framing of the Record Linkage Consent Question." *International Journal of Public Opinion Research* 28(1): 142–152. Doi: http://dx.doi.org/10.1093/ijpor/edv006.

Little, R.J. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*, (2nd ed.). John Wiley & Sons.

Meinfelder, F. 2013. "Datenfusion: Theoretische Implikationen und praktische Umsetzung." In *Weiterentwicklung der amtlichen Haushaltsstatistiken*, edited by T. Riede, N. Ott and S. Bechthold, 83–98. Berlin: GWI Wissenschaftspolitik Infrastrukturentwicklung.

Moriarity, C. and F. Scheuren. 2001. "Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure." *Journal of Official Statistics* 17(3): 407–422. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-matching-a-paradigm-for-assessing-the-uncertainty-in-the-procedure.pdf.

Moriarity, C. and F. Scheuren. 2003. "A Note On Rubin's Statistical Matching using File Concatenation." *Journal of Business and Economic Statistics* (21): 65–73. Doi: http://dx.doi.org/10.1198/073500102288618766.

Mostafa, T. 2016. "Variation within Households in Consent to Link Survey Data to Administrative Records: Evidence from the UK Millennium Cohort Study." *International Journal of Social Research Methodology* 19(3): 355–375. Doi: http://dx.doi.org/10.1080/13645579.2015.1019264.

Ness, A.R. 2004. "The Avon Longitudinal Study of Parents and Children (ALSPAC) – A Resource for the Study of the Environmental Determinants of Childhood Obesity." *European Journal of Endocrinology* 151(Suppl 3): U141–U149. Doi: http://dx.doi.org/10.1530/eje.0.151u141.

Oberski, D.L., A. Kirchner, S. Eckman and F. Kreuter. 2017. "Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models."

Journal of the American Statistical Association. Doi: http://dx.doi.org/10.1080/01621459.2017.1302338.

Paass, G. 1985. "Statistical Record Linkage Methodology: State of the Art and Future Prospects." *Bulletin of the International Statistical Society. Proceedings of the 45th Session*. Voorburg, Netherlands: ISI.

Rässler, S. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer Science & Business Media.

Rässler, S. 2003. "A Non-Iterative Bayesian Approach to Statistical Matching." *Statistica Neerlandica* 57(1): 58–74. Doi: http://dx.doi.org/10.1111/1467-9574.00221.

Rässler, S. and H. Kiesl. 2009. "How Useful are Uncertainty Bounds? Some Recent Theory with an Application to Rubin's Causal Model." Proceedings of the 57th Session of the International Statistical Institute, August 16–22, 2009, Durban, South Africa. Available at https://www.isi-web.org/index.php/publications/proceedings.

Renssen, R.H. 1998. "Use of Statistical Matching Techniques in Calibration Estimation." *Survey Methodology* 24: 171–184. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/1998002/article/4354-eng.pdf.

Rodgers, W.L. 1984. "An Evaluation of Statistical Matching." *Journal of Business & Economic Statistics* 2(1): 91–102. Doi: http://dx.doi.org/10.1080/07350015.1984.10509373.

Rubin, D.B. 1976. "Inference and Missing Data." *Biometrika* (3): 581–592. Doi: http://dx.doi.org/10.2307/2335739.

Rubin, D.B. 1978. "Multiple Imputation in Sample Surveys – a Phenomological Bayesian Approach to Nonresponse." *Proceedings of the Survey Research Method Section of the American Statistical Association*: Joint Statistical Meetings 1978, San Diego, U.S.A.: 20–30. Available at: http://www.asasrms.org/Proceedings/index.html.

Rubin, D.B. 1986. "Statistical Matching using File Concatenation with Adjusted Weights and Multiple Imputations." *Journal of Business & Economic Statistics* 4(1): 87–94. Doi: http://dx.doi.org/10.1080/07350015.1986.10509497.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Sakshaug, J.W., M.P. Couper, M.B. Ofstedal and D.R. Weir. 2012. "Linking Survey and Administrative Records: Mechanisms of Consent." *Sociological Methods & Research* 41(4): 535–569. Doi: http://dx.doi.org/10.1177/0049124112460381.

Sakshaug, J.W. and M. Huber. 2016. "An Evaluation of Panel Nonresponse and Linkage Consent Bias in a Survey of Employees in Germany." *Journal of Survey Statistics and Methodology* 4(1): 71–93. Doi: http://dx.doi.org/10.1093/jssam/smv034.

Sakshaug, J.W., S. Hülle, A. Schmucker and S. Liebig. 2017. "Exploring the Effects of Interviewer- and Self-administered Survey Modes on Record Linkage Consent Rates and Bias." *Survey Research Methods* 11(forthcoming): 171–188. Doi: http://dx.doi.org/10.18148/srm/2017.v11i2.7158.

Sakshaug, J.W. and F. Kreuter. 2012. "Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data." *Survey Research Methods* 6(2): 113–122. Doi: http://dx.doi.org/10.18148/srm/2012.v6i2.5094.

Sakshaug, J.W. and B. Vicari. 2017. "Obtaining Record Linkage Consent from Establishments: The Impact of Question Placement on Consent Rates and Bias." *Journal of Survey Statistics and Methodology*. Doi: http://dx.doi.org/10.1093/jssam/smx009.

Sala, E., J. Burton and G. Knies. 2012. "Correlates of Obtaining Informed Consent to Data Linkage: Respondent, Interview, and Interviewer Characteristics." *Sociological Methods & Research* 41(3): 414–439. Doi: http://dx.doi.org/10.1177/0049124112457330.

Schulte Nordholt, E., J. Van Zeijl and L. Hoeksma. 2014. *Dutch Census 2011, Analysis and Methodology*, Technical report, Statistics Netherlands. ISBN: 978-90-357-1948-4. Available at: https://www.cbs.nl/NR/rdonlyres/5FDCE1B4-0654-45DA-8D7E-807A0213DE66/0/2014b57pub.pdf.

Sims, C. 1972. "Comments on Okner (1972)." *Annals of Economic and Social Measurement* (1): 343–345.

Singh, A., H. Mantel, M. Kinack and G. Rowe. 1993. "Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption." *Survey Methodology* 19(1): 59–79. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199300114475.

Sozialgesetzbuch. 1997. SGB Drittes Buch (III) – "Arbeitsförderung".

Sozialgesetzbuch. 2003. SGB Zweites Buch (II) – "Grundsicherung für Arbeitsuchende".

Trappmann, M., J. Beste, A. Bethmann and G. Müller. 2013. "The PASS Panel Survey After Six Waves." *Journal for Labour Market Research* 46(4): 275–281. Doi: http://dx.doi.org/10.1007/s12651-013-0150-1.

Van Buuren, S. and K. Groothuis-Oudshoorn. 2011. "MICE: Multivariate Imputation By Chained Equations in R." *Journal of Statistical Software* 45(3). Doi: http://dx.doi.org/10.18637/jss.v045.i03.

Wu, C. 2004. "Combining Information from Multiple Surveys through the Empirical Likelihood Method." *Canadian Journal of Statistics* 32(1): 15–26. Doi: http://dx.doi.org/10.2307/3315996.

Received June 2017
Revised May 2018
Accepted June 2018

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 935–960, http://dx.doi.org/10.2478/JOS-2018-0046

Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics

Maarten Vanhoof[1], *Fernando Reis*[2], *Thomas Ploetz*[3], *and Zbigniew Smoreda*[4]

Mobile phone data are an interesting new data source for official statistics. However, multiple problems and uncertainties need to be solved before these data can inform, support or even become an integral part of statistical production processes. In this article, we focus on arguably the most important problem hindering the application of mobile phone data in official statistics: detecting home locations. We argue that current efforts to detect home locations suffer from a blind deployment of criteria to define a place of residence and from limited validation possibilities. We support our argument by analysing the performance of five home detection algorithms (HDAs) that have been applied to a large, French, Call Detailed Record (CDR) data set (∼18 million users, five months). Our results show that criteria choice in HDAs influences the detection of home locations for up to about 40% of users, that HDAs perform poorly when compared with a validation data set (resulting in 35°-gap), and that their performance is sensitive to the time period and the duration of observation. Based on our findings and experiences, we offer several recommendations for official statistics. If adopted, our recommendations would help ensure more reliable use of mobile phone data vis-à-vis official statistics.

Key words: Mobile phone data; home location; home detection algorithms; official statistics; big data.

1. Introduction

By now, big data has well and truly arrived and their potential as well as the challenges it poses for official statistics have become much more evident. Consequently, there has been a clear demand to invest in pilot projects that explore how big data can be integrated into official statistics (Eurostat 2014; Glasson et al. 2013).

From a practical perspective, pilot projects are useful not only to identify practical issues (e.g., legal issues, data management). They are also particularly useful when

[1] Open Lab, Urban Sciences Building, 1 Science Square, Science Central, Newcastle upon Tyne NE4 5TG, United Kingdom. Email: m.vanhoof1@newcastle.ac.uk
[2] Task Force Big Data, Eurostat, European Commission, Joseph Bech Building, 5, Rue Alphonse Weicker, L-2721 Luxembourg. Email: fernando.reis@ec.europa.eu
[3] School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, U.S.A., and Open Lab, Urban Sciences Building, 1 Science Square, Science Central, Newcastle upon Tyne NE4 5TG, United Kingdom. Email: thomas.ploetz@gatech.edu
[4] Department SENSE, Orange Labs, Orange Gardens, 44 Avenue de la République, 92320 Châtillon, France. Email: zbigniew.smoreda@orange.com
Acknowledgments: The authors wish to thank Orange Labs France, especially the SENSE department, for making available the mobile phone data. We also like to thank INSEE, especially Stéphanie Combes, Marie-Pierre de Bellefon, Pauline Givord and Vincent Loonis for the construction of the validation dataset. Ultimately, we like to thank Morag Ottens and Andy Garbett for the help in preparin the article.

critically assessing the reliability of data sources and methodologies. It is reassuring that, regarding such assessments, Karlberg et al. (2015, 1) observe that: "There is a clear trend towards a more reflective approach, with an emphasis not only on producing high-quality statistics, but also on rendering explicit details on exactly how this is being achieved". When it comes to big data the importance of providing explicit details is not to be underestimated as big data sources, typically, do not adhere to official statistics' standards and principles – such as issues on coverage, representativity, quality, accuracy and precision (Daas et al. 2015) – and, consequently, neither do their methodologies.

In this article, we present a pilot study focusing on, arguably, the most important step for the application of mobile phone data in official statistics: identifying where someone lives, that is, detecting their home location. Current home detection methods for mobile phone data do not adhere to official statistics standards (or even to what could reasonably be expected from academic standards). We elaborate our argument by means of an extensive review of literature and an empirical analysis based on a large-scale, French, Call Detailed Records (CDR) data set. In doing so, we aim to show how current home detection practices came to be, how they are bound by limited validation possibilities and how they are sensitive to criteria choice or decision rule development. Given the lack of research on these problems, we argue that there is no clear framework on which to apprise the performance or the uncertainty of current home detection methods.

Our analysis evaluated the performance of five different home detection algorithms using a mobile phone data set from France. The case study allows us to reflect on the findings from a more practical point of view, whilst also contributing to our discussions and recommendations on the various uncertainties that underlie current home detection practices. We hope our contribution will help other researchers and practitioners to recognise the difficulties of integrating information on home locations sourced from mobile phone data into official statistics.

2. Mobile Phone Data, Official Statistics and the Role of Home Detection Methods

2.1. Mobile Phone Data and Official Statistics

Before looking at the methods used to identify home location, let us quickly consider how mobile phone data can be of interest for official statistics.

Over the last decade, the analysis of mobile phone data has grown into a mature research field with a wide array of applications that are being developed and applied (Blondel et al. 2015). One line of interest is that mobile phone data have the potential to capture temporal patterns of user presence (Deville et al. 2014), which could be used to estimate population density (Ricciato et al. 2015). In turn, these estimations could usefully support official statistics in developing countries (Blondel et al. 2012; de Montjoye et al. 2014).

Another line of interest relates to the large-scale recording of mobility patterns. As mobile phones can capture individual mobility for millions of users, applications have been developed that estimate nationwide commuting figures (Kung et al. 2014), long-distance trips (Janzen et al. 2016; Janzen et al. 2018), inbound tourism trips (Raun et al. 2016) and even domestic tourism trips (Vanhoof et al., 2017b).

These and similar developments have the potential to enhance official statistics in fields such as the delineation of urban areas (Vanhoof et al. 2017a), the understanding of migration patterns (Blumenstock 2012), or to complement tourism statistics (Ahas et al. 2008). They could even perform *nowcasting* of macro-economic and socio-economic aspects of populations (Baldacci et al. 2016; Marchetti et al. 2015; Giannotti et al. 2012; Pappalardo et al. 2016, and Vanhoof et al. 2018).

2.2. The Role and Method of Home Location

Common to many, if not all, mobile phone data research is the need to identify the home location of mobile phone users before proceeding to more advanced analysis. For example, knowing the place of residence is a prerequisite before analysing the amount of time spent at home and commuting patterns, which in turn fuel mobility and epidemiological models (Rubrichi et al. 2017). Besides its relevance within mobile phone data analysis, knowledge of home location also forms the crucial link between mobile phone data and other data sources, such as census data, making it a key enabler for the combination of information.

The method of pinpointing where someone lives consists of attributing a supposed home location to every single user in the database from the geographical metadata obtained from their mobile phone records. In practice, identifying a person's home means that a single cell tower is allocated as their home location. This allocation is based on the calling and movement patterns of each individual user. The spatial resolution of cell towers is used because most mobile phone data sets only have geographical data of the towers' positions. The assumption then is not that a user lives at that exact cell tower location, but rather somewhere in the area covered by the tower. It is remarkable that even though detecting the home location now forms a cornerstone of mobile phone research, home detection methods are often obscured in literature: details on their exact application, related uncertainties, perceived performance or even the validation processes are only rarely communicated.

In the following section, we show why current home detection practices are problematic. In an extended literature review, we show how, over time, methodologies for home detection have been simplified to single-step approaches using decision rules that are based on simple, *a priori* defined criteria of what defines a "home". Such methods are questionable because the possibilities to validate are limited, and there is a lack of knowledge on their sensitivity, specifically in respect to criteria choice. Our empirical work with a large, French, mobile phone data set exemplifies several of the problems we raise. It allows us to put the problems in a more practical context and outline their consequences in more detail.

3. Identifying Homes from Large-Scale Location Traces

Given the enormity of the data sets that capture geolocated traces of users, literature explains the automated methods developed for identifying the homes, or other meaningful places such as the workplace, of users. Here, it is necessary to distinguish between continuous location traces (e.g., GPS data) and noncontinuous location traces (e.g., mobile

phone data) where the latter do not provide a similar high-volume, high-resolution capture of location traces in time or space compared to the former.

As our main interest is to outline the deficits in the methods used for noncontinuous location traces, this section will start by reviewing the literature on automated home location.

3.1. *Identifying Meaningful Places from Continuous Location Traces*

The analysis of continuous location traces has been the focus of early developments in the automated identification of meaningful places. Related work typically used small-scale data sets, most commonly from continuous GPS traces, but also from Bluetooth, or Wi-Fi positioning (Wolf et al. 2001; Shen and Stopher 2014). The general methodology used to identify meaningful places from continuous location traces consists of a two-step approach.

In the first step, location traces are clustered in space (and sometimes in time) in order to detect *important places*. Techniques for clustering continuous location traces range from manual GIS analysis (Wolf, Guensler, and Bachman 2001; Gong et al. 2012) to automated, unsupervised analysis using, for example, k-means clustering (Ashbrook and Starner 2003), nonparametric Bayesian approaches (Nurmi and Bhattacharya 2008), or fingerprinting of the radio environment (Hightower et al. 2005).

In a second step, the important places identified are then annotated as *meaningful places* (such as home, work, recreation area). Annotation can be done either through interpretation, for example by expert judgment, by surveying the user that produced the traces, or through automation, mainly by means of time-space heuristics (Nurmi and Bhattacharya 2008).

3.2. *Identifying Meaningful Places from Noncontinuous Traces*

In contrast to the above-mentioned continuous location traces, the use of noncontinuous location traces has recently become very popular. Examples of activities that produce noncontinuous location traces in large-scale data sets are mobile phone usage, credit card transactions, or check-ins through location-based services (e.g., Foursquare) and online social networks (e.g., Twitter).

The identification of meaningful places from noncontinuous location traces poses substantial challenges, most notably due to the less frequent observations and the larger spatial resolution in which observations are captured (e.g., mobile phone data are only captured at the location of the cell tower used). However, these challenges are outweighed by the presumed advantages associated with the larger coverage, in terms of users, timespan and spatial extent of the data sources (Järv, Ahas, and Witlox 2014; Kung et al. 2014).

The following analysis will focus on one example of how to identify one meaningful place – the location of a user's home, using one prominent example of noncontinuous traces: Call Detailed Record (CDR) data.

CDR data are mobile phone data captured by the network operator every time a user makes or receives a text or call (hence the noncontinuous tracing). Note that the methods

and problems described in the following sections are not limited to CDR data, but are relevant for all data sets covering noncontinuous location traces.

3.2.1. Two-Step Approaches for Noncontinuous Traces

As with the two-step approaches for continuous traces, initial methods to detect home locations from CDR data also clustered location traces into important places before annotating them as meaningful places. For example, in Isaacman et al. (2011) individual traces from CDR data are clustered using Hartigan's leader algorithm. Clusters are then annotated into meaningful places by means of a logistic regression model that is trained on data from 18 persons for which ground truth was available. Next, and for each user, the cluster with the highest score on the logistic regression model is chosen to be the presumed home area.

3.2.2. Single-Step Approaches for Noncontinuous Traces

However, two-step approaches for noncontinuous location traces quickly gave way to single-step approaches that are now widely deployed in literature (Calabrese et al. 2014; Calabrese et al. 2011; Kung et al. 2014; Phithakkitnukoon et al. 2012). The difference between two-step and single-step methods is that the latter skips the clustering into important locations and thus acts directly on individual cell towers instead of groups of cell towers.

One of the reasons for switching to single-step approaches is that the standard clustering methods used in the two-step approaches make it difficult to construct consistent spatial traces when combined with noncontinuous location traces. Nevertheless, the main drawback of this switch to a single-step approach is that the spatial pattern of the location traces is largely neglected, as only single cell tower annotation is targeted. This increases the uncertainty of fixing home location, because single events at individual cell towers may be sufficient to undermine the method.

In practical terms, detecting a home in a single-step approach is done by using a decision rule that is based on an *a priori* definition of home – the *home criterion* as we call it – in order to produce a list of one or several cell towers that could be the home location. A standard example of a home criterion for the case of CDR data is "home is where calls are made during the night". The problem with single-step approaches is that such decision rules are being applied as *heuristics*, meaning that one general rule is applied to the location traces of all users even though a different set of decision rules could potentially lead to better results.

In terms of identifying home location, applying heuristics implies that meaningful places (like the home) can be described similarly for all users in the data set, regardless of the user's characteristics as observed in their movements and calling patterns. It seems logical that the imposition of this assumption can only be done when a proper evaluation and validation of their movements has been carried out, or when clear evidence exists for the use of a specific criterion or decision rule. For this reason, the following paragraphs will discuss how to define decision rules for one-step home detection methods and which criteria to use.

3.3. Defining Decision Rules for Single-Step Home Identification

3.3.1. Simple Decision Rules for Single-Step Home Detection

The core challenge for single-step home detection is in defining a decision rule that is simultaneously capable of i) distinguishing between different important places, and ii) annotating the correct home location. Most research employs simple decision rules that are either based on information from official statistics or rely on precedents found in literature.

When examining the existing decision rules in research literature, the most popular are: time-based limitations for the night ("home is the location that has the most activity between x p.m. and y a.m."), time-based aggregations ('home is where the most distinct days, or weekend-days are spent), and spatial groupings ('home is the location with the most activity in a spatial radius of x km around it), (Calabrese et al. 2011; Phithakkitnukoon et al. 2012; Frias-Martinez and Virseda 2012; Kung et al. 2014; Tizzoni et al. 2014). One example, using time-interval statistics from a Boston data set drawn from the American Time Use Survey (Calabrese et al. 2011), uses the highest distinct number of observations between 6 p.m. and 8 a.m. to derive home locations.

Almost all studies using simple decision rules rely on census data. They depend either on specific surveys and questionnaires to define the criteria deployed, (Calabrese et al. 2011) or, for high-level validation, on aggregated population density data (Phithakkitnukoon et al. 2012) or commuting Figures (Kung et al. 2014).

3.3.2. Complex Decision Rules for Single-Step Home Detection

A few studies have elaborated more complex decision rules for home detection. The seminal work of Ahas et al. (2010), for example, uses a tree-based approach that combines a set of criteria including distinct days of activities on a cell tower, the starting times of calls, deviations of starting time of calls, durations of calls, and this all for a training set of 14 people for which the ground truth was known. The decision rules, as defined by the classification tree, were consequently deployed to all users in an Estonian data set (as heuristics in other words), raising the question of how representative a training set of 14 people could possibly be for a large population.

The problem of small training sets was overcome in Frias-Martinez et al. (2010), who used a training set of 5,000 users to construct a complex decision rule for home detection. Deploying a Genetic Algorithm technique, they focus on finding the best combination of temporal criteria to denote home locations in an emerging economy. Their best performance is a correct prediction of around 70% for a subset of 50% of the users. Users were filtered on the basis of having at least a 20% difference in the percentage of total calls between the first and second eligible cell tower. The complex decision rule they use to obtain this result is to select the cell tower logging the most activity during the nights of Friday, Saturday, Sunday, Monday and Tuesday from 5:15 p.m. to 8:30 a.m.

The individual ground truth data in Frias-Martinez et al. (2010) are retrieved from users' contracts with the provider. This data is not available in most countries due to legal obligations to anonymise users or bans on linking individual information to CDR data.

As a consequence, Csáji et al. (2013), tried to derive a temporal decision rule, but this time without a training/validation data set at individual level. Applying an unsupervised

k-means algorithm to the temporal activity patterns of frequently used cell towers in Portugal, they found clusters that are interpretable as temporal patterns typically relating to presence at home, at work, or as not interpretable at all. Consequently, their decision rule to detect home locations was based on these temporal patterns interpreted as home presence. Compared to Frias-Martinez et al. (2010), one of the drawbacks of their approach is that they did not construct their criteria based on individual observations. This raises the question as to the degree to which such criteria are realistic for different subsets of users.

In a way, the subset representativity problem persists for all single-step approaches, regardless of whether their decision rules are defined in a complex or simple way. If the same decision rules is applied to all phone users, careful investigation into the effect at individual level, or at population subset level should be carried out, in order to know the degree to which generalisation favours or disfavours subsets of users. In other words, if decision rules are applied generically, indepth validation of the single-step approaches is important.

3.4. Validating Large-Scale Home Detection Methods

The use of a particular decision rule, whether derived from a census, borrowed from literature or defined by training sets, is often based on comparing population counts from mobile phone data with census data. However, such high-level validation does not offer a direct evaluation of performance at individual user level, nor does it allow for comparison between cases. In fact, assessing the performance of different decision rules by comparing the resultant population counts with census data is, strictly speaking, a rather limited alternative solely justified by the absence of individual level validation data.

The absence of validation data at individual level is a common problem in published research, and is therefore often taken for granted. However, the absence of validation data has several consequences. First and foremost, it impedes the creation of evaluation metrics that can assess the performance of home detection at individual level. Such an individual level evaluation could allow us to better understand the workings of different decision rules on a specific data set and user subsets, which in turn could enable a comparison between different decision rules, data sets, users and areas.

Secondly, the absence of validation data at individual level is implicitly why single-step approaches apply decision rules as heuristics. In the absence of individual level validation data, it is impossible to understand which decision rules works best for any individual user. Consequently, case-adjusted, adaptive algorithms cannot be developed. This implicitly forces researchers and practitioners to adhere to a one-size-fits-all solution in order to be clear and consistent.

It is worth noting that, currently, high-level validation is still assumed to be a good solution in the absence of individual level validation data. In particular, two observations stand out.

Firstly, census data is often used for high-level validation. For example, comparisons for small geographical areas can be made between the counts of home locations identified from mobile phone data and the aggregated counts of peoples' residential locations obtained from censuses. This is a very opportunistic, if not naïve, validation attempt as

census data has never specifically been gathered to serve this purpose and little or no information exists on how, for example, different spatial delineations or the distorted market shares of mobile phone operators could influence this kind of validation.

Secondly, it is noteworthy that no studies have used high-level validation to compare the performance of different decision rules. Nor are there studies that evaluate the sensitivity of high-level validation to criteria choice. This absence is probably because high-level validation is not informative enough to properly understand the differences between criteria, decision rules, and their performances. Given this, we are far from obtaining a consensus on which criteria are best, or on how to construct optimal decision rules. In fact, we are far from understanding the strengths and weaknesses of different home detection methods altogether. Given this, we should question the degree to which high-level validation contributes to the development and trustworthiness of home detection.

3.5. Current Deficits of Home Detection Using Noncontinuous Location Traces

In conclusion, we find a clear framework is missing to allow us to understand the performance, uncertainty and sensitivity of the criteria choice or decision rule development, especially at individual level, when using noncontinuous location traces to detect home location. Despite their widespread use, no clear reasoning exists as to why single-step approaches should be chosen over two-step approaches. Nor does a consensus exist on which criteria should be used, or how optimal decision rules for a given data set should be defined.

Similarly, it is striking that no work investigates the sensitivities of single-step approaches to criteria choice. Additionally, we find that the validation of large-scale home detection methods is severely limited because of the absence of ground truth data at individual level. As a result, current assessments of home detection methods are based on high-level validation, but the trustworthiness and exact contribution of this practice is rather dubious.

In summary, our findings indicate that the current methods to identify users' home locations for official statistics are rather questionable. We illustrate some of the aforementioned problems by means of a case study for identifying home locations using French CDR data.

4. Investigating Home Detection Algorithms for French CDR Data

To explore the application of single-step home detection methods on a French CDR data set, we start by constructing five home detection algorithms that incorporate different popular home criteria in simple decision rules. We apply these algorithms to the French data set, perform high-level validation, and investigate sensitivity to criteria choice. This allows us to demonstrate some of the aforementioned problems in an applied context.

4.1. The French CDR Data Set

CDR data are the most widely-used examples of mobile phone data in research. CDR data are passively gathered by operators for billing and maintenance purposes and are collected

every time a mobile phone user makes or receives a text or a call. Apart from technical metadata on the workings of the network, CDR data contain information on the time, the location (the cell tower used), as well as the caller and the call receiver.

For our analysis, we use an anonymised CDR data set from the mobile phone carrier Orange. The data covers the mobile phone usage of ~18 million users on the Orange network in France during a period of 154 consecutive days in 2007 (13 May 2007 to 14 October 2007). At that time, mobile phone penetration was estimated at 86% (ARCEP 2008). Given a population of 63.9 million inhabitants during the observed period (counted as the average of monthly estimates between May and October 2007 as obtained from the INSEE Website: www.insee.fr), this data set covers about 32.8% of all French mobile phone users and 28.6% of the total population.

The Orange France 2007 CDR data set is one of the largest CDR data sets available in terms of population-wide coverage and has been extensively studied before (Grauwin et al. 2017; Sobolevsky et al. 2013; Deville et al. 2014). It is the latest CDR data set available for France that allows for long-term, temporal continuous tracking of mobile phone users. Access to more recent data sets is limited by The French Data Protection Agency (CNIL), which is anticipating the EU General Data Protection Regulation and does not allow individual traces for periods of more than 24 hours to be collected, before being irreversibly recoded.

Some of the typical characteristics of CDR data sets that pose substantial challenges for their automated analysis are the temporal sparsity in observations and the spatially uneven distribution of the areas covered. The former results in only a few records per user per day. For example: for an arbitrary day in the French data set (Thursday, 1 October 2007), the median number of records per user was four, relating to only two different locations. Such statistics are representative for CDR based studies and can be deemed rather high compared to other large-scale noncontinuous datasets like credit-card transactions or Flickr photos (Bojic et al. 2015). The latter is the result of a demand-driven, nonuniform distribution of cell tower locations (higher densities of cell towers are found in more densely populated areas, such as cities or coastlines), meaning that the spatial accuracy of the dataset is restricted to the network's spatial resolution.

On the other hand, it is very attractive to have the possibility of researching the large-scale CDR data sets at population level, without users needing to share their locations. This increases the feasibility of automated applications such as home location. In addition, continuous data collection allows us to observe over extended periods, which in turn enables complex analysis and lessens any influence emanating from singular events and/or nonroutine behaviour.

4.2. Applying Five HDAs to the French CDR Data

4.2.1. Constructing Five HDAs with Simple Decision Rules Based on Popular Home Criteria

To perform home detection, we construct five basic Home Detection Algorithms (HDAs). Each incorporates one or two popular home criteria that are applied by means of simple decision rules. In order to select criteria, we took into account literature that dealt with

single-step approaches (e.g., Ahas et al. 2010; Isaacman et al. 2011; Calabrese et al. 2011; Tizzoni et al. 2014; Chen et al. 2014; Phithakkitnukoon et al. 2012; Csáji et al. 2013; Kung et al. 2014). We also used distilled criteria that were sometimes used independently (e.g., Tizzoni et al. 2014), sometimes combined (e.g., Ahas et al. 2010), sometimes within simple decision rules (e.g., Phithakkitnukoon et al. 2012), and sometimes within complex decision rules (e.g., Csáji et al. 2013 and Frias-Martinez 2010).

The HDAs we construct use the decision rules that 'home' is in the area of the cell tower where:

1. The majority of both outgoing and incoming calls and texts were made (amount of activities criterion),
2. The maximum number of distinct days with phone activities – both outgoing and incoming calls and texts – was observed (amount of distinct days criterion),
3. Most phone activities were recorded during 7 p.m. and 9 a.m. (time constraints criterion),
4. Most phone activities were recorded, implementing a spatial perimeter of 1,000 meters around a cell tower that aggregates all activities within (space constraints criterion) and
5. The combination of 3) and 4), thus most phone activities recorded during 7 p.m. and 9 a.m. and implementing a spatial perimeter of 1,000 meter (time constraints and space constraint criterion).

Note that throughout this article, we will estimate cell tower areas by means of the Voronoi tessellation of the cell tower network. The use of Voronoi polygons to describe the spatial patterns of cell tower coverage has disadvantages. Although widely used in literature, Voronoi polygons are a simplification of the actual capacity of cell towers to cover areas. In reality, capacity is dependent on factors such as humidity, urban environment, elevation of the cell tower, and orientation. Theoretically, developing estimation models for the coverage of cell towers should be possible, but such models need extensive field surveys for validation, surveys that are expensive and thus rarely available. Therefore, there exists an unobservable measurement error when using Voronoi polygons and most findings (including ours) are dependent on the assumption that this error has an insignificant impact.

Note also that Bojic et al. (2015) uses similar HDAs when assessing and comparing home detection methods for a credit card transaction and Flickr data set. This shows that the relevance of these algorithms goes beyond the case of CDR data and also serves other data sets with non-continuous location traces.

4.2.2. Applying Five HDAs to the French CDR Dataset

We apply all five HDAs to the Orange France 2007 CDR data set to detect the cell tower that covers the presumed home location (L1) for all users during all months in the data set (May to October). Besides the L1 cell tower, we gather information about the second (L2) and the third (L3) most plausible cell tower to cover the home location following the particular decision rule applied.

Table 1 shows the total number of times each HDA could detect an L1, L2 or L3 cell tower based on the CDR data of ~18 million users and when applied to each month in the

Table 1. Number of times (in millions) for which an L1, L2 or L3 cell tower could be detected from an individual user's CDR data by the various Home Detection Algorithms (HDAs) when applied per month in the data set. Percentages are column-wise and with respect to the number of L1 detections.

Number of users (in million) with	Amount of activities (algorithm 1)	Amount of distinct days (algorithm 2)	Time restraints (amount of act.) (algorithm 3)	Space restraints (amount of act.) (algorithm 4)	Time and Space restraints (algorithm 5)
Detected L1	109.4 (*100%*)	109.4 (*100%*)	98.4 (*100%*)	109.4 (*100%*)	98.4 (*100%*)
Detected L2	102.2 (*93.5%*)	102.2 (*93.5%*)	78.0 (*81.3%*)	102.0 (*92.8%*)	78.4 (*79.6%*)
Detected L3	96.1 (*87.9%*)	96.1 (*87.9%*)	65.0 (*66.1%*)	66.1 (*86.6%*)	62.3 (*63.3%*)

data set. Given the availability of six different months (mid-May to mid-October), non-restrictive algorithms (such as algorithm 1 and 2) will be capable of detecting an L1 cell tower for about 109.4 million users (~18 000 000*6). Restrictive algorithms, such as the time-constraining algorithm 3, have fewer users for which a presumed home cell tower (L1) can be detected. The reason is that some users might not have made or received calls or texts during the restricted timeframe, so no CDR records exist and therefore the algorithm cannot identify an L1 cell tower.

For example, when we compare the number of times algorithm 1 (all activities) was capable of detecting an L1 compared to algorithm 3 (only nighttime activities), we can derive that up to 10% (98.4/109.4) of the users did not have mobile phone activities during the night. This made it impossible for the time-constraint HDAs to detect a cell tower presumably covering the home location. It is also interesting to note that, depending on decision rule of the algorithm, between 79.6 and 93.5% and between 62.3 and 87.9% of users have an L2 or L3 cell tower that could also be nominated as the home location cell tower, as they only varied by a slight degree compared to the L1 (or L2) cell tower(s). In other words, the decision rules applied do not overly discriminate between the eligibility of different cell towers to be the presumed home location. This raises the question of whether the French data set would not have benefited from a two-step approach.

4.3. *Comparison of HDAs at Individual Level*

One intriguing question is whether, for the same individual user, different HDAs would detect different home locations (L1 cell towers). We assess to which degree two different algorithms detect similar home locations for all individual users in the data set by evaluating the Simple Matching Coefficient (SMC) (Bojic et al. 2015):

$$\%SMC\ (algorithm_A, algorithm_B) = 100 * \frac{\sum_{i=1}^{n} \delta(Home_{A,i}, Home_{B,i})}{N} \quad (1)$$

where $i = 1, \ldots, N$ denotes the N users analysed, and $\delta(\mathrm{Home_{A,i}}, \mathrm{Home_{B,i}})$ is the Kronecker delta which is equal to 1 when the home detected by algorithm A for the i-th user is identical to the home detected by algorithm B for the same user. The Kronecker delta becomes 0 otherwise. Values of per cent SMC thus range between 0 and 100 and can be interpreted as the percentage of individual cases for which both algorithms detected the same home locations. When calculating SMC values, we omit all cases where one of the algorithms failed to detect a home location (e.g., when no observations were left after implementing a time constraint).

Figure 1 shows the SMC values for all pair combinations of HDAs during the different months in the data set. In general, pair accordance ranges between 61.5% and 96.4% of the detected homes, resulting in discordance rates between about 40% and 4%. In absolute numbers, this means that different decision rules predict different homes for between 6.8 and 0.6 million users. The patterns of (dis)similarities between HDAs are rather clear. Algorithms that incorporate time-constraints (algorithms 3 and 5) have a high degree of variance with algorithms that count the amount of activities (algorithm 1), distinct days (algorithm 2), or perform spatial groupings (algorithm 4), all of which show rather high degrees of pair accordance. The different results for the time-constraints algorithms might

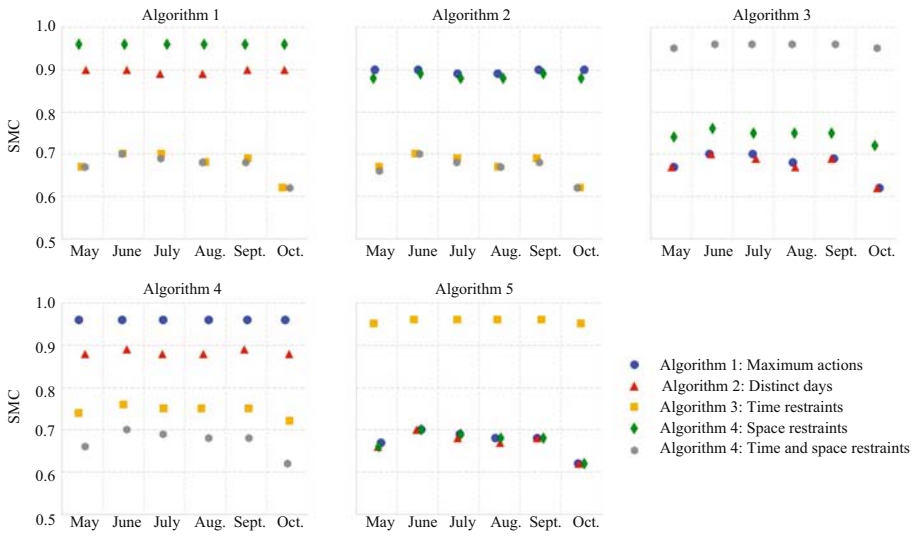

Fig. 1. SMC values for all pair combinations of HDAs, for each month in the data set. SMC values express the ratio of users for which two HDAs detect the same home.

stem from sparser observations or different movement patterns during the night, but exact reasons are unknown.

4.4. High Level Validation of Home Detection Algorithms

Given that different HDAs give different results for a considerable share of all individual users, the question becomes which decision rule should be preferred. As discussed previously, no consensus exists in literature on which decision rule(s) are best. This is partly because of the absence of comparative studies, but mainly because of the lack of proper validation data at individual level. In our case too, individual-level ground truth data was not available and so our assessment is at high-level, comparing census figures with population counts produced by HDAs.

4.4.1. National Statistics Validation Data Set

In contrast to related works, our high-level validation is based on a unique validation data set that was created in collaboration with the French National Statistics Institute (INSEE). To construct the validation data set, the Public Finances Directorate General (DGFIP) collected individual (or household) home locations from revenue declarations, housing taxes and the directory of taxable individuals. It then aggregated this information into population counts at the resolution of the Orange cell tower network (see also Figure 3a). In other words, an estimation of the population numbers, based on census data, for the geographical areas created by the Voronoi polygons of the Orange cell towers was produced and made available to the research project under a non-disclosure agreement.

It is a huge advantage to have access to a validation data set that has the same spatial resolution as the mobile phone network. It avoids the spatial translation of statistical zones

to the cell tower Voronoi areas, which is complicated and prone to errors (Frias-Martinez et al. 2010), given the spatially uneven distribution of cell towers.

Unfortunately, the individual (or household) home locations used to construct the validation data set could only be made available for the year 2010. However, for reasons explained in the previous paragraph we do opt to use this validation data set with its temporal mismatch (the mobile phone data set covers 2007) over the low resolution, publicly available census data that are updated every year. Since we only use the validation data set for relative comparisons between HDAs (i.e., no absolute validation is attempted), the assumption we introduce concerning this temporal mismatch is that relative population patterns do not change drastically within three years.

4.4.2. Validation of HDA Results at Cell Tower Level

To compare results from HDAs with the proposed validation data set, we evaluate the degree of similarity in population counts attributed to all cell tower areas. Note that we do not target an absolute assessment of similarity, as this is impossible given the unknown spatial distribution of the 28.7% sample of Orange users and the differences in times of collection between the CDR data set (2007) and our validation data (2010). Instead, we compare general patterns of estimated populations by means of vector comparison.

In our case, a first vector denotes the estimated population by one HDA for all cell tower areas and is compared to a second vector that describes the validation population count for exactly the same cell tower areas. Both vectors thus have an equal length ($n = 18\,273$ the amount of cell towers in the Orange network). To quantify the similarities and differences between both vectors, we use a standard Cosine Similarity Metric (CSM). According to Ye (2011, 91): "The cosine similarity is a classic measure used in information retrieval and is the most widely reported measure of vector similarity", and it is based on the angle between two vectors described by its cosine:

$$\cos(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} * \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2}$$

where x_i and y_i are components of vectors $\vec{x}$ and $\vec{y}$ respectively and n is the total number of cell tower areas.

Values of the cosine will range between -1 and 1. A value of 1 indicates the highest similarity in orientation (the angle between $\vec{x}$ and $\vec{y}$ is zero degrees), 0 indicates the lowest similarity in orientation (the angle between $\vec{x}$ vector and $\vec{y}$ vector is 90 or -90 degrees) and -1 indicates an opposite orientation (the angle between $\vec{x}$ and $\vec{y}$ is 180 degrees). Deriving the angle between two vectors and expressing it in degrees (°) consequently gives us the CSM value we want:

$$\mathrm{CSM}(\vec{x}, \vec{y}) = \left| \cos^{-1}\left(\frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} * \sqrt{\sum_{i=1}^{n} y_i^2}} \right) * \frac{180}{\pi} \right| \tag{3}$$

A CSM value of 0° denotes the highest possible similarity, 90° indicates the lowest similarity in orientation whereas 180° degrees refers to an opposite orientation.

4.4.3. Validation with Census Data: CSM

Figure 2 shows the calculated CSM values for all HDAs and for different months. The distinct days that the algorithm performs best in replicating the population pattern of the validation data set, followed by the number of activities and the time-constrained number of activities. The HDAs that involve grouping in space perform worst, even though the applied perimeter (1 kilometre) in reality does not correspond to a substantial distance. It is worth noting that the performance of all HDAs range between 34° and 38°. This is substantially different from the intended 0°, which would signify a perfect match with the validation set. In other words, a 'gap' of about 35 degrees exists when using the CSM measure. This is indicative for the limited performance of our HDAs and raises the question of whether there is a structural limitation on the performance of single-step HDAs when applied to the French data set or to CDR data in general.

Interestingly, the performance of all HDAs is rather similar. Especially in their temporal patterns, where lower CSM values for June and September, and higher values for May, July, August and October are observed. A possible explanation for the high SMC values for May and October is the limited number of available days for these months in the data set (18 and 14 days respectively). This indicates that data should be collected for a certain duration for the HDA to perform properly.

The highest CSM values are observed during summer (July and August). All algorithms are sensitive to this temporal change, most likely because of the changing spatial behaviour of users who go on holiday (see also Deville et al. 2014; Vanhoof et al., 2017b). Time-limited criteria are more sensitive to temporal changes, which raises questions about their widespread adaptation in literature. In addition, it is interesting to note that differences between each algorithm are smaller than the differences of each algorithm

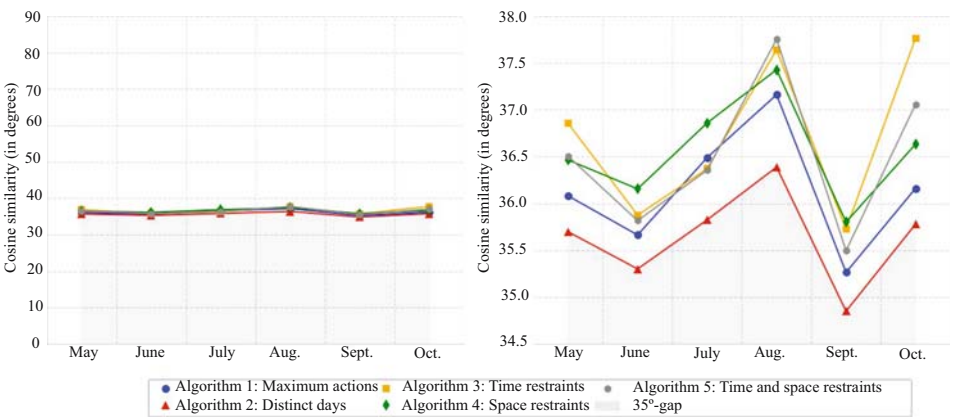

Fig. 2. CSM values (in degrees) of the comparison with ground truth data, for all HDAs applied to all months in the dataset. CSM values were calculated at cell tower level. The 35°gap is denoted as the difference between the best performing HDA and the expected CSM of 0° in the case of a perfect match between population counts from home location and the validation data set.

over time. Future analysis of HDA performance should therefore take into account the time period.

4.4.4. Spatial Patterns of Population Count

Although the CSM values for all HDAs are within a rather small range, it is important to realise that small differences in CSM values can imply major differences in the related spatial patterns of population counts.

Figures 3c and 3d for instance, show the spatial patterns of population counts obtained by the number of activity algorithms for June and August respectively. The difference in CSM values between both is a mere 1.08°, but their spatial pattern, as emphasised by the Getis-Ord G_i* statistic (Getis and Ord 1992), is rather different. This statistic shows statistically significant clusters of high (hotspots) or low (coldspots) population counts. In August, for instance, the detected hotspots illustrate clear clusters of high numbers of home locations near sea and mountain areas. This is in contrast to an expected spatial pattern, where high clusters of population counts are found near cities and in urban areas, as can be seen from the spatial pattern of the validation data set in Figure 3b.

The spatial pattern of the differences between the validation data sets and detected homes in June and August are given in Figures 3e and 3f and visualise this contrast. Note that in Figure 3, the centre of Paris is often denoted as a coldspot because of the high density of cell towers, so each tower has a lower number of users, resulting in apparent coldspots. This effect is also visible in other city centres where cell tower density is high.

5. Discussion

5.1. Differences at Individual Level and the Absence of Ground Truth Data

Our results showed high discordance rates between different HDAs (ranging from 4% to 40% of the individual mobile phone users). This finding challenges the use of single-step home detection approaches for the French CDR data set when done without fully justifying the home location criteria used and the decision rules involved in the HDAs. As we argued, such justification is currently absent in the aforementioned literature, mainly because of the absence of ground truth data at individual level. Our case study clarifies how the absence of individual ground truth data necessitates a heuristic application of decision rules in current home detection methods. By this, we mean that one decision rule is applied to all users in the data set, regardless of the nature of their CDR traces. Obviously, the better approach would be to have non-generic algorithms that could flexibly select decision rules (and validation) based on the characteristics of individual user traces. Such a solution, however, would require large training samples (individual ground truth) to learn how to switch between different decision rules. As yet, these are not available.

5.2. Sensitivity of Performance Considering Time and Decision Rule Choice

Performing high-level validation on five HDAs, by comparing population counts with census data, unveiled rather poor performances (CSM values between 34° and 38°), and a clear sensitivity to the chosen time period. In fact, for the French data set, defining a time

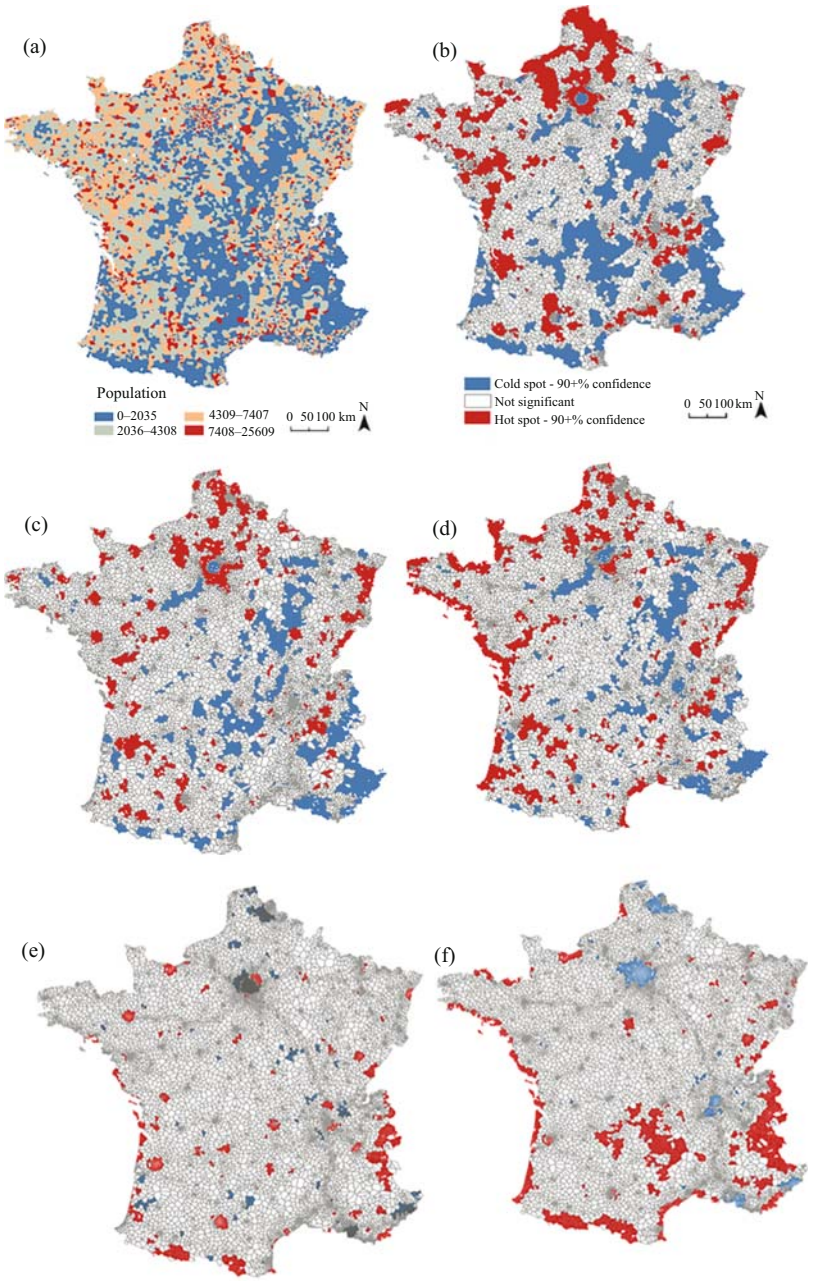

Fig. 3. Population counts of the validation data set: (a) Hotspots (red) and coldspots (blue) as defined by the 90+% interval of the Getis-Ord G_i^* statistic for the population counts of the validation data set (b); for the number of detected homes using the amount of activities algorithm in June (c); for the number of detected homes using the amount of activities algorithm in August (d); for the log(ratio) between the amount of activities algorithm in June and the population counts of the validation data set (e); and for the log(ratio) between the amount of activities algorithm in August and the population counts of the validation data set (f). All maps are compiled from Voronoi tessellation of the Orange cell towers. Figures b, c, d, e, and f share the legend of Figure b.

period to carry out home detection seems as important as criteria choice. As was illustrated by the spatial pattern of population counts in Figure 3d, and by increasing CSM values in July and August in Figure 2, summer periods should be avoided when running HDAs. Additionally, shorter observation periods (like May and October in our case) also seem to influence the performance of HDAs.

When comparing criteria, it is clear that the space constraints criterion is outperformed by all others. The main logic behind grouping close locations together (in this case, within a 1 km perimeter) is to avoid frequent handovers between close cell towers. However, on a large scale, such a precaution, seems to have a negative impact. Furthermore, the extreme volatility of the performance of the time constraints criteria is remarkable. Clearly, this criterion is not able to cope with (changing) user behaviour during the summer months, resulting in the worst performance obtained.

5.3. The 35°-gap in High-Level Validation

The most telling result of our analysis, is that all tested HDAs have CSM values which are still far off from the intended 0°, as can be observed in Figure 2. Additionally, it is remarkable that CSM values for all HDAs occur in the same, rather small, range of CSM values (even though a small difference in CSM can induce a rather profound change in spatial patterns). The *35°-gap* observed is indicative for the (current) limits of single-step approaches based on simple decision rules, at least at cell tower level (as aggregation to higher levels might diminish the gap considerably).

The *35°-gap* also adds to the discussion of high-level validation. As mentioned before, the absence of individual-level validation is hindering a clear understanding of why the performance between algorithms may differ. Playing devil's advocate, one couldn't care less about individual correctness, as long as the statistical performance at nationwide level is sufficient. However, given the considerable differences between census and mobile phone home location data at cell tower level, as reveled by both the *35°-gap,* in Figure 2, and the clear differences in spatial patterns in Figure 3, it seems inevitable that investigations at individual or subset level will need to be undertaken to improve insights into the workings of HDA and, ultimately, the performance of home detection methods in general.

It is clear that the *35°-gap* requires further exploration so as to understand its constituent parts. We consider, at least, the following elements to be of importance:

- *Distorted local market shares*: Local market shares for individual mobile phone operators can be highly volatile and are often unknown. This causes a lot of uncertainty when it comes to high-level validation with census data, as the percentage of the population that the operator actually captures in different regions is unclear. Unknown local market shares therefore impede both validation techniques that perform in pairs and/or absolute comparisons between population estimations and ground truth figures collected by nationwide censuses. They also most probably hinder validation techniques that are based on relative differences (like the CSM metrics).
- *Diversity of mobile phone use*: Differences in mobile phone use between users and/or regions can structurally influence the validation of single-step HDAs. When concentrated, differences in mobile phone use influence high-level validation in the

same way as distorted market shares would. Additionally, it is clear that mobile phone usage changes rapidly over time. It can be argued that the use of phones for professional or private purposes was different in 2007 than it is today. Unfortunately, such usage contexts are not available in CDR data. Neither can they be easily derived since, in general, privacy regulations ban the linking of CDR data and customer databases that gather for example, billing addresses or type of payment information. In other words, traces in CDR data will be of a different nature at different times because of differences in mobile phone usage. This implies that information on mobile phone usage is necessary to understand the effect on home detection performance.

- *Differing definitions of home*: Differences between the definition of home in census data and the definition of home by HDAs may cause structural discordance when validating the latter by the former. Even though official statistical practices have a tradition of distinguishing between different definitions of home, such as 'usual resident population' and 'second home population', it remains unclear to what degree mobile phone data is capable of capturing such concepts of home and to what degree different decision rules would favour the detection of different types of homes.
- *Technical aspects of the data collection and methodology*: Research has paid wide attention to the technical aspects of mobile phone data, especially when it comes to the estimation of cell tower areas and their translation into statistical areas (Ricciato et al. 2015). In our case, we avoided the translation problem by constructing a validation data set at cell tower level, but for many cases this is not an option. Estimation of cell tower areas was done by Voronoi polygons, which introduces errors at a local scale, but could also structurally influence high-level validation if, for example, areas of cell towers in high population density locations were consistently underestimated. Here too, the effects on high-level validation practices are currently unknown, but we expect them to be minor compared to previous points.

For all the points raised above, no quantification of their effect(s) has yet been explored. Additionally, it is worth nothing that some of these points may become more or less relevant over time due to, factors such as technical advancements or regulations. The EU General Data Protection Directive 2018, for example, will probably make it harder to work at the individual level (individual mobile phone use, different types of home definition). This will make high-level validation techniques more relevant, and thus increase the need to have proper knowledge of local market shares. For this reason, it is difficult to assess the relevance of the findings given in this article. However, we strongly suggest that all are the topics of future research and consideration.

6. Recommendations

Throughout this work, we have given suggestions concerning the use of HDAs for mobile phone data. In summary, we believe we can compile these suggestions into a set of three recommendations, which are relevant at different levels.

1. **Individual level:** Currently, the biggest problem in ensuring the reliable use of HDAs for mobile phone data (and, in extension, other similar data sources like

location based services or geotagged online social networks) is the absence of ground truth data at the individual level. We strongly recommend the collection of ground truth data linking mobile phone usage, the related CDR data, and movement patterns of individual users. Even if collected for only small samples of users, this step is essential for giving proper estimations of error and performance of HDAs at individual level. It also would help in understanding the differences between decision rules, and plausibly allow for the development of non-generic HDAs that could switch between decision rules based on the characteristics of individual traces, instead of being generically applied to all users. Additionally, the availability of individual level ground truth could shed light on the structural effects that currently obscure high-level validation practices, such as the changing usage of mobile phones and the differences between declared homes and lived-in homes as captured by census and mobile phone data respectively.

2. **National level:** Apart from ensuring nationally representative sampling of individual level ground truth data, we believe it to be important either to understand local market shares of single operators, or to collect mobile phone data from all operators in the territory. Without this information, high-level validation of population estimations at nationwide level will remain flawed, making it impossible to describe correctly the performance of HDAs at a larger geographical scale. In addition, resolving the local market share issue is a crucial step in the investigation of the (spatial) representativity of available mobile phone data sets, as unknown market shares at local level impede the analysis of subset populations in data sets.
3. **International level**: Finally, we believe that one of the key components to ensure reliable use of mobile phone data in official statistics is the opportunity to test ideas and methodologies on different data sets, which contain differing populations and cover various time periods. This is not necessarily a matter of testing for uniformity. On the contrary, it is a matter of understanding the limits of current methodologies, assessing the true potential for applications and anticipating the wider challenges posed by fast-evolving technology usage and deployment. All of these factors are necessary to ensure the future applicability of mobile phone data sources in official statistics.

As we reflect on the direction that further investigation should take, together with the feasibility of carrying out the recommendations proposed, we realize that this is a larger intervention than any single researcher, research group, national statistics office or even operator can be expected to take. Therefore, it is encouraging to see that collaborations are being formed to address different parts of the problem.

In France, for example, a collaboration between the operator Orange and the national statistics office INSEE is investigating different aspects of the high-level validation of home detection practices, such as translating Voronoi polygons into existing statistical grids (Sakarovitch et al. in preperation.).

On a European scale, the ESSnet Big Data project has been organising the exchange of best practices for the integration of mobile phone data (and multiple other big data sources) in official statistics. Its goal is directly in line with the recommendations previously described (especially recommendation 3: international level), facilitating the

uniformity of quality and methodologies for the use of big data sources in European official statistics (ESSnet Big Data 2018).

As a last example, the Open Algorithm project (OPAL) is a collaboration between operators, academia, and institutional partners who are building a platform to allow the use of *Open Algorithms* on mobile phone data sets from different operators (OPAL 2018). The idea is that users can launch a predefined set of algorithms (such as home detection algorithms), which are then run behind the firewalls of the operators before returning the aggregated results back to the user. Although the project is currently still in its test phase (with pilots in Senegal and Colombia), hopes are that it could facilitate cooperation between different operators in sharing basic statistical information from their data sets (as captured by the predefined set of algorithms). If all of a country's mobile phone operators would engage in this form of cooperation, the problem of dealing with a distorted market share, for example, would be solved.

Hence, the bottom line is that although the home location problem is mainly a methodological one, the paths to address the problem are much more complex. They require the combination of collaborative, technical, methodological, institutional and strategic actions. Optimistically, we believe that official statistics offices are in a good position to (continue to) play a prominent role, because of their organisational structure, methodological knowledge and recognised institutional role within a country.

7. Conclusion

Big data sources in general, and mobile phone data in particular, create intriguing new opportunities and challenges for official statistics. Because of this, there has been a clear call for exploratory pilot projects to be carried out, as well as a trend towards critical investigation and transparency of methodologies to produce high-quality statistics. This article adhered to both of these calls in its analysis of home detection practices for non-continuous location traces, focusing mainly on mobile phone data.

Based on a critical review of literature, we discussed how existing methods to identify home locations using non-continuous location traces mainly consist of single-step approaches that deploy simple decision rules and use high-level validation only.

We argued that, given the absence of ground truth data at individual level, i) it is unclear why one-step approaches are preferred over two-step approaches that are typically used for continuous location traces; ii) no consensus in literature exists on which criteria are best to deploy when creating decision rules for home detection methods, nor has work been done to investigate the sensitivity of the results to these decision rules and criteria; and iii) the trustworthiness of high-level validation and its added value to the home detection practices are questionable at best.

By deploying five algorithms with simple decision rules to a large French CDR data set, we demonstrated several of the problems. At individual level, we found home detection methods to be rather sensitive to criteria choice, with pair comparison of different home detection algorithms resulting in different identified homes for up to 40% of users. When looking at high-level validation, we found that five different home detection algorithms performed in a similar range (34°–38°) with a similar sensitivity to the time period and the duration for which the mobile phone data was collected. Even though we found that the

sensitivity to time and the differences between different HDA algorithms does not seem large when expressed in CSM values, we showed how small changes to CSM values influence substantive and nationwide changes in the spatial patterns of population counts.

Our most noteworthy finding is the magnitude of the mismatch (the *35°-gap*) between population counts constructed from mobile phone-based data on home location and a validation data set based on census data. This large mismatch is indicative of the severity of the home location problem and challenges the validity of single-step approaches in literature. In our discussion, we listed several elements that plausibly effect this mismatch but go unnoticed when only high-level validation is undertaken. We believe that these (structural) elements, such as unknown market shares and differences in mobile phone usage, need further investigation if ever home detection methodologies are to comply with official statistics' standards.

Finally, we compiled our findings, insights, and experiences into a set of specific recommendations, ranging from the collection of individual ground truth data to the testing of methods on multiple data sets. Given the nature of these recommendations and the tasks at hand, we think that it is unlikely that individual researchers, research groups, national statistics offices, or even mobile phone operators can, or will, invest in them. Therefore, we call on and support any ongoing, collaborative actions that tackle these problems, while recognising the prominent role official statistics can (continue to) play in this area.

8. References

Ahas, R., A. Aasa, A. Roose, Ü. Mark, and S. Silm. 2008. "Evaluating Passive Mobile Positioning Data for Tourism Surveys: An Estonian Case Study." *Tourism Management* 29(3): 469–486. Doi: http://dx.doi.org/10.1016/j.tourman.2007.05.014.

Ahas, R., S. Silm, O. Järv, E. Saluveer, and M. Tiru. 2010. "Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones." *Journal of Urban Technology* 17(1): 3–27. Doi: http://dx.doi.org/10.1080/10630731003597306.

ARCEP. 2008. "Le Suivi Des Indicateurs Mobiles – Les Chiffres Au 31 Décembre 2007 (Publication Le 4 Février 2008)." Available at: http://www.arcep.fr/index.php?id=9545 (Last accessed February 2018).

Ashbrook, D. and T. Starner. 2003. "Using GPS to Learn Significant Locations and Predict Movement across Multiple Users." *Personal and Ubiquitous Computing* 7(5): 275–286. Doi: http://dx.doi.org/10.1007/s00779-003-0240-0.

Baldacci, E., D. Buono, G. Kapetanos, S. Krische, M. Marcellino, G. Mazzi, and F. Papailias. 2016. "Big Data and Macroeconomic Nowcasting: From Data Access to Modelling." Luxembourg: Eurostat. Doi: http://dx.doi.org/10.2785/360587.

Blondel, V.D., M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. 2012. "Data for Development: The D4D Challenge on Mobile Phone Data." *arXiv:1210.0137*. Available at: http://arxiv.org/abs/1210.0137 (Last accessed February 2018).

Blondel, V.D., A. Decuyper, and G. Krings. 2015. "A Survey of Results on Mobile Phone Datasets Analysis." *EPJ Data Science* 4(10): 1–57. Doi: http://dx.doi.org/10.1140/epjds/s13688-015-0046-0.

Blumenstock, J.E. 2012. "Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda." *Information Technology for Development* 18(2): 107–125. Doi: http://dx.doi.org/10.1080/02681102.2011.643209.

Bojic, I., E. Massaro, A. Belyi, S. Sobolevsky, and C. Ratti. 2015. "Choosing the Right Home Location Definition Method for the given Dataset." In *Social Informatics – 7th International Conference, SocInfo 2015, Beijing, China, December 9–12, 2015, Proceedings*, edited by Tie-Yan Liu, Christie Napa Scollon, and Wenwu Zhu, 9471: 194–208. Bejing: Springer. Doi: http://dx.doi.org/10.1007/978-3-319-27433-1_14.

Calabrese, F., L. Ferrari, and V.D. Blondel. 2014. "Urban Sensing Using Mobile Phone Network Data: A Survey of Research." *ACM Computing Surveys* 47(2): 1–20. Doi: http://dx.doi.org/10.1145/2655691.

Calabrese, F., G. Di Lorenzo, L. Liu, and C. Ratti. 2011. "Estimating Origin-Destination Flows Using Mobile Phone Location Data." *IEEE Pervasive Computing* 10(4): 36–44. Doi: http://dx.doi.org/10.1109/MPRV.2011.41.

Chen, C., L. Bian, and J. Ma. 2014. "From Traces to Trajectories: How Well Can We Guess Activity Locations from Mobile Phone Traces?" *Transportation Research Part C: Emerging Technologies* 46: 326–337. Doi: http://dx.doi.org/10.1016/j.trc.2014.07.001.

Csáji, B.C., A. Browet, V.A. Traag, J.C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V.D. Blondel. 2013. "Exploring the Mobility of Mobile Phone Users." *Physica A: Statistical Mechanics and Its Applications* 392(6): 1459–1473. Doi: http://dx.doi.org/10.1016/j.physa.2012.11.040.

Daas, P.J.H., M.J. Puts, B. Buelens, and P.A.M. van den Hurk. 2015. "Big Data as a Source for Official Statistics." *Journal of Official Statistics* 31(2): 249–262. Doi: http://dx.doi.org/10.1515/JOS-2015-0016.

de Montjoye, Y.-A., Z. Smoreda, R. Trinquart, C. Ziemlicki, and V.D. Blondel. 2014. "D4D-Senegal: The Second Mobile Phone Data for Development Challenge." *arXiv:1407.4885*. Available at: http://arxiv.org/abs/1407.4885 (Last accessed February 2018).

Deville, P., C. Linard, S. Martin, M. Gilbert, F.R. Stevens, A.E. Gaughan, V.D. Blondel, and A.J. Tatem. 2014. "Dynamic Population Mapping Using Mobile Phone Data." *Proceedings of the National Academy of Sciences* 111(45): 15888–15893. Doi: http://dx.doi.org/10.1073/pnas.1408439111.

Eurostat. 2014. "ESS Big Data Action Plan and Roadmap 1.0. Approved by the 22nd Meeting of the European Statistical System Committee." Available at: https://ec.europa.eu/eurostat/cros/content/ess-Big-DataAction-Plan-and-Roadmap-10_en (Last accessed February 2018).

ESSnet Big Data. 2018. "ESSnet Big Data." Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data (Last accessed February 2018).

Frias-Martinez, V. and J. Virseda. 2012. "On the Relationship between Socio-Economic Factors and Cell Phone Usage." In Proceedings of the Fifth International Conference on Information and Communication Technologies and Development, March 12–15, 2015. 76–84. New York, NY: ACM Press. Doi: http://dx.doi.org/10.1145/2160673.2160684.

Frias-Martinez, V., V. Jesus, A. Rubio, and E. Frias-Martinez. 2010. "Towards Large Scale Technology Impact Analyses." In Proceedings of the 4th ACM/IEEE

International Conference on Information and Communication Technologies and Development – ICTD '10, December 13–16, 2010. 1–10. New York, NY: ACM Press. Doi: http://dx.doi.org/10.1145/2369220.2369230.

Getis, A. and J.K. Ord. 1992. "The Analysis of Spatial Association by Use of Distance Statistics." *Geographical Analysis* 24(3): 189–206. Doi: http://dx.doi.org/10.1111/j.1538-4632.1992.tb00261.x.

Giannotti, F., D. Pedreschi, A. Pentland, P. Lukowicz, D. Kossmann, J. Crowley, and D. Helbing. 2012. "A Planetary Nervous System for Social Mining and Collective Awareness." *The European Physical Journal Special Topics* 214(1): 49–75. Doi: http://dx.doi.org/10.1140/epjst/e2012-01688-9.

Glasson, M., J. Trepanier, V. Patruno, P. Daas, M. Skaliotis, and A. Khan. 2013. "What Does 'Big Data' Mean for Offical Statistics?" Available at: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614 (Last accessed February 2018).

Gong, H., C. Chen, E. Bialostozky, and C.T. Lawson. 2012. "A GPS/GIS Method for Travel Mode Detection in New York City." *Computers, Environment and Urban Systems* 36(2): 131–139. Doi: http://dx.doi.org/10.1016/j.compenvurbsys.2011.05.003.

Grauwin, S., M. Szell, S. Sobolevsky, P. Hövel, F. Simini, M. Vanhoof, Z. Smoreda, A.-L. Barabasi, and C. Ratti. 2017. "Identifying and Modelling the Structural Discontinuities of Human Interactions." *Scientific Reports* 7: 46677. Doi: http://dx.doi.org/10.1038/srep46677.

Hightower, J., S. Consolvo, A. LaMarca, I. Smith, and J. Hughes. 2005. "Learning and Recognizing the Places We Go." In *UbiComp 2005: Ubiquitous Computing*, edited by M. Beigl, S. Intille, J. Rekimoto, and H. Tokuda, 159–176. Berlin, Heidelberg: Springer Berlin Heidelberg. Doi: http://dx.doi.org/10.1007/11551201.

Isaacman, S., R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. 2011. "Identifying Important Places in People's Lives from Cellular Network Data." In *Pervasive Computing: Pervasive 2011*, edited by K. Lyons, J. Hightower, and E.M. Huang, 133–151. Berlin, Heidelberg: Springer Berlin Heidelberg. Doi: http://dx.doi.org/10.1007/978-3-642-21726-5_9.

Janzen, M., M. Vanhoof, and K.W. Axhausen. 2016. "Estimating Long-Distance Travel Demand with Mobile Phone Billing Data." In Proceedings of the 16th Swiss Transport Research Conference (STRC 2016), May 18–20, 2016. Available at: http://www.strc.ch/conferences/2016/Janzen_EtAl.pdf (Last accessed February 2018).

Janzen, M., M. Vanhoof, Z. Smoreda, and K.W. Axhausen. 2018. "Closer to the Total? Long-Distance Travel of French Mobile Phone Users." *Travel Behaviour and Sociey* 11: 31–42. Doi: http://dx.doi.org/10.1016/j.tbs.2017.21.001.

Järv, O., R. Ahas, and F. Witlox. 2014. "Understanding Monthly Variability in Human Activity Spaces: A Twelve Month Study Using Mobile Phone Call Detail Records." *Transportation Research Part C: Emerging Technologies* 38: 122–135. Doi: http://dx.doi.org/10.1016/j.trc.2013.11.003.

Karlberg, M., S. Biffignandi, P.J.H. Daas, A. Holmberg, B. Hulliger, P. Jacques, R. Lehtonen, R.T. Münnich, N. Shlomo, R. Silberman, and I. Stoop. 2015. "Preface." *Journal of Official Statistics* 31(2): 149–153. Doi: http://dx.doi.org/10.1515/jos-2015-0011.

Kung, K.S., K. Greco, S. Sobolevsky, and C. Ratti. 2014. "Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data." *PLoS ONE* 9(6): e96180. Doi: http://dx.doi.org/10.1371/journal.pone.0096180.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli. 2015. "Small Area Model-Based Estimators Using Big Data Sources." *Journal of Official Statistics* 31(2): 263–281. Doi: http://dx.doi.org/10.1515/JOS-2015-0017.

Nurmi, P. and S. Bhattacharya. 2008. "Identifying Meaningful Places: The Non-Parametric Way." In *Pervasive Computing*, edited by J. Indulska, D. Patterson, J. Rodden, and M. Ott, 111–127. Berlin: Springer Berlin.

OPAL. 2018. "The OPAL project." Available at: http://www.opalproject.org/ (Last accessed February 2018).

Pappalardo, L., M. Vanhoof, L. Gabrielli, Z. Smoreda, D. Pedreschi, and F. Giannotti. 2016. "An Analytical Framework to Nowcast Well-Being Using Mobile Phone Data." *International Journal of Data Science and Analytics* 2(1–2): 75–92. Doi: http://dx.doi.org/10.1007/s41060-016-0013-2.

Phithakkitnukoon, S., Z. Smoreda, and P. Olivier. 2012. "Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data." *PloS One* 7(6): e39253. Doi: http://dx.doi.org/10.1371/journal.pone.0039253.

Raun, J., R. Ahas, and M. Tiru. 2016. "Measuring Tourism Destinations Using Mobile Tracking Data." *Tourism Management* 57: 202–212. Doi: http://dx.doi.org/10.1016/j.tourman.2016.06.006.

Ricciato, F., P. Widhalm, M. Craglia, and F. Pantisano. 2015. "Estimating Population Density Distribution from Network-Based Mobile Phone Data." Luxembourg: Publications Office of the European Union. Doi: http://dx.doi.org/10.2788/162414.

Rubrichi, S., Z. Smoreda, and M. Musolesi. 2017. "A Comparison of Spatial-Based Targeted Disease Containment Strategies Using Mobile Phone Data." *arXiv:1210.0137*. Available at https://arxiv.org/pdf/1706.00690.pdf (Last accessed February 2018).

Sakarovitch, B., P. Givord, M.-P. de Bellefon, and M. Vanhoof. In Preparation. "Allô t'es où ? Estimer la population résidente à partir de données de téléphonie mobile, une première exploration." *Economie et Statistique/Economics and Statistics*. (Preprint available upon request to authors).

Shen, L. and P.R. Stopher. 2014. "Review of GPS Travel Survey and GPS Data-Processing Methods." *Transport Reviews* 34(3): 316–334. Doi: http://dx.doi.org/10.1080/ 01441647.2014.903530.

Sobolevsky, S., M. Szell, R. Campari, T. Couronné, Z. Smoreda, and C. Ratti. 2013. "Delineating Geographical Regions with Networks of Human Interactions in an Extensive Set of Countries." *PloS One* 8(12): e81707. Doi: http://dx.doi.org/10.1371/journal.pone.0081707.

Tizzoni, M., P. Bajardi, A. Decuyper, G.K.K. King, C.M. Schneider, V. Blondel, Z. Smoreda, M.C. Gonzalez, and V. Colizza. 2014. "On the Use of Human Mobility Proxies for Modeling Epidemics." *PLoS Computational Biology* 10(7): e1003716. Doi: http://dx.doi.org/10.1371/journal.pcbi.1003716.

Vanhoof, M., S. Combes, and M.-P. de Bellefon. 2017a. "Mining Mobile Phone Data to Detect Urban Areas." In *Proceedings of the Conference of the Italian Statistical Society*

(SIS), edited by A. Petrucci and R. Verde, 1005–1012. Florence: Firenze University Press. ISBN (online) 978-88-6453-521-0.

Vanhoof, M., L. Hendrickx, A. Puussaar, G. Verstraeten, T. Ploetz, and Z. Smoreda. 2017b. "Exploring the Use of Mobile Phones during Domestic Tourism Trips." *Netcom* 31(3/4): 335–372.

Vanhoof, M., W. Schoors, A. Van Rompaey, T. Ploetz, and Z. Smoreda. 2018. "Correcting Mobility Entropy for Regional Comparison of Individual Movement Patterns." *Journal of Urban Technology* 25(2): 27–61. Doi: http://dx.doi.org/10.1080/10630732.2018.1450593.

Wolf, J., R. Guensler, and W. Bachman. 2001. "Elimination of the Travel Diary: Experiment to Derive Trip Purpose from GPS Travel Data." *Transportation Research Record* 1768: 125–134. Doi: http://dx.doi.org/10.3141/1768-15.

Ye, J. 2011. "Cosine Similarity Measures for Intuitionistic Fuzzy Sets and their Applications." *Mathematical and Computer Modelling* 53(1–2): 91–97. Doi: http://dx.doi.org/10.1016/j.mcm.2010.07.022.

Received June 2017
Revised March 2018
Accepted April 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 961–979, http://dx.doi.org/10.2478/JOS-2018-0047

Megatrend and Intervention Impact Analyzer for Jobs: A Visualization Method for Labor Market Intelligence

Rain Opik[1], *Toomas Kirt*[2], *and Innar Liiv*[1]

This article presents a visual method for representing the complex labor market internal structure from the perspective of similar occupations based on shared skills; and a prototype tool for interacting with the visualization, together with an extended description of graph construction and the necessary data processing for linking multiple heterogeneous data sources. Since the labor market is not an isolated phenomenon and is constantly impacted by external trends and interventions, the presented method is designed to enable adding extra layers of external information. For instance, what is the impact of a megatrend or an intervention on the labor market? Which parts of the labor market are the most vulnerable to an approaching megatrend or planned intervention? A case study analyzing the labor market together with the megatrend of job automation and computerization is presented. The source code of the prototype is released as open source for repeatability.

Key words: Labor market; megatrends; big data; visualization; network theory.

1. Introduction

New approaches and tools are needed to understand the complex phenomena of the labor market (e.g., a mismatch between the jobs that job seekers desire or have qualifications for, and actual vacancies), and to analyze the different megatrends impacting the labor market, such as technological change, the future of professions, the automation and computerization of jobs, robots, urbanization, refugee crises, and so on. Megatrends are great forces in societal development that will impact business, economy, society, culture and individual people for the next 10–15 years (Mogensen et al. 2014). Every new megatrend creates the need for a new policy and every successful policy starts an intervention. Therefore, it is necessary to develop methods for visualizing and mapping the implications of megatrends and interventions.

Recent advances in artificial intelligence (LeCun et al. 2015) and automation have raised fears of a significant impact on the job market (Mitchell and Brynjolfsson 2017). For example, it was found that across the OECD countries, on average 9% of jobs are automatable (Arntz et al. 2016). On the other hand, this does not mean that certain jobs are disappearing completely, but rather that they are transformed into other industries and jobs requiring a different set of skills. As Lerman and Schmidt have found regarding the

[1] Tallinn University of Technology, Akadeemia 15A, 12618 Tallinn, Estonia. E-mails: rain.opik@gmail.com, and innar.liiv@gmail.com
[2] Statistics Estonia, Tatari 51, 10134 Tallinn, Estonia. E-mail: toomas.kirt@gmail.com

appearance of the first personal computers in the mid-seventies and in 1983, computer industry jobs in the United States have grown by almost 80% while total employment in the US manufacturing industry has increased by only 4% (Lerman and Schmidt 2005). Yet recent developments in technology affect too many industries simultaneously, potentially causing an accumulation of problems, as was the case with the year 2000 problem (Jones 1997), which required substantial investments to review and upgrade existing computer systems.

We consider the computerization of jobs to be one of the most important megatrends affecting the labor market and have therefore taken this as our case study to exemplify the application of the visual method proposed in this article. Our method and the prototype tool help to visualize the complex structure of the labor market and to link job demand and vacancy data to a published hypothesis on how susceptible different jobs are to computerization (Frey and Osborne 2016). The focus of this article is not on presenting new estimates of computerization, but on developing a visual method for making sense and better understanding the connectedness and the impact of megatrends on the labor market. The presented visualization method and the prototype tool itself are universal and could be used for different data sets of megatrends and interventions.

In this article, we propose a method for representing the complex internal structure of the labor market from the perspective of occupations that are similar based on shared skills. In addition, we have developed, and present here, a prototype tool, together with an extended description of its graph construction and the related necessary data processing for linking multiple heterogeneous data sources. The method is applied to a case study of visualizing the labor market along with external information (i.e., the jobs susceptible to computerization according to Frey and Osborne 2016) in order to understand the interplay between and the joint patterns in several data sets. The source code of the prototype is released as open source for repeatability at (Opik 2017b) and the prototype is available online at (Opik 2017a).

The article includes a detailed description of the steps needed to develop the visual method and implement the prototype tool. In Section 2, we describe the methods and data used, as well as the relations between the data sets. Section 3 provides the details of how we constructed the graph of occupations and how similarity between the occupations is defined. The general technical architecture of the prototype tool and the data processing pipeline is covered in Section 4. This section also discusses the visualization capabilities of the tool and how it can be used to reveal the demand and supply imbalance of occupations. In the final section, the limitations of the prototype tool are explained, followed by conclusions and directions for future research.

2. Methods and Data

The visual method for representing the complex labor market internal structure and the prototype for interacting with the data were developed using a hackathon approach. The word hackathon is combined from the words *hack* and *marathon*, where hack is used in the sense of exploratory and investigate programming (Briscoe and Mulligan 2014).

The main contributions of this article are based on an entry for the European Big Data Hackathon, organized by the European Commission and Eurostat (European Commission

2017b). Teams were gathered from all over Europe to compete for the best data product that combines official statistics and big data to support policymakers in pressing policy questions facing Europe. The policy question for the 2017 hackathon was: "How would you support the design of policies for reducing mismatch between jobs and skills at regional level in the EU through the use of data (European Commission 2017a)?" This article took a more general approach to focus on the interconnectivity of the labor market, the supply and demand in certain segments of the labor market (Weiling and Borghans 2001), and to develop a visual method for representing the complex labor market internal structure from the perspective of similar occupations based on shared skills.

The participants of a hackathon collaborate intensively over a short period of time, and the design of such events encourages and rewards creativity and innovation (Zukin and Papadantonakis 2017). Therefore, despite inherent limitations due to the short time frame, hackathon as a methodology provides feedback and validation mechanisms for ideas and results.

The European Big Data Hackathon 2017 had two independent panels of evaluators – a statistical panel and an industry panel – who were responsible for the evaluation of results presented by the competing teams. The statistical panel was composed of ten members ranging from policymakers with responsibilities in the domain of the policy question (i.e., employment and skills, big data and data economy), official statisticians and academia. The industry panel was composed of ten representatives from all the sponsors of the Hackathon (European Commission 2017d). The evaluation criteria were the same for both panels: relevance, methodological soundness, communication, innovative approach, and replicability (European Commission 2017a).

The methodological basis of the presented method is formed by graph theory (West 2001), a node-link representation (Ghoniem et al. 2005), analytic task taxonomy (Amar et al. 2005) and a value-driven framework for visualizations (Stasko 2014). The hackathon format and constant feedback from mentors and co-participants enabled the development of a visualization method that would maximize the number of low-level components of analytical activity (Amar et al. 2005), following guidelines to maximize the value of visualization (Stasko 2014). The method and the prototype tool were designed to support the following low-level components of analytical activity: clustering, finding anomalies, filtering, finding similarities and extrema.

2.1. Connecting Different Data Sets and Classification Systems

To connect all data sets, we needed to convert the US–based O*NET-SOC job classifier into the international system. For that purpose, we used an occupation classifications crosswalk table, which maps an O*NET-SOC occupation to a job in ISCO (Hardy et al. 2016). While jobs in ISCO are organized into a clearly defined set of groups according to the tasks and duties undertaken in the course of the job (International Labor Organization, 2008), the classifier does not explicitly provide a list of those tasks and duties in a machine-readable format.

O*NET-SOC is a taxonomy based on the Standard Occupational Classification 2010 (U.S. Bureau of Labor Statistics 2010), which defines a set of occupations across the working world (U.S. Department of Labor/Employment and Training Administration 2010).

ESCO (European Skills, Competences, Qualifications and Occupations) is a relatively new classification system (European Commission 2013) that provides occupational profiles that show the relationships between occupations, skills, competences and qualifications in an RDF (Resource Description Framework) format. It contains 619 ISCO and 2,950 ESCO occupations, with references for mapping an occupation in the ESCO system to a corresponding job in ISCO. In addition to organizing occupations, ESCO provides a hierarchy of skills and competences. This article has greatly benefited from the 65,814 relationships in the ESCO system, which connects skills to occupations.

2.2. *Different Data Sets*

The following heterogeneous data sources were combined for the visualization method:

- EURES CV and job vacancy data set (European Commission 2017c)
- ESCO classifier in RDF format (European Commission 2013)
- List of jobs susceptible to automation/computerization (Frey and Osborne 2016)
- Occupation classifications mapping table from Occupation classifications crosswalks – from O*NET-SOC to ISCO (Hardy et al. 2016)

The basis of this article is a list of jobs susceptible to automation/computerization (Frey and Osborne 2016), which outlines 702 occupations, classified in SOC (U.S. Bureau of Labor Statistics 2010), along with their probability of computerization in the near future.

For measuring the impact of computerization, we chose to use the EURES data set (European Commission 2017c), which provides insight into the jobs offered by employers and sought by jobseekers across Europe. The EURES data set consists of two main tables, one on anonymized curricula vitae (4.7 million lines) proposed by jobseekers and another on job vacancies (35 million lines) published by potential employers. The vacancy data set was extracted from the EURES database on December 2, 2016. As the organizers of the hackathon did not want to use all the data, the same job vacancies were aggregated. The CV data set included monthly downloaded CVs from the period of March 2015 to November 2016 and contained the data of 297,940 unique jobseekers. Records in the CV table are classified by ESCO occupation identifiers, but the vacancy table is classified by ISCO identifiers.

Figure 1 illustrates how all the data sets are connected.

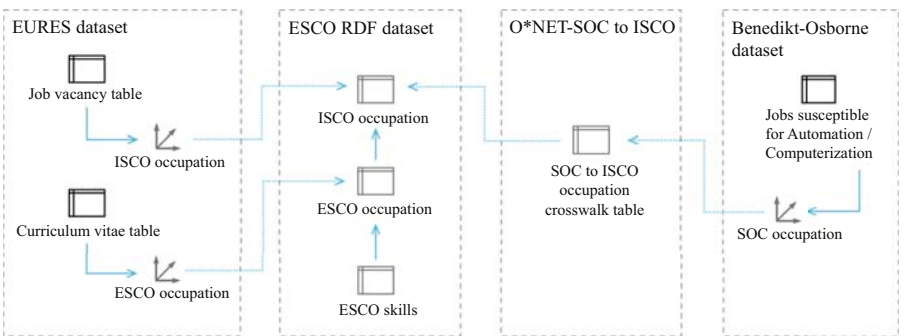

Fig. 1. Overview of data sets and their relationships.

3. Constructing a Graph

In order to visualize the complexity of a labor market, we propose to use graph theory (West 2001) to construct the node-link diagram (Ghoniem et al. 2005) in order to represent the similarity and interrelations between different occupations in the ESCO classification and hackathon data sets, according to shared skills required for a particular job. In the next subsections we will present our approach and data modelling choices for constructing the graph.

3.1. Graph of Occupations

An occupation graph is defined by two entities – node and link. A node in the graph denotes an occupation in the ESCO classifier (e.g., *bus driver*). A link is defined between two nodes (occupations) when they are similar to one another. The link in the occupation graph does not represent a match between jobs and skills, but rather a similarity of occupations based on skill information in the ESCO.

3.2. Linking Similar Occupations

The similarity between two occupations is defined as follows. For a given ESCO occupation o_1, we enumerated all the skills required for that occupation (SK_{o_1}). Then we matched all occupations that require at least one skill from SK_{o_1}, which produces a mapping from ESCO occupation o_1 to ESCO occupation o_2. We define the similarity measure as the ratio of the number of shared skills between two occupations to the number of all skills required by the first occupation (Figure 2).

To avoid ending up with a large number of skills with varying relevance, we only chose skills that were marked as *essential* for the given occupation in the ESCO classifier.

For example, let us take two occupations: *bus driver* (ESCO occupation identifier: *00cee175-1376-43fb-9f02-ba3d7a910a58*) and *private chauffeur* (*e75305db-9011-4ee0-ab62-8d41a98f807e*) and enumerate all the skills that are essential for both occupations (Table 1).

The skills in this table can be divided into three groups:

- Skill that is only required for the first occupation (e.g., *bus driver*)
- Skill that is only required for the second occupation (e.g., *private chauffeur*)
- Skill that is required for both occupations.

When we count the number of distinct skills that are required for both occupations (22 in this example) and divide it by the number of distinct skills required for the first occupation (35), we get a percentage of matching skills, which we use as a similarity measure between these two occupations.

$$similarity(o_1, o_2) = \frac{n\left(SK_{o_1} \cap SK_{o_2}\right)}{n(SK_{o_1})}$$

Fig. 2. Occupation similarity.

Table 1. A sample of essential skills for an occupation pair.

Skills required for *bus driver*	Skills required for *private chauffeur*
provide first aid	N/A
manoeuvre bus	N/A
N/A	maintain personal hygiene standards
N/A	park vehicles
drive in urban areas	drive in urban areas
keep time accurately	keep time accurately
provide information to passengers	provide information to passengers

The resulting matrix is very large, as it contains all occupation pairs that are loosely connected by a very generic, albeit essential, skill. For example, both *bus driver* and *physiotherapy assistant* have *use different communication channels* as an essential skill, which connects them in the occupation graph. However, when we calculate the skills match ratio, we get a modest 2%. In addition, the connection between these occupations does not translate into real life, as it is difficult to imagine that a person skilled in operating heavy vehicles could easily apply for a position that requires medical skills.

Not every link in the graph is important, especially when representing the graph visually. To reduce the visual clutter, we decided to prune the graph of weakly connected occupation pairs and take only the three most similar occupations for every occupation. This has also been researched in social network analysis, where the number three has been considered sufficient to represent structurally important connections, while revealing the variation of inter-node relations across the graph and allowing efficient clustering of the graph into subgroups (Burt 1984 and Merluzzi and Burt 2013). Table 2 shows an example of the pruned graph for two selected occupations and Figure 3 contains an illustration of how the graph would look.

When this algorithm is run for all occupations (e.g., *private chauffeur*), we get new links in the graph, yielding at least three links for every node (Figure 4).

3.3. Annotating Occupations With Megatrend and Supply-Demand Data

In its simplest form, a graph node contains only one attribute, which is the title of the occupation. However, we can treat the list of nodes as a data table and attach additional attributes that explain the phenomena being investigated.

Table 2. The pruned occupation graph for two occupations.

From occupation	To occupation	Skill match
bus driver	trolley bus driver	80%
bus driver	tram driver	77%
bus driver	private chauffeur	63%
cargo vehicle driver	dangerous goods driver	60%
cargo vehicle driver	bus driver	55%
cargo vehicle driver	private chauffeur	45%

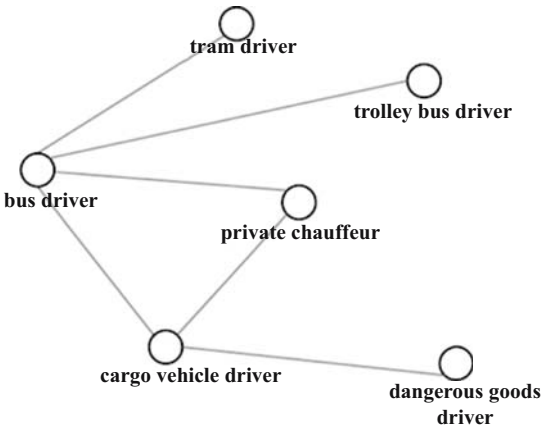

Fig. 3. The graph for two occupations.

First, we want to understand how the megatrend (automation) affects the occupation graph. The list of jobs susceptible to automation/computerization originally had SOC occupation codes. However, our occupation graph was based on occupations classified by ESCO, which can be mapped to ISCO occupation codes. We need a mapping table to link these two data sets. A mapping of ISCO to SOC (Hardy et al. 2016) is unfortunately one-to-many, which means that some ISCO occupations (e.g., *8332 – Heavy truck and lorry drivers*) are associated with several SOC occupations (*53-1031 – Driver/Sales Workers and 53-3032 – Heavy and Tractor-Trailer Truck Drivers*). As a result, they also have different probabilities for automation (0.98 and 0.79 respectively). To solve this ambiguity, we calculated two probabilities, maximum and average.

After knowing which jobs are going to be impacted, we wanted to assess how many people would be affected by this trend. Since we based our tool on the EURES CV and job vacancy data set, we could readily count the number of vacancies and the number of unique persons that have marked this occupation as their desired job. For example, based

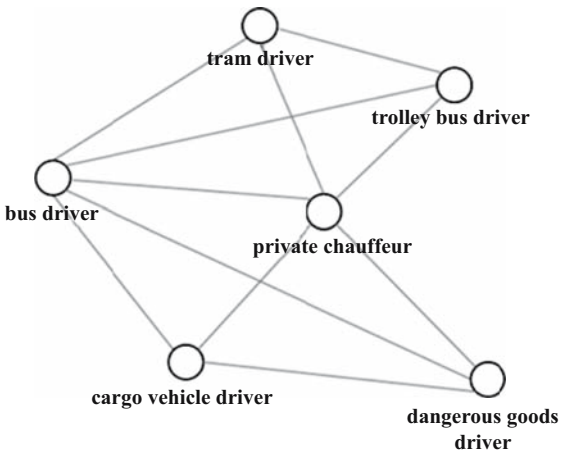

Fig. 4. A fragment of the full occupation graph.

Table 3. Node data attributes for two selected occupations.

Occupation	Prob. of automation	Vacancies total	CVs total	Vacancies in Austria	CVs in Austria	Vacancies in Belgium	CVs in Belgium	. . .
bus driver	0.89	53 936	535	1 426		1 925	5	
cargo vehicle driver	0.79	666 061	13 305	13 305		35 475	15	
. . .								

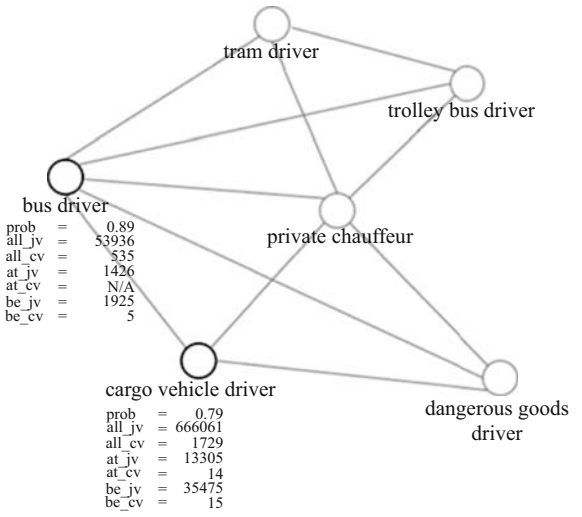

Fig. 5. An occupation graph with annotations.

on the EURES data, there are 1,925 job vacancies for *bus driver* in Belgium and five job seekers have marked *bus driver* as their desired occupation – see Table 3 for examples. Figure 5 shows the annotated nodes in a visual graph representation.

4. Visualization Tool Prototype

4.1. Technical Architecture

The majority of the integrated data sets were initially received as text files in a CSV format. After estimating the size of the main data set (EURES), which was approximately 26 million records, we decided to use Apache Hive (Thusoo et al. 2009; Huai et al. 2014) for large queries and data aggregation and PostgreSQL (Smith 2010) for more complex queries. We chose PostgreSQL as the primary database engine for data management and exploration. In such a data exploration phase, if there is a relatively small amount of data, relational databases have many benefits over specialized parallelized databases like Apache Hive. The most important advantage provided by PostgreSQL is the ease of ad-hoc querying and the expressiveness of the SQL language. For example, a simple SELECT query on a table with a couple of joins or calculating aggregates such as totals over millions of rows will be more efficient on a specialized big data backend. However, most big data query languages tend to have very limited support for more advanced operations such as subqueries or common table expressions, and the analyst is thus forced to fall back on expressing the query in a programming language. Moreover, evaluating different schema alternatives and developing a suitable data model is an inconvenient task in most big data databases, as ad-hoc schema modifications are slow and cumbersome. For that reason, we decided to perform the data exploration and schema discovery phase in PostgreSQL, and then create the final schema in Apache Hive, where we also ran the main queries for aggregating the occupation data.

The first draft of the occupation graph was drawn with the Python graph-tool (Peixoto 2014), which produced static image files. Since pre-rendered image files give a good overview of the graph, but lack in providing effective methods for filtering and obtaining details, we decided to implement the visualizer in d3.js (Bostock et al. 2011). The d3.js application can be viewed in a modern web browser without any additional dependencies.

The visualizer tool was designed to run without a server backend or online connection to a database. This makes it easy to host the tool on a static website (like GitHub) without any running costs. The final table of similar occupations and the list of all nodes in the graph were exported to text files so they can be served statically.

We used Amazon Web Services to host the infrastructure in a cloud environment. This gave us the flexibility to easily create a computational environment capable of processing the hackathon data sets and dispose of the resources after the computation is completed in order to minimize running costs. Amazon Relational Database Services (RDS) provides various SQL database engines such as MySQL or PostgreSQL, and Amazon Elastic MapReduce (EMR) facilitates running Hadoop workloads with preconfigured big data frameworks such as Apache HBase, Spark or Hive. However, due to our decision to prioritize open-source components, the backend can also be set up in an on-premise datacenter, without relying on cloud service providers.

4.2. *Data Processing*

We built the occupation graph with PostgreSQL queries. The resulting graph was stored in two denormalized tables: *node* containing a list of all occupations and their metadata, and *link* containing connections between similar occupations.

During the construction of the node table, we observed that the amount of data in the EURES data set makes direct querying inefficient – counting the number of unique job seekers and vacancies by occupations and different countries was the most time-consuming task. Since this type of workload is more suitable for databases using the MapReduce programming model, we used Apache Hive to calculate the country-based aggregates for each occupation. This resulted in a tenfold increase in query performance.

The ESCO classifier was originally presented in an RDF format, which is a list of semantic triples in the subject-predicate-object format. While specialized graph databases have support querying data in the triplet format (e.g., SPARQL or Gremlin), writing queries that join data across SQL and a graph database is very inefficient in terms of performance. Therefore, we decided to parse the RDF file and convert it to a relational structure suitable for SQL.

The serverless design of the visualizer mandates that the data files are accessible without a database. We have used flat CSV files for feeding data to the visualizer. Figure 6 shows the data processing pipeline.

4.3. *Calculating Graph Layout*

Our experience with d3.js has shown that real-time calculation of graph layouts (i.e., how to position nodes on a two-dimensional plane) may be slow for graphs with a non-trivial structure. Our occupation graph has 2,950 nodes and 8,838 links and after some experimentation we decided to pre-calculate the positions of the graph nodes. We used the

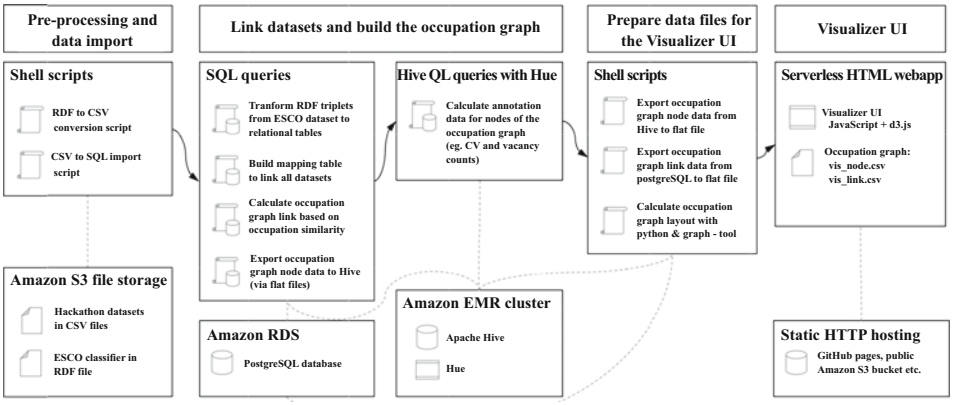

Fig. 6. The data processing pipeline.

SFDP layout algorithm (Hu 2005) from the graph-tool (Peixoto 2014) for calculating the position of the nodes and re-indexing node identifiers to a format that is suitable for a visualizer.

Besides performance gains, this also has a second benefit – the visualization can be easily shared with fellow analysts. Most graph layout algorithms are non-deterministic in nature due to random initialization and produce a different layout after each run. By using

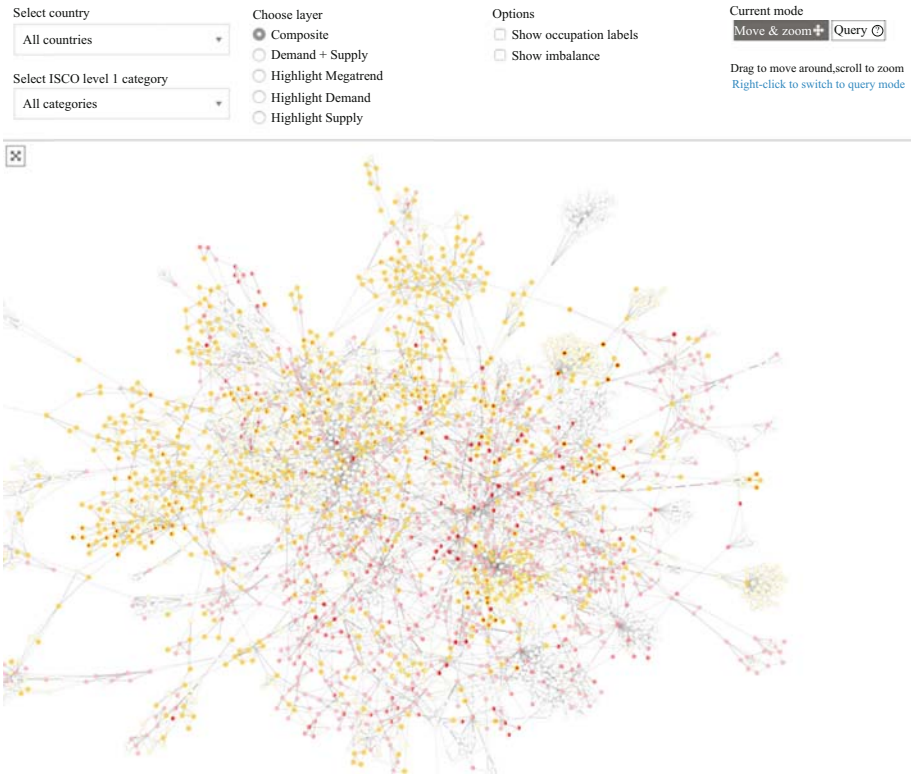

Fig. 7. The visualizer prototype.

pre-calculated node coordinates, we can ensure that the visualizer produces output that looks exactly the same in every browser given the same set of input parameters.

4.4. Visualizer UI

The user interface for the visualizer is built with d3.js, which renders a zoomable and scrollable SVG document for browsing the graph online. The prototype application (Figure 7) can be viewed in a modern web browser, preferably Google Chrome.

4.5. Prototype Interaction Models

Initially the visualizer displays the complete occupation graph. To facilitate the possibility of obtaining more detailed information, the application has two modes:

- Move and zoom mode – an analyst can click and drag the mouse to move around the graph. Scrolling the mouse wheel zooms in and out.
- Query mode – when an analyst moves the mouse cursor over a node, a small tooltip with demand and supply numbers will be displayed. Hovering also highlights connected jobs and fades out the rest of the graph. A click on the right mouse button allows for switching between Move and Query modes. See Figure 8 for an example of a query mode activated for the bus driver node.

The full occupation graph has enough nodes to appear as an impenetrable hairball when zoomed in. To reduce the clutter, we have added a filter tool to show only a relevant subset

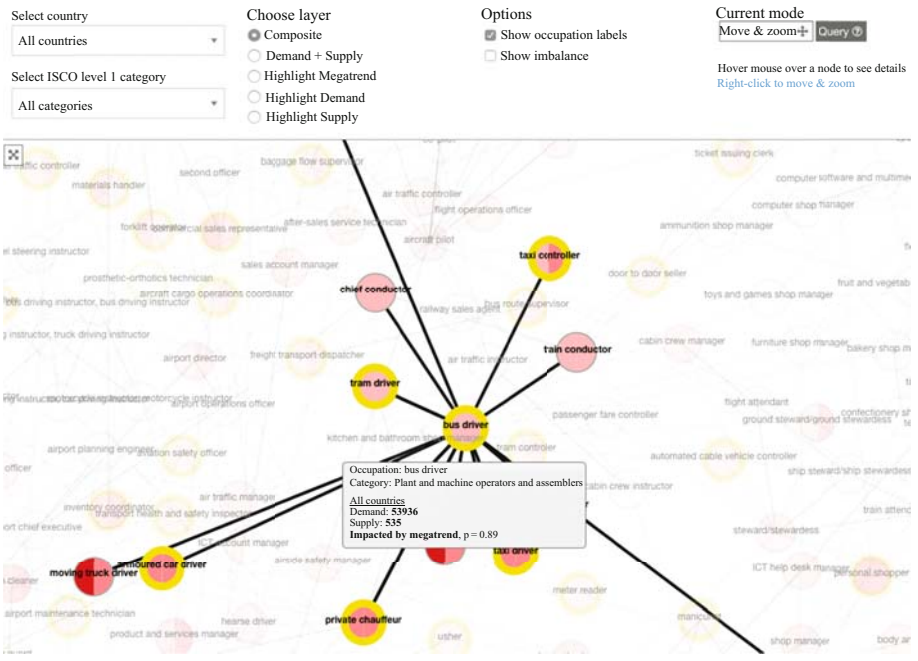

Fig. 8. A query mode is activated for a node.

of occupations. The filter tool allows an analyst to choose an ISCO level 1 occupation category (e.g., *Plant and machine operators and assemblers*) and render only these occupations that have this categorization while hiding the rest of the graph. See Figure 9 and Figure 10 for the effect of the filter tool.

The filter tool is effective due to the nature of the data set – since nodes represent ESCO occupations which can be linked to hierarchical ISCO classification, the top level of the ISCO classifier produces a meaningful subset of the graph with the same semantics.

4.6. Visualizing Node Metadata

We have used color to encode various metadata attributes that were attached to graph nodes. The visualizer supports several types for color coding – we call them layers.

- Composite layer. The left half of a node is colored by the number of vacancies available for that job (demand). Starting with white (no vacancies) to light pink (low demand) and ending with red color denoting high demand. The right half is colored by the number of job seekers who have listed a particular job in their desired job list. Color gradation is similar to the left half. Additionally, the node is marked with a yellow halo when the relevant job is affected by the Megatrend, that is, the job in the

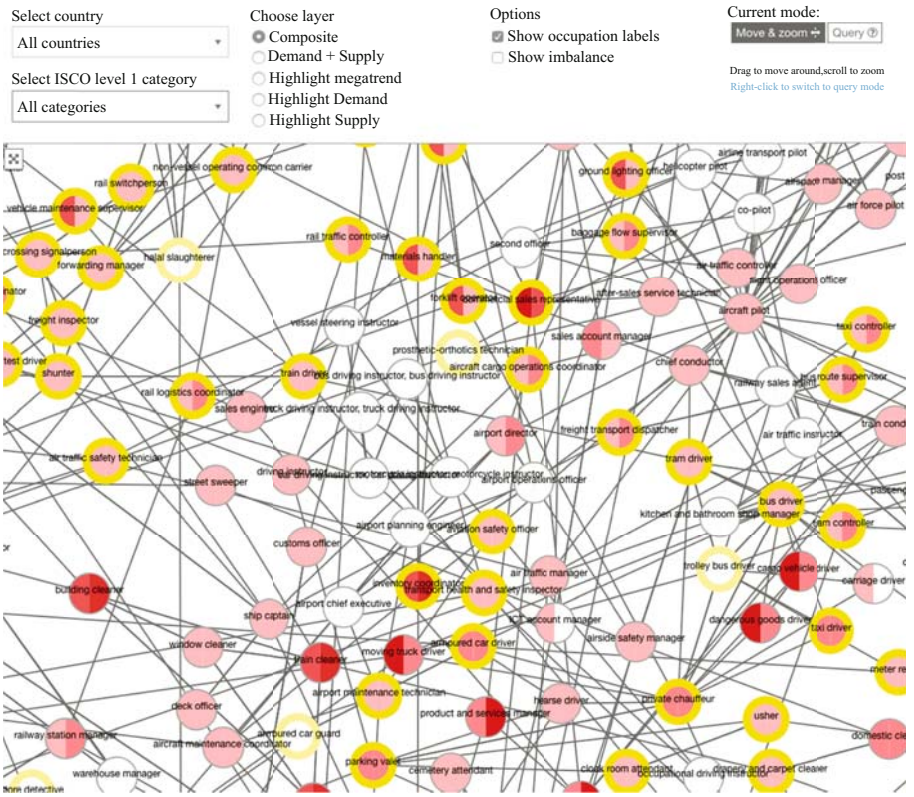

Fig. 9. A close-up of the occupation graph.

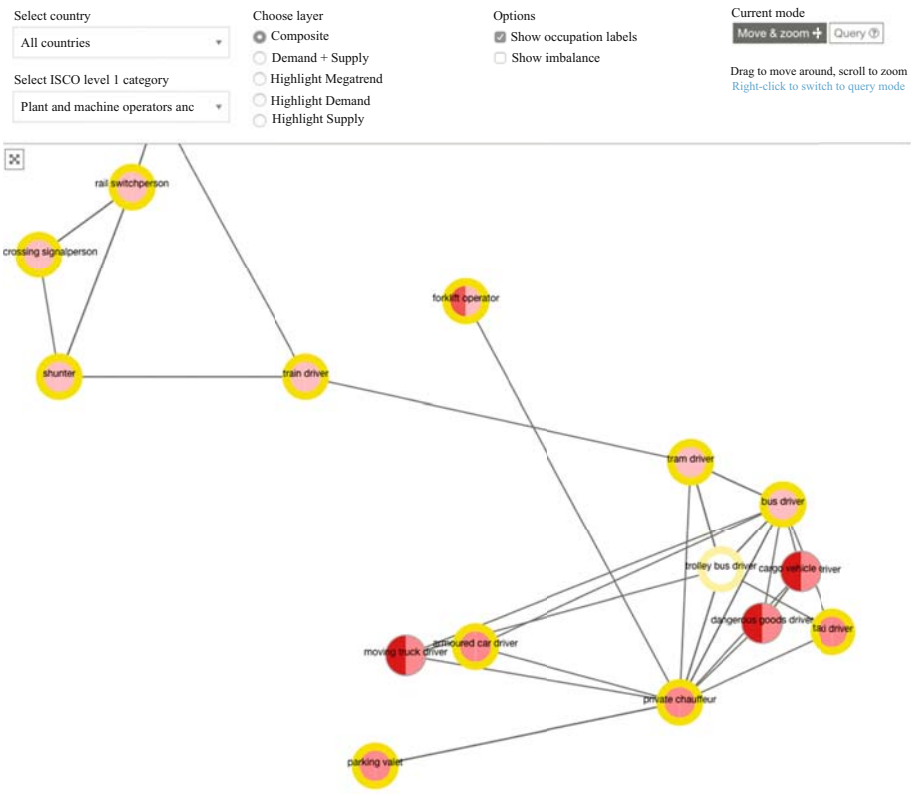

Fig. 10. The filter tool is applied.

list of jobs susceptible to automation/computerization. When the yellow halo is not present, the occupation is unaffected by the Megatrend.

- Demand and Supply layer. This is essentially the same visualization as Composite, except that the Megatrend markers (the yellow halo around the nodes) are not drawn.
- Highlight Megatrend layer. The node is colored red when the job is affected by the Megatrend. Non-affected jobs are colored white.
- Highlight Supply layer. The node is colored red when at least one job seeker has listed this job in their desired job list. White nodes denote jobs that no one desires.
- Highlight Demand layer. The node is colored red when a particular job is listed in at least one job vacancy. White nodes denote jobs with no demand.

4.7. Demand and Supply Imbalance

Color values for the left and right half (demand and supply) are normalized separately due to a huge imbalance in the EURES data. For example, some countries have no job seekers in EURES, while showing lots of vacancies and vice versa.

To overcome this issue, we implemented an alternative way for coloring nodes (refer to Subsection 4.6 for details). The default mode (i.e., *Show imbalance unchecked*) calculates the saturation ("brightness" of the red color) of the left and the right half of the node on the same scale. This helps to identify the most sought-after jobs – the analyst needs to look for

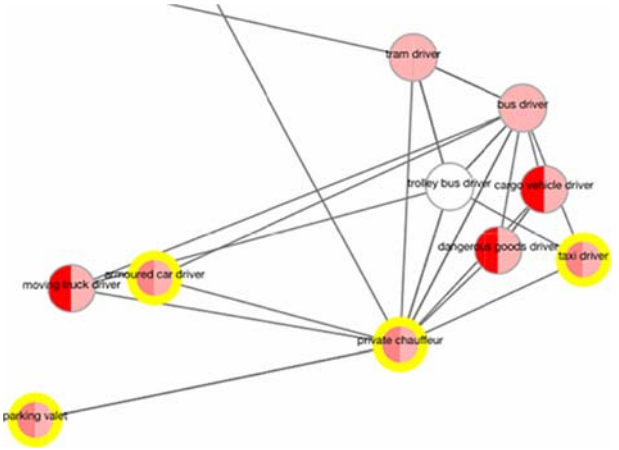

Fig. 11. Default coluring of nodes.

nodes with a bright red left half. Similarly, jobs with the largest supply of job seekers have a bright right half. For example, on Figure 11, the occupation *Private chauffeur* has a total demand (across all EU countries) of 215,677 and a total supply of 1,674. When these counts are normalized across the whole graph, both numbers are assigned the same color.

Enabling the *Show imbalance* mode normalizes both colors on the same scale. This visualizes the imbalance – when the left half of the node is a brighter red compared to the right, the job has unsatisfied demand. Conversely, a brighter right half marks jobs with an excessive number of job seekers. Figure 12 illustrates this.

Note: the EURES data contains huge discrepancies between supply and demand across different countries. Some countries have no job seekers in EURES while showing lots of vacancies and vice versa. Therefore, the *Show imbalance* mode may reveal only the extremities.

5. Limitations

Currently, the position of the graph nodes is determined by the graph layout algorithm, which generally tends to improve aesthetics by optimizing certain criteria that reduce visual clutter, for example, by minimizing the number of crossings, ensuring the even

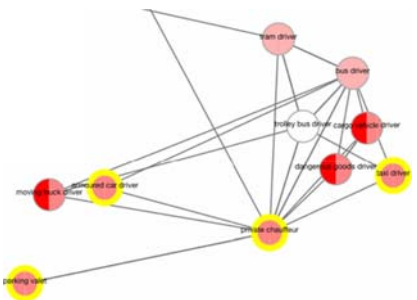

Fig. 12. Show imbalance mode activated.

distribution of nodes, and so on (Battista 1998). For the purposes of analysis, it would be beneficial to utilize the position of the nodes to encode the semantics of the underlying structure. Since the ESCO classifier contains a four-level category structure that effectively clusters the graph, we were able to use this information for the initial positioning of the nodes, providing additional visual cues to the analyst navigating the graph.

The visualizer does not currently distinguish strong links (i.e., high similarity between occupations) from weak ones. Exploring the options for visual representation of similarity (e.g., coloring intra- and inter-category links; different encoding for a link, whether it is based on cross-sector or sector-specific skills; a cut-off point for weak links) and choosing the optimal visualization remains an exercise for the future.

The tool does not, at present, provide a straightforward means of examining the common skills that were the basis for making a connection between occupations. Providing information about common skills inside the visualizer, as shown in Table 1, facilitates understanding why unusual relationships are present in the occupation graph.

6. Conclusion

Rapid changes in society in the information age can pose challenges to the national statistics offices. Registering time series might not be sufficient to face those challenges, and the need is likely to arise for different approaches, which can indicate potential future changes, as well as for using new data sources for this purpose. As novel big data sources are often heterogeneous, there are numerous steps in between to link them, many of which are not known to people entering into the space of big data.

The main contribution of this article is to provide a novel visual representation of all occupations in the labor market, which makes it possible to see similarities and patterns; and the rendering of information about job supply and demand along with external information about the trends on that same visualization. A prototype tool with the necessary data processing is proposed for interacting with the visual representation.

Our method is universal and allows for adding extra layers of information. For instance, what is the impact of a megatrend or an intervention on the labor market? Which segments of the labor market are the most vulnerable to an approaching megatrend or planned intervention?

The computerization of jobs as a megatrend was chosen as an example for using our method. What occupations are the most susceptible to computerization? Is it potentially going to impact labor market demand and skills mismatch, or further increase unemployment? These are only a few of the questions for the exploratory data analysis approach presented in this article. The real value of visualization methods and different visualizations lies in their ability to spur on and discover insights and/or insightful questions about the data (Stasko 2014).

In addition to addressing the limitations highlighted in Section 5, several interesting and different directions for future research are opened up by this work. As a lot of pre-processing of data was done manually here, a production version of such a tool should consider integrating existing data wrangling (Kandel et al. 2011) tools to optimize the time spent on introducing new data sets or scenarios. Frey and Osborne have recently published an opinion (Frey and Osborne 2018) on revisiting their seminal study (Frey and Osborne

2016), which clearly demonstrates that more research and different scenarios for automation and the future of work will be available. Research groups with different assumptions, approaches and methodologies make it ever more difficult to compare scenarios. Investigating the various options of representing several different scenarios using the visual method presented in this article could help researchers and policymakers to grasp the different results. In addition, conducting user studies with policymakers could help enhance the visual method, its interaction and the prototype tool more generally, and researchers could get valuable novel insights from policymakers about computerization or other megatrends. Finding interesting insights from data is always a dialogue and enabling policymakers to visually navigate the complex internal structure of the labor market can introduce wholly new forms of knowledge transfer.

7. References

Amar, R., J. Eagan, and J. Stasko. 2005. "Low-level components of analytic activity in information visualization." IEEE Symposium on Information Visualization, 2005. INFOVIS 2005, October 23–25 2005. 111–117. Minneapolis, MN, U.S.A. IEEE.

Arntz, M., T. Gregory, and U. Zierahn. 2016. "The Risk of Automation for Jobs in OECD Countries." *OECD Social, Employment and Migration Working Papers*. Doi: http://dx.doi.org/10.1787/5jlz9h56dvq7-en.

Battista, G.D., P. Eades, R. Tamassia, and I.G. Tollis. 1998. *Graph Drawing: Algorithms for the Visualization of Graphs*. Englewood Cliffs, NJ: Prentice Hall.

Bostock, M., V. Ogievetsky, and J. Heer. 2011. "D3: Data-Driven Documents." *IEEE transactions on visualization and computer graphics (Proc. InfoVis)*. Available at: http://vis.stanford.edu/papers/d3/ (accessed April 2017).

Briscoe, G. and C. Mulligan. 2014. "Digital Innovation": The Hackathon Phenomenon. Creative works London Working Paper. Queen Mary University of London. Available at: http://qmro.qmul.ac.uk/jspui/handle/123456789/7682 (accessed December 2017).

Burt, R.S. 1984. "Network items in the general social survey." *Social Networks* 6: 293–339. Doi: http://dx.doi.org/10.1016/0378-8733(84)90007-8.

European Commission. 2013. *ESCO – European Classification of Skills/Competences, Qualifications and Occupations – The first public release*. Luxembourg: Publications Office of the European Union. Available at: http://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=7676 (accessed April 2017).

European Commission. 2017a. *Description of the European Big Data Hackathon*. Eurostat. Available at: https://ec.europa.eu/eurostat/cros/system/files/european_big_data_hackathon_-_description.pdf (accessed December 2017).

European Commission. 2017b. *European Big Data Hackathon*. Eurostat. Available at: https://ec.europa.eu/eurostat/cros/EU-BD-Hackathon_en. (accessed December 2017).

European Commission. 2017c. *Hackathon Data Catalogue*. Eurostat. Available at: https://ec.europa.eu/eurostat/cros/content/hackathon-data-catalogue_en (accessed December 2017).

European Commission. 2017d. *Panel of evaluators*. Eurostat. Available at: https://ec.europa.eu/eurostat/cros/content/panel-evaluators_en (accessed December 2017).

Frey, C.B. and M.A. Osborne. 2016. "The future of employment: How susceptible are jobs to computerisation?" *Technological Forecasting and Social Change* 114: 254–280. Doi: http://dx.doi.org/10.1016/j.techfore.2016.08.019.

Frey, C.B. and M.A. Osborne. 2018. *Automation and the Future of Work – Understanding the Numbers*. Available at: https://www.oxfordmartin.ox.ac.uk/opinion/view/404. (accessed April 2018).

Ghoniem, M., J.D. Fekete, and P. Castagliola. 2005. "On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis." *Information Visualization* 4(2): 114–135. Doi: http://dx.doi.org/10.1057/palgrave.ivs.9500092.

Hardy, W., D. Autor, and D. Acemoglu. 2016. "Occupation classifications crosswalks – from O*NET- SOC to ISCO." [Online]. Available at: http://ibs.org.pl/en/resources/occupation-classifications-crosswalks-from-onet-soc-to-isco/. (accessed April 2017).

Hu, Y. 2005. "Efficient, high-quality force-directed graph drawing." *Mathematica Journal* 10(1): 37–71. Available at: http://www.mathematica-journal.com/issue/v10i1/graph_draw.html (accessed October 2018).

Huai, Y., A. Chauhan, A. Gates, G. Hagleitner, E.N. Hanson, O. O'Malley, J. Pandey, Y. Yuan, R. Lee, and X. Zhang. 2014. "Major technical advancements in apache hive." In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, June 2014*: 1235–1246. New York. NY, U.S.A., ACM.

International Labour Organization. 2008. *ISCO – International Standard Classification of Occupations*. Switzerland Geneva: International Labour Office. Available at: http://www.ilo.org/public/english/bureau/stat/isco/index.htm (accessed April 2017).

Jones, C. 1997. *The Year 2000 Software Problem: Quantifying the Costs and Assessing the Consequences*. ACM Press/Addison-Wesley Publishing Co. Available at: https://dl.acm.org/citation.cfm?id=267961 (accessed December 2017).

Kandel, S., J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N.H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. 2011. "Research directions in data wrangling: Visualizations and transformations for usable and credible data." *Information Visualization* 10(4) : 271–288. Doi: http://dx.doi.org/10.1177/1473871611415994.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep learning." *Nature* 521(7553): 436–444. Doi: http://dx.doi.org/10.1038/nature14539.

Lerman, R.I. and S.R. Schmidt. 2005. *Trends and challenges for work in the 21st century*. Future Work, US Dept. of Labor, The Urban Institute, Washington DC. Available at: https://www.dol.gov/oasam/programs/history/herman/reports/futurework/report.htm (accessed April 2017).

Merluzzi, J. and R.S. Burt. 2013. "How many names are enough? Identifying network effects with the least set of listed contacts." *Social Networks* 35(3): 331–337. Doi: http://dx.doi.org/10.1016/j.socnet.2013.03.004.

Mitchell, T. and E. Brynjolfsson. 2017. "Track how technology is transforming work." *Nature* 544(7650): 290. DOI: http://dx.doi.org/10.1038/544290a.

Mogensen, K.A., K. Brown, A.D. Baedkel, K. Gu, M. Fert-Malka, N.T. Hemmingsen, L. Borgstrom-Hansen, C.S. Petersen, and O. Denysenko. 2014. *Trends for Tomorrow*. Member's Report 4/2014. Copenhagen Institute for Futures Studies. Available at: https://cifs.dk/publications/members-reports/ (accessed October 2018).

Opik, R. 2017a. "The prototype." Available at: https://rainopik.github.io/eubdhack-megatrend/ (accessed October 2018).

Opik, R. 2017b. "The source code of the prototype." Available at: https://github.com/rainopik/eubdhack-megatrend/ (accessed October 2018).

Peixoto, T.P. 2014. "The graph-tool python library." *figshare*. Doi: http://dx.doi.org/10.6084/m9.figshare.1164194.

Smith, G. 2010. *PostgreSQL 9.0: High Performance*. Packt Publishing Ltd.

Stasko, J. 2014. "Value-driven evaluation of visualizations." In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. (pp. 46–53). ACM.

Thusoo, A., J.S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy. 2009. "Hive: a warehousing solution over a map-reduce framework." *Proceedings of the VLDB Endowment* 2(2): 1626–1629. Doi: http://dx.doi.org/10.14778/1687553.1687609.

West, D.B. 2001. *Introduction to Graph Theory*. New York: Prentice Hall.

Wieling, M., and L. Borghans. 2001. "Descrepancies between supply and demand and adjustment processes in the labour market." *Labour* 15(1): 33–56. Doi: http://dx.doi.org/10.1111/1467-9914.00154.

U.S. Bureau of Labor Statistics. 2010. *Standard Occupational Classification*. Washington DC: Bureau of Labor Statistics. Available at: https://www.bls.gov/soc/. (accessed April 2018).

U.S. Department of Labor/Employment and Training Administration. 2010. *The O*NET-SOC Taxonomy*. Available at: https://www.onetcenter.org/taxonomy.html (accessed December 2017).

Zukin, S. and M. Papadantonakis. 2017. "Hackathons as Co-optation Ritual: Socializing Workers and Institutionalizing Innovation in the 'New' Economy." In *Precarious Work*. (pp. 157–181). Emerald Publishing Limited.

Received June 2017
Revised April 2018
Accepted May 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 981–1010, http://dx.doi.org/10.2478/JOS-2018-0048

Augmenting Statistical Data Dissemination by Short Quantified Sentences of Natural Language

Miroslav Hudec[1], *Erika Bednárová*[1], *and Andreas Holzinger*[2]

Data from National Statistical Institutes is generally considered an important source of credible evidence for a variety of users. Summarization and dissemination via traditional methods is a convenient approach for providing this evidence. However, this is usually comprehensible only for users with a considerable level of statistical literacy. A promising alternative lies in augmenting the summarization linguistically. Less statistically literate users (e.g., domain experts and the general public), as well as disabled people can benefit from such a summarization. This article studies the potential of summaries expressed in short quantified sentences. Summaries including, for example, "*most visits from remote countries are of a short duration*" can be immediately understood by diverse users. Linguistic summaries are not intended to replace existing dissemination approaches, but can augment them by providing alternatives for the benefit of diverse users of official statistics. Linguistic summarization can be achieved via mathematical formalization of linguistic terms and relative quantifiers by fuzzy sets. To avoid summaries based on outliers or data with low coverage, a quality criterion is applied. The concept based on linguistic summaries is demonstrated on test interfaces, interpreting summaries from real municipal statistical data. The article identifies a number of further research opportunities, and demonstrates ways to explore those.

Key words: Linguistic summaries; linguistic quantifiers; fuzzy sets; database queries; user interface.

1. Introduction

Businesses, public administrations, researchers, journalists, and the general public are increasingly interested in data and information that describe various aspects of our society. National Statistical Institutes (NSIs) are sources that are generally regarded as credible, due to their profound and reliable methodologies for data collection, production and dissemination, explained through the Generic Statistical Information Model (e.g., GSIM 2013; Scanu and Casagrande 2016). Various approaches to data dissemination have already been developed; applications such as Contestina (Zottoli et al. 2017) provide interfaces for creating questions, interpreting answers in tables, graphs and on maps, and storytelling based on specific parameters chosen by the users. However, all these approaches, although powerful, often require up-to-date information and communication technologies (web browser versions and fast internet connection running on up-to-date hardware) which are still not available to everybody. Consequently, the dissemination

[1] Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia. Emails: miroslav.hudec@euba.sk and bednarovaa.erika@gmail.com
[2] Holzinger Group HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Auenbruggerplatz 2, 8036 Graz, Austria. Email: andreas.holzinger@medunigraz.at

should also be accessible to those who rely on "low-tech" information and communication technologies, on a variety of platforms.

While larger businesses prefer raw data and analyze them by their own methods, smaller businesses rather look for information and prefer simple presentations and short descriptions (Bavdaž 2011). One reason might be that smaller businesses usually cannot afford specialists in data mining and statistics, or expensive consultancy, and their statistical and computational skills may not be sufficient to effectively interpret the data produced by NSIs. The same might hold true for some journalists searching for statistical information to support their articles (while data journalists would prefer access to the data). Disabled people frequently have to overcome obstacles when searching for data and information on websites: blind people need content that can be expressed by sound or voice instead of graphs and tables; people who are dyslexic or have cognitive impairments may benefit from the use of simpler language (Disability Rights Commission 2004; Heimgärtner et al. 2008).

Inspiration for an alternative approach emerged from the following five observations: (i) graphical interpretation is a valuable way of summarization; however, it is not always effective (Disability Rights Commission 2004; Lesot et al. 2016); (ii) users (e.g., small businesses) are often interested in summarized information rather than data (Bavdaž 2011); (iii) summaries should not be as terse as means (Yager et al. 1990), and should hold for any data type and distribution; (iv) a natural way for humans to communicate, compute and conclude is natural language (Zadeh 2001); and, (v) existing approaches in data dissemination are typically based on precise (crisp) conditions or questions, for example, "*find towns that accommodated more than 1,000 visitors*". The alternative is summaries of short quantified sentences of natural language, or Linguistic Summaries (LSs). For example, we can express: "*the mean value is 235.4 with a standard deviation of 123.3*", or linguistically: "*few observations are near the mean value*". The linguistic case clearly illustrates that the mean value is not a sufficiently representative characteristic in this example. The other option is interpreting the summary between attributes, for example, "*most visits from remote countries are of a short duration*". Such a summary, although neither based on traditional mathematical methods nor on visualisation, contains very valuable information for accommodation providers, marketers, journalists and local authorities. In addition, linguistic summaries can be interpreted by text-to-speech synthesis systems. They are especially useful whenever the users' visual attention is focused on something else (Arguelles and Triviño 2013), or for the aforementioned disabled and/or elderly people (Holzinger 2002). Kacprzyk and Zadrożny (2005, 282) recognized the benefits of linguistic summaries by emphasizing that "*Data summarization is one of [the] basic capabilities that is now needed by any 'intelligent system' that is meant to operate in real life*". People ask, evaluate and conclude by linguistic terms, which are vague, but on the other hand very effective. Here, "vague" means nonsharp boundaries of concepts (linguistic terms) expressed by fuzzy sets, whereas "effective" means that we distinguish elements by intensity of belonging to a set without adding further properties. This observation led Zadeh (2001) to formalize the concept known as *computing with words*.

In this article, we provide a more theoretical view on dissemination by linguistic summaries for the users of official statistics. The "test interfaces" have been developed

merely to illustrate our idea, to demonstrate applicability and to show procedures for calculating and interpreting linguistic summaries from real-world data.

The remainder of this article is organized as follows: Section 2 introduces linguistically quantified sentences and a theoretical basis consisting of related works and our observations, all of which is necessary for the subsequent sections. Section 3 is dedicated to dissemination through LSs, supported by illustrations and examples on data from the Municipal Statistics Database of the Slovak Republic. Section 4 is focused on discussing our findings, problems, challenges, potential obstacles and suggestions for future research topics, while Section 5 concludes the article. Moreover, Section 6 (Appendix A) addresses theoretical aspects of fuzzy logic and quality measures, whereas Section 7 (Appendix B) provides a list of symbols used.

2. Linguistic Summaries, Formalization and Quality

This section studies relevant theoretical aspects of flexible linguistic data summarization, which are used throughout the article.

2.1. Basic Types of Linguistic Summaries

Linguistic summaries summarize information from data into a concise and easily understandable interpretation. Lesot et al. (2016) divided prototype forms (protoforms) of linguistic summaries into the following three main groups:

1. classic protoforms,
2. protoforms of time series, and
3. temporal protoforms.

The classic protoforms summarize attribute(s) on the whole data set, or relations among attributes (Kacprzyk and Zadrożny 2005; Rasmussen and Yager 1997; Yager 1982). These summaries are of the structure *Q entities are S* and *Q R entities are S*, respectively, where Q is a flexible linguistic quantifier, S is a summarizer and R is a restriction. The former structure is illustrated by the sentence "*most houses have high gas consumption*". An illustrative example of the latter structure is: "*most old houses have high gas consumption*".

The protoforms of time series linguistically express behavior of attributes over time (Almeida et al. 2013; Kacprzyk et al. 2006). These summaries are divided into summaries describing a time series of the structure *Q Bs are A*, and summaries considering several time series together of the structure *Q Bs are A* Q_T *time*, where Q_T is a quantifier applied to the time attribute, Q is a relative fuzzy quantifier, A and B are the examined concepts. Illustrative sentences are: "*most trends of topic B are of low variability*" and "*about half small businesses have small response rate most of the time*", respectively.

However, the temporal protoforms do *not* use linguistic quantifiers, but a mode of behavior for creating periodic summaries. This kind of summaries is of the structure *P, the data are A*, where P is a temporal adjustment and A is a fuzzy modality. An illustrative example would be: "*regularly in autumn, the participation is high*". Here, the term "*regularly*" describes the extent to which a summary holds in considering a particular temporal adjustment (Moyse et al. 2013).

While the other protoforms are also promising for data dissemination, and could be examined and applied in a similar manner, this work is focused on the classic ones in order to examine their applicability.

2.2. *Linguistic Variables and Quantifiers*

Linguistic summaries rely on the theories of fuzzy sets and fuzzy logic, where belonging to a set is a matter of degree. A fuzzy set F over the universe of discourse X is defined by the membership function μ_F that matches each element of X with its degree of membership to the set F (Zadeh 1965)

$$\mu_F(x)\colon X \rightarrow [0,\ 1] \tag{1}$$

where $\mu_F(x) = 0$ means that an element x does not belong at all to F, while $\mu_F(x) = 1$ means that x is a full member of F. A value of $\mu_F(x)$ between 0 and 1 indicates the intensity by which the element x belongs to F. The concept of fuzzy sets is further discussed in Section 6, Appendix A.

The first major concept required for our work is Linguistic Variable (LV). An LV is a variable, whose values (often called labels) are words of natural language determined by a quintuple $(L, T(L), X, G, H)$ (Zadeh 1975), where

- L is the name of the variable,
- $T(L)$ is a set of all linguistic labels related to variable L,
- X is the universe of discourse,
- G is the syntactic rule for generating $T(L)$ values, and
- H is the semantic rule that relates each linguistic label of $T(L)$ to its meaning $H(L)$.

An example of LV is any attribute whose domain can be divided into overlapping granules, for example *pollution* and *number of visits*. The LV "*pollution*" consisting of labels *low*, *medium* and *high* is plotted in Figure 1. For a finer granulation we can construct more labels, for example *very low*, *low*, *medium*, *high* and *very high*. The syntactic rule explains the required number of linguistic labels and their names, whereas the semantic rule assigns the context dependent meaning to each label by fuzzy sets. For instance, the

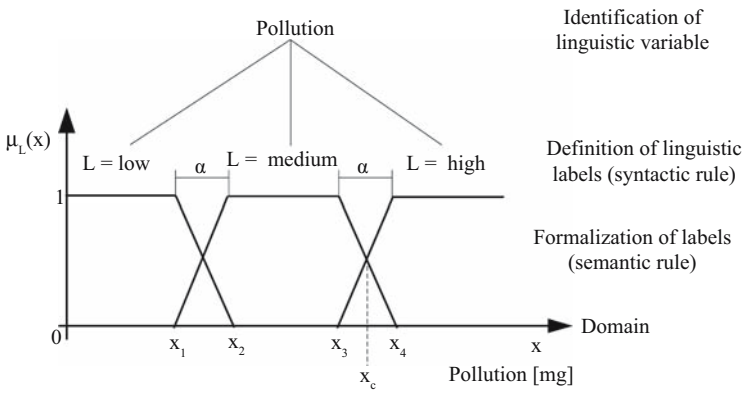

Fig. 1. Linguistic variable "pollution" and its labels.

fuzzy set *high* (using the x-values defined in Figure 1) is expressed as

$$\mu_{high}(x) = \begin{cases} 1 & x \geq x_4 \\ (x - x_3)/(x_4 - x_3) & x_3 < x < x_4 \\ 0 & x \leq x_3 \end{cases} \tag{2}$$

Value x_c is the maximal uncertainty point. In a smooth transition from sets *medium* to *high*, x_c belongs to both with 0.5 degree, that is, we cannot be sure whether this value is more *medium* than *high*. The intervals having the width α are uncertain areas. When $\alpha = 0$, these sets are *crisp* (an element is either a member of the set or not).

Generally, fuzzy sets can be formalized by non-linear functions. In this article, we adopted the linear ones due to their simplicity for the end users. In the case of non-linear functions, the users have to specify the shapes of fuzzy sets, which is not a simple task for the less mathematically literate users, as is, for example, the case in the medical domain (Holzinger et al. 2017).

The next element in LSs is the fuzzy quantifier. Fuzzy quantifiers are discussed in detail by, for example, Glöckner (2006). The formalization of fuzzy relative quantifiers can be realized by three approaches: sigma counts (Zadeh 1983), Ordered Weighted Averaging (OWA) operator (Yager 1988) and Competitive Type Aggregation (Yager 1984). For reasons of simplicity, the sigma count approach is chosen for this article. In this way, summarizer and restriction (explained later), as well as quantifier are modelled by the same approach, which is, in addition, more intuitive for diverse users. Within that approach, the quantifier *most of* is formalized by an increasing (usually linear) function where $\mu_Q(0) = 0$ and $\mu_Q(1) = 1$ as (Kacprzyk and Yager 2001; Kacprzyk and Zadrożny 2005)

$$\mu_Q(y) = \begin{cases} 1 & y \geq 0.8 \\ 2y - 0.6 & 0.3 < y < 0.8 \\ 0 & y \leq 0.3 \end{cases} \tag{3}$$

where y is the proportion of units fully or partially satisfying a predicate in a summary expressed by fuzzy sets. In our application, we modified the parameters in (3) in such a way that the membership degree becomes higher than zero only for the proportions higher than 0.5 to meet the usual meaning of *most of* and *majority*, that is

$$\mu_Q(y) = \begin{cases} 1 & y \geq 0.8 \\ (y - 0.5)/0.3 & 0.5 < y < 0.8 \\ 0 & y \leq 0.5 \end{cases} \tag{4}$$

Analogously, the quantifier *about half* is a symmetric triangular or trapezoidal fuzzy set centered around the value of 0.5 ($\mu_Q(0.5) = 1$). The quantifier *few* is expressed by a decreasing function ($\mu_Q(0) = 1$, $\mu_Q(1) = 0$). Thus, a possible family of relative quantifiers plotted in Figure 2 is also a LV.

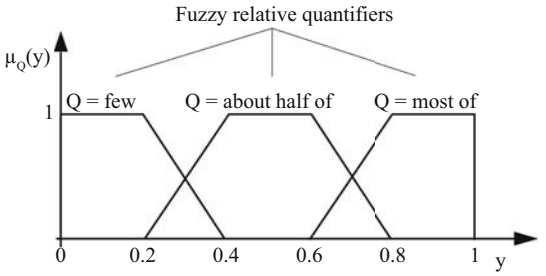

Fig. 2. A possible family of relative quantifiers for a proportion y.

2.3. *Formalization of Classic Protoforms and Their Quality Aspects*

A basic structure of LS for summarizing attributes is *Q entities in database are (have) S* (Yager 1982). Quantifier Q and summarizer S are usually both formalized by linguistic terms (fuzzy sets), for example "*most agricultural companies have a high turnover*". The proportion of records in a data set $\mathbf{X}$ that fully and partially satisfies the predicate S are defined as

$$y_{LSb}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \mu_S(x_i) \tag{5}$$

where n is the number of units in a data set and the membership function μ formalizes summarizer S for the units. The validity (truth value) of the summary is calculated as

$$v_{LSb}(\mathbf{X}) = \mu_Q(y_{LSb}(\mathbf{X})) \tag{6}$$

where the function μ formalizes quantifier Q for the summary. Both $y_{LSb}(\mathbf{X})$ and $v_{LSb}(\mathbf{X})$ assume values in the interval [0, 1].

Linguistic Summaries with restrictions take the form *Q R entities in database are (have) S*, where restriction R, also expressed in linguistic terms, focuses on a part of data set relevant for the summarization task (Rasmussen and Yager 1997), for example, "*most highly polluted municipalities have a high number of respiratory diseases*". The proportion of records in a data set $\mathbf{X}$ that fully or partially satisfies the restriction R and also fully or partially satisfies the summarizer S, is defined as

$$y_{LSr}(\mathbf{X}) = \frac{\sum_{i=1}^{n}(\mu_S(x_i) \wedge \mu_R(x_i))}{\sum_{i=1}^{n}\mu_R(x_i)} \tag{7}$$

where n is the number of units in $\mathbf{X}$ and the membership function μ formalizes, in term, both S and R. The "and operator" in the numerator is expressed by a triangular norm (Section 6, Appendix A). The convention $0/0 = 0$ is used in order to avoid undefined proportions; this situation occurs when not a single record meets R (and as a logical consequence, not a single record simultaneously meets R and S). Analogously to (6), the validity of the summary is calculated as

$$v_{LSr}(\mathbf{X}) = \mu_Q(y_{LSr}(\mathbf{X})) \tag{8}$$

The concept of LSs was introduced by Yager (1982). Since then, the theory of LSs has been improved and applied in a variety of fields. Boran et al. (2016) provide an overview of recent developments. The linguistic terms used in S and R can be formalized by fuzzy sets having functions of different shape (as illustrated in Figure A.1 of Section 6, Appendix A), ensuring the smooth transition between belonging and nonbelonging to the set.

The basic quality criterion (validity or truth value as defined in (6) and (8)) does not cover all aspects of quality (Kacprzyk and Yager 2001). Due to the complexity of quality measures, problems with their aggregation and particularities of the considered data dissemination (see Section 6, Appendix A), we adopted a simplified quality measure that integrates two of the most important measures: validity and coverage introduced by Hudec (2017) for LSs with restriction

$$Q_c = \begin{cases} t(v, C) & C \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where C is data coverage (defined as a function (A.4) of the proportion of the whole data set affected by the summarizer and restriction) and t is a nonidempotent t-norm, for example, a product t-norm (A.2). A discussion related to quality measures and the rationale for choosing the measure (9) is held in Section 6, Appendix A. This simplified measure, which is calculated from the data, contributes to the decreased complexity of interfaces, because users do not need to intervene. In (5) and (6), the whole data set is covered due to n (the cardinality of the data set) being in the denominator. It means that the data coverage is implicitly calculated in $y_{LSb}(\mathbf{X})$.

2.4. A Case Study for Interpreting Data by Crisp and Fuzzy Logic

Hypothetical values of pollution measured over all 30 days of a month in two districts are shown in Table 1. The authorities wish to disseminate information regarding the pollution dispersion. Let us have crisp set "*high pollution*" (*HP*) defined as $HP = \{x : x > 20\}$. When we apply this set in a query: *select districts where high pollution was recorded*, then district *D1* is selected, whereas *D2* is not. However, a quick glance at Table 1, applying common sense reasoning, leads to the conclusion that *D2* is more polluted than *D1*. Furthermore, it might happen that the recorded values for *D1* in days 10 and 14 are incorrect due to measurement errors. In that case, the disseminated information does not correspond with reality. Dissemination by proportion says that for *D1* pollution was high in 7% of the days, whereas high pollution was not recorded for *D2*.

Let us examine this problem from the fuzzy logic perspective. The concept "*high pollution*" can be expressed by a fuzzy set (2) as follows

$$\mu_{FHP}(x) = \begin{cases} 1 & x \geq 20 \\ (x - 15)/5 & 15 < y < 20 \\ 0 & x \leq 15 \end{cases} \tag{10}$$

where pollution above 20 units is still considered high without any doubt, but slightly lower values belong to the concept "*high pollution*" with membership degrees smaller than 1 (Table 1).

Table 1. Measured pollution for two districts (illustrative data).

District D1						District D2					
Day	Measured pollution [mg]	Matching degree to high pollution $\mu_{FHP}(x)$	Day	Measured pollution [mg]	Matching degree to high pollution $\mu_{FHP}(x)$	Day	Measured pollution [mg]	Matching degree to high pollution $\mu_{FHP}(x)$	Day	Measured pollution [mg]	Matching degree to high pollution $\mu_{FHP}(x)$
1	2.950	0	16	8.375	0	1	16.577	0.315	16	18.925	0.785
2	6.740	0	17	8.079	0	2	16.923	0.385	17	17.223	0.445
3	1.669	0	18	9.183	0	3	15.102	0.020	18	14.465	0
4	5.887	0	19	7.104	0	4	19.383	0.877	19	18.465	0.693
5	2.621	0	20	16.005	0.2010	5	18.606	0.721	20	11.530	0
6	9.106	0	21	5.630	0	6	12.981	0	21	17.281	0.456
7	8.239	0	22	10.286	0	7	16.589	0.318	22	16.084	0.217
8	7.036	0	23	4.569	0	8	19.038	0.808	23	19.969	0.994
9	5.438	0	24	8.877	0	9	14.043	0	24	15.023	0.005
10	21.232	1	25	8.150	0	10	19.346	0.869	25	16.003	0.201
11	4.285	0	26	16.256	0.2512	11	18.443	0.689	26	17.226	0.445
12	7.494	0	27	4.456	0	12	19.889	0.978	27	18.099	0.620
13	2.831	0	28	3.187	0	13	19.886	0.977	28	18.402	0.680
14	20.006	1	29	2.041	0	14	18.359	0.672	29	12.049	0
15	1.810	0	30	2.950	0	15	19.039	0.808	30	18.077	0.615

Next, we calculate y_{LSb} and adopt suitable quantifiers. The proportion for *D1* (i.e., the sum of matching degrees divided by 30 – i.e., the number of days) is obtained as 0.07 (as for the *crisp logic* example above), whereas the proportion is 0.51 for *D2*. Now, we are able to disseminate by proportions: "*for D1, in 7% of the days, pollution was high*, *for D2, in 51% of the days, pollution was high*". We may continue to elaborate a more sophisticated linguistic interpretation by the following two sentences: "*in D1, for a few days, pollution is high*; *in D2, for about half of the days, pollution is high*" (using the parameters shown in Figure 2, the validities of both sentences are obtained as $\mu_{few}(y_{D1}) = 1$ and $\mu_{about\ half}(y_{D2}) = 1$, respectively). The second sentence can be further summarized into "*in D2, for slightly above half of the days, pollution is high*", when we formalize the quantifier *slightly above half*.

3. Linguistically Summarizing Statistical Data

In this section the innovative potential of LSs for the official statistics data dissemination is demonstrated on the illustrative data, as well as on the real data from the Municipal Statistics Database managed by the Statistical Office of the Slovak Republic. This database consists of more than 800 attributes for 2,927 municipalities. The test interfaces were developed for the sole purpose of illustrating applicability and procedures for calculating linguistic summaries and have therefore not yet been tested on users. The interfaces were developed in Visual Studio 2013 and MS Access 2013, while the data was stored in an MS Access relational database.

3.1. An Option of Representing Data by a Set of High-Validity Sentences

The extent to which observations are spread around their mean value is expressed by dispersion functions. However, these functions can be overlooked, especially by people with a lower level of statistical literacy, who may conclude that all essential information about a variable is encapsulated in its mean value. In the following example, we will illustrate how linguistic summaries could help remedy this.

Example 1

A fictive data set contains seven respondents with their respective ages {26, 28, 32, 40, 54, 56, 57} (Hudec 2016). Summarization by statistical methods reveals that the average age (arithmetic mean) is 41.9, the median age is 40, and the standard deviation is 13.7. The arithmetic mean and median lead us to the conclusion that the typical age of a respondent is around 40, but standard deviation shows that this is not the case.

The interpretation by linguistic summaries says the same, but differently. Three labels: *young*, *middle-aged* and *old* of the LV "*age*" required for summarizer *S* are formalized as follows

$$\mu_{young}(x) = \begin{cases} 1 & x \leq 30 \\ (35 - x)/5 & 30 < x < 35 \\ 0 & x \geq 35 \end{cases}$$

$$\mu_{middle\_aged}(x) = \begin{cases} (x-30)/5 & 30 < x < 35 \\ 1 & 35 \leq x \leq 50 \\ (55-x)/5 & 50 < x < 55 \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_{old}(x) = \begin{cases} 0 & x \leq 50 \\ (x-50)/5 & 50 < x < 55 \\ 1 & x \geq 55 \end{cases}$$

The LV expressing the family of quantifiers: *few*, *about half of* and *most of* is depicted in Figure 2. With three labels and three quantifiers, $3{\cdot}3 = 9$ possible LSs exist. The high validity sentences and their respective validities are shown in Table 2.

From Table 2 it is clear that about half of respondents are young, and about half are old, whereas few are middle-aged (although the mean value is around 40). It is worth noting that a histogram would provide the same message visually; this is the corresponding verbal summary.

Linguistic summaries are able to generate all relevant sentences regarding the attributes under consideration and merge them to create a simple story. In our case, the story is*:* "*half of respondents are old, about half are young and few are middle-aged*". Moreover, such summaries might be supportive for automated or computational journalism, that is, technologically oriented journalism focused on the application of computational intelligence to the practices of information gathering and information presentation (Coddington 2015). Graefe (2016, 15) states that "Current solutions range from simple code that extracts numbers from a database, which are then used to fill in the blanks in prewritten template stories, to more sophisticated approaches that analyse data to gain additional insight and create more compelling narratives." The LSs concept presented in this article is situated between these two extremes.

This discussion naturally leads to the question of automated creation of relevant LSs, which is an important future topic in machine learning. This task is formalized by Liu (2011) as

$$\begin{aligned} &\text{find } Q, S, R \\ &\text{subject to} \\ &Q \in \bar{Q}, S \in \bar{S}, R \in \bar{R}, v(Q,S,R) \geq \theta \end{aligned} \tag{11}$$

Table 2. Summaries of high validity, which express age of respondents.

Linguistic summary	Validity as defined in (6)
About half respondents are old	1.0000
Few respondents are middle-aged	0.8575
About half respondents are young	0.8570

where $\bar{Q}$ is a set of quantifiers of interest, $\bar{R}$ and $\bar{S}$ are sets of relevant linguistic expressions for restriction and summarizer, respectively, and θ is a threshold value from the interval]0, 1]. In this case, all feasible solutions ($Q*$ $R*$ *are* $S*$) create a story.

In Example 1, we have the following task:

find Q, S

subject to

$Q \in \{few,\ about\ half,\ most\ of\}, S \in \{young,\ middle-aged,\ old\}, v(Q,S,R) \geq 0.75$

To also take the quality aspect into account, the constraint related to the validity threshold in (11) could be replaced by

$$Q_c(Q,S,R) \geq \theta_k \tag{12}$$

where θ_k is a threshold value from the interval]0, 1] related to quality expectations.

The following question naturally arises: how can we efficiently obtain LSs from large data sets? When the number of records and their attributes is relatively large, the computation might take much more time, hence might be costly. For instance, when having 2,927 records described by 800 attributes, it is necessary to compute 2927·800 membership degrees (Niewiadomski et al. 2006). We can avoid such an amount of computation by optimization procedures based on the calculated proportions using matching degrees and involving users to select sets of relevant attributes for $\bar{S}$, $\bar{R}$ and $\bar{Q}$ (11). Moreover, the processor power and memory size of modern computers ensure that the response time is not too high (the examples on municipal statistics were executed within a few seconds on an ordinary desktop computer).

3.2. Linguistically Expressing Data Distributions Around the Mean Value

The well-known and often used SQL query language contains a function for computing arithmetic mean, abbreviated as AVG, as well as a function for calculating standard deviation, abbreviated as STDEV from databases or data warehouses.

We have extended this functionality for LSs. The procedure is as follows: In the first step, the SQL query retrieves the mean value M, standard deviation and number of records of a chosen attribute with the following SELECT statement

```
SELECT AVG(chosen_att) as M, STDEV(chosen_att) as st_dev,
COUNT(id_record) as n
```

where *chosen_att* stands for the attribute selected by the user. The retrieved mean value M is a modal element of a triangular fuzzy set plotted in Figure 3. This fuzzy set is created by a widening factor *wf* of a membership function to get symmetric and convex fuzzy number "*around the mean value M*". The lowest and the highest values of support are calculated in the following way

$$a = M - wf \cdot M; \quad b = M + wf \cdot M \tag{13}$$

In the next step, all values of *chosen_att* that belong to the support of the fuzzy set "*around the mean value M*" (]*a*, *b*[– Figure 3) are selected from the database by the following SELECT statement

SELECT chosen_att FROM municipalities WHERE chosen_att BETWEEN (a, b)

Consequently, matching degrees for these values to the fuzzy set "*around mean value M*" are calculated, summed and divided by *n* for y_{LSb}. In the final step, matching degrees to the respective quantifiers are calculated; the relevant linguistic interpretation is constructed and shown in the interface.

The benefit of triangular fuzzy numbers against the interval *around the mean value M* is in the intensity of belonging. The closer an element is to the boundary, the lower matching degree the element has, and accordingly, its influence on the proportions y_{LSb} and y_{LSr} is lowered. The next example, illustrating our procedure, is based on the Municipal Statistics Database.

Example 2

A historian wishes to examine the mean value of *the year of the first written notice* (an attribute in the aforementioned municipal database). In addition, the historian has divided municipalities into two sets: "*population less than 12,000*" and "*population greater than or equal to 12,000*". The interface for interpreting solutions is shown in Figure 4. In this interface, the user can choose the relevant attribute and relative dispersion *wf* around the mean value; *wf* is set to 10% by default. The user can add further conditions merged by the logical "and operator" for focusing on the more restrictive subset of municipalities, or merged by the logical "or operator" for the less restrictive subset.

Via this interface, the historian can discover that the mean value of *the year of first written notice* for the municipalities with low population is the year 1363, and also that about half of them have their year of first written notice in the vicinity of the mean value (Figure 4 – upper interface). Hence, the mean value is a suitable generalization. For the municipalities with high population, the situation is the opposite. The mean value is the year 1147, but few municipalities fully or partially belong to the neighbourhood of this mean value (Figure 4 – lower interface).

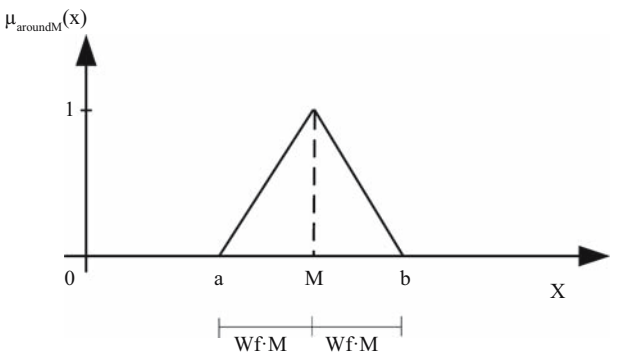

Fig. 3. Triangular fuzzy set "around the mean value M".

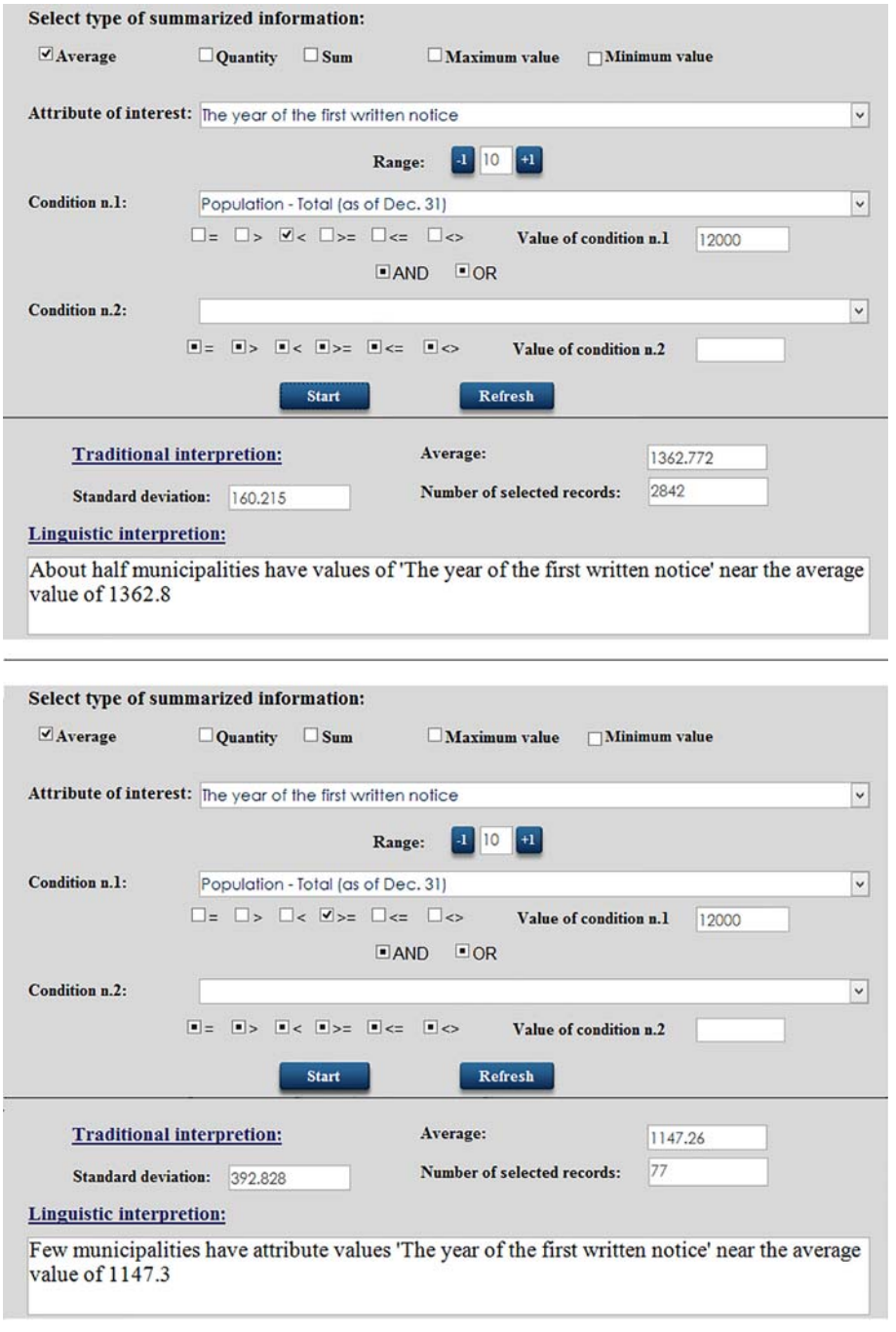

Fig. 4. The interface for linguistically interpreting data distribution around the mean value (upper interface for population $<$ 12,000, lower interface for population $\geq$ 12,000).

These interpretations are suitable for advanced, as well as for less advanced (in terms of statistical literacy and IT skills) users of official statistics, because the well-known statistical measures are disseminated together with their verbal interpretations. Additional functionalities can be added when required. For systems based on fuzzy logic, the following observation holds: "With any given system, it is easy to layer on more functionality without starting again from scratch" (Meyer and Zimmermann 2011, 432).

3.3. *Quantified Sentences as Nested Query Conditions*

This class of queries is suitable for the 1:N relationships in relational databases, or dimensions and facts in data warehouses, such as DISTRICT-RESPONDENT (one district contains multiple respondents, but each respondent is settled in one district).

An example of a quantified query condition is: "*find regions where most of the municipalities have a high amount of waste produced per inhabitant*". The algorithm is not complicated, but it might take more time depending on the number of entities on the "1" side of the considered relationship. The formula for calculating validities for each class j on the "1" side is created as the extension of (5) and (6) in the following way (Hudec 2016)

$$v_{LSbj}(x) = \mu_Q\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\mu_S(x_{ij})\right), \quad j = 1 \ldots K \tag{14}$$

where n_j is the number of entities in class j (e.g., municipalities belonging to the region j), K is the number of classes in a database (e.g., regions) and v_{LSbj} is the validity of LS for jth class. Similarly, the nested query condition expressed by LS with restriction can be constructed by extending (7) and (8).

Example 3

A small enterprise is interested in extending its business activities related to agricultural equipment into the areas of low altitude and high ratio of arable land, but it is not sure which regions to favor. Hence, the SQL-like flexible query is: *SELECT regions WHERE most of the municipalities have low altitude and high ratio of arable land*. The decision maker considers an altitude of less than 200 m to perfectly match, between 200 m and 270 m to partially match and above 270 m to be out of the question. Thus, we formalize this user's linguistically expressed requirement by the fuzzy set *low*, where $m = 200$ and $b = 270$ (Figure A.1 (see Subsection 6.1)). The high ratio of arable land is formalized by the fuzzy set *high* plotted in Figure A.1, where $a = 40$ and $m = 60$. The quantifier *most of* is formalized by (4). For the "and operator" in the summarizer the minimum t-norm (A.1) was used. The result is presented in Table 3, where two regions (out of eight) partially meet the condition.

Table 3. Selected regions by quantified query condition: "most of the municipalities have low altitude above sea level and high ratio of arable land".

Region	Validity as defined in (14)
Nitra	0.930
Trnava	0.603

In the case of a classical database query, none of the regions meet the query condition and therefore the result is the empty set.

A further benefit for users may be to disseminate these results on thematic maps, for example, highlighting territorial units by hues, the intensity of which would be determined by the validity calculated by (14).

This type of summaries displays records on the higher hierarchy level, not data on lower levels. This is convenient when data on lower levels are sensitive. Hence, the risk of data disclosure is reduced, but care should be taken when summarizing from a small data set.

3.4. Summaries about Attributes

The basic structure of LSs (6) provides a summary across the database for a particular subset of attributes. In order to practically illustrate this, we have developed a procedure and an illustrative interface (see Figure 5) for the aforementioned municipal database. The user selects the desired quantifier, chooses a relevant attribute from the database and selects the desired LV label (Subsection 2.2 and Figure 1). Consequently, the suggested parameters of a chosen label (fuzzy set) are shown under the picture of LV.

Example 4

A journalist examines distances to the nearest train stop for the purpose of writing an article regarding the train network coverage in municipalities. As shown in Figure 5, the interface requires the user to select a relevant quantifier, in our case *about half*, an attribute *Distance in km to the nearest passenger train station* and a label *low*. The value of 0 km means that the train station is situated within the municipality in question, whereas a

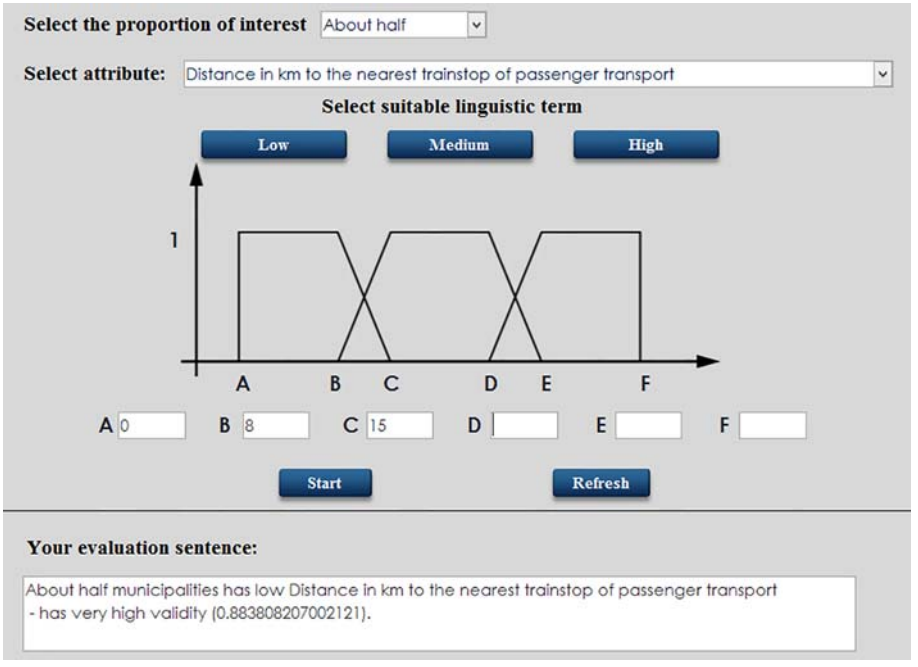

Fig. 5. The illustrative interface for creating a LS and interpreting its validity.

Table 4. A possible mapping from validities (6) and (8) into a linguistic interpretation.

Validity	Linguistic explanation
0	Sentence is irrelevant
]0, 0.15]	Sentence has very low validity
]0.15, 0.4]	Sentence has low validity
]0.4, 0.6]	Sentence has medium validity
]0.6, 0.85]	Sentence has high validity
]0.85, 1[	Sentence has very high validity
1	Sentence excellently explains data

distance greater than 0 indicates how far from the municipality the nearest train stop is. When the journalist chooses the label *low*, the lowest value from the database is shown (parameter A) and initial parameters for the fuzzy set *low* (parameters B and C) are suggested. In the next step, the user can modify the parameters to values more suitable for a particular task. Finally, the linguistic interpretation is shown in the explanation box. In our case, the sentence "*about half of the municipalities have a low distance to the nearest passenger train station*" has a very high validity. The validity value in brackets is just shown for the purpose of illustration, and would be hidden by default. The rating of a linguistic explanation depends on the validity of quantified sentences. A possible mapping from v_{LSb} or v_{LSr} into a linguistic interpretation is shown in Table 4.

3.5. *Summaries for Subsets Expressed by Linguistic Summaries with Restrictions*

To demonstrate summaries based on (8), we have developed a procedure and an interface for summarizing from the aforementioned municipal database.

Example 5

An environmental agency is interested in learning whether "*the majority of municipalities with a high ratio of arable land have a low population density*". The interface is shown in Figure 6. On the upper left-hand side, the user chooses the relevant quantifier from a drop-down list (and modifies its parameters if needed). In the main part, the user selects

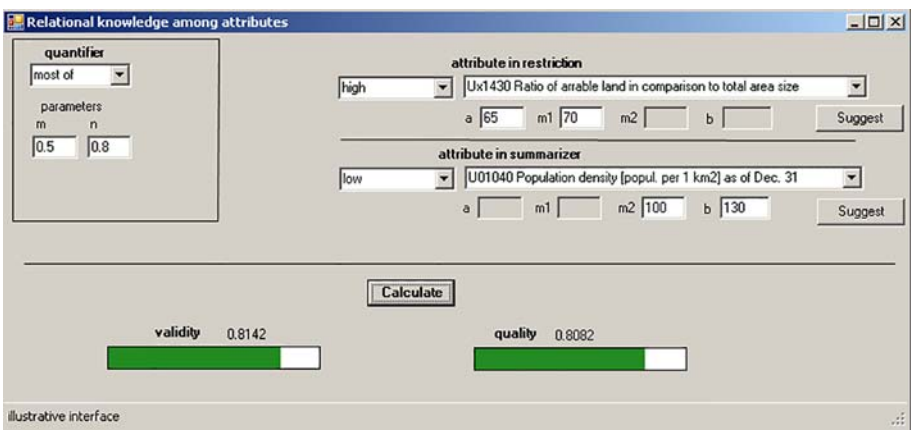

Fig. 6. The illustrative interface for analysing LS with restriction.

attributes and desired linguistic labels. The user can directly assign values to the respective parameters of labels, or ask for suggestions. In the latter case, parameters are mined from the database and presented to the user, who has the choice to either accept or modify them. Parameters a, m_1, m_2, b correspond to the fuzzy sets parameters shown in Figure A.1. For the fuzzy set expressing the term *high* we have that $m_1 = m$, and for the fuzzy set formalizing the term *low* we have that $m_2 = m$. The chosen parameters are shown in Figure 6. The validity of this summary is 0.814 (defined by (8)) and its quality is 0.8082 (defined by (9)). These high values lead us to the conclusion that the summarized sentence is of a high quality.

If an agency is interested in investigating whether "*most of the municipalities with a high number of warm days (temperature above 25°C) have a small amount of waste produced per inhabitant*", then the validity (8) of this summary is equal to 1. However, the coverage (A.4) is equal to 0 and therefore the quality (9) is 0, for example, this summary is not representative. Focusing only on validity might lead us to draw inappropriate conclusions.

4. Discussion

This section provides the main features of the suggested approach and a reflection on its advantages and drawbacks, as well as some further research opportunities.

4.1. The Main Features of the Suggested Approach

In this article, we adjusted well-known approaches for formalizing flexible predicates, quantifiers and quality measures, and provided the rationale for our choices. The suggested approach may be beneficial for NSIs due to the following features:

- It is less sensitive to the imprecise nature of some data and to inliers (i.e., erroneous values that lie in the normal range of a variable). When the measured value is not far from the real one, then this approach eliminates sharp jumps between belonging and not belonging to a set (Figures 1 and A.2 (see Subsection 6.1)).
- The suggested approach reveals summaries from the data, not the data itself. Generally speaking, the data disclosure would not be a problem; however, care should be taken when summarizing from small data sets. The decision regarding which data sources might be available for users to realize summaries should meet regulations and other relevant rules.
- The less complex interpretation of the data is especially welcome for less statistically literate users and disabled people, for whom the summarized sentences may be interpreted by voice.
- The *computing with words* concept can easily be applied to any human language. Adjectives such as *high* and quantifiers such as *most of* are always expressed by increasing functions, regardless of their translation to the other languages and examined concepts.
- LSs are able to offer an alternative answer when the initial sentence (summary) is of insufficient validity. For instance, if the proportion for the sentence "*most short visits are from countries with high GDP*" is 0.06, the answer is not only that the validity is

zero, but we can provide an alternative summary: "*few short visits are from countries with high GDP*".

- Statistical offices typically refrain from disseminating dispersion measures, although this information is valuable. Our suggested approach includes the linguistically interpreted deviation, which is suitable for all users, especially for the less statistically literate ones.

4.2. Further Research and Development Opportunities

The following subsections identify opportunities for future research topics.

4.2.1. Reflections on User Input

The quality measures for LSs are usually calculated from the data, excluding human intervention. While this might sometimes be convenient, it might sometimes be useful to develop an additional measure, for which the user would be able to assign relevance to each summary of interest.

The interfaces introduced in Section 3 were created for illustrative purposes. The interfaces in Figures 4 and 5 might be suitable for both types of users, since well-known statistical measures and linguistic interpretations are provided. The interface for summaries among attributes (Figure 6) may be difficult to use for less skilled users. On the other hand, experienced ones might welcome the possibilities of adjusting all relevant parameters of linguistic labels. The option provided to the less skilled users by the test interface is the automated support. Further options might be inspired by ReqFlex – a "fuzzy query engine for everyone", developed by Smits et al. (2013), where the users assign parameters by moving sliders rather than filling input boxes. Future research should include sophisticated usability testing and adjusting various designs according to the user feedback in order to meet the expectations of both advanced and less advanced users. While this approach is applicable for summarizing, for example at the European Union Member State level, the benefit of LSs is in general higher for larger data sets, such as on levels of Nomenclature of Territorial Units for Statistics (NUTS).

4.2.2. Linguistic Quality

A possible obstacle might be the structure of short quantified sentences (indicated in italics throughout this article). Although such structures are widely used, the order of terms and the structure itself might not fully meet the usual terminology in official statistics, general public expectations and grammar rules. A mechanical construction of sentences may lead to grammatically incorrect expressions. Thus, there is room for experts from different fields, including linguists, to identify sound and practical solutions, but interactive machine learning could also be of help here. Moreover, verbal explanations are extremely important for the emerging field of "explainable artificial intelligence" (Goebel et al. 2018), which opens additional application fields.

4.2.3. Applying SDMX to Summaries

The Statistical Data and Metadata eXchange (SDMX) standard was initially developed for the dissemination and exchange of data (SDMX 2012). The dimensional data structure is

solid, because it is based on a clear methodology and is therefore suitable for inclusion into business intelligence questions. This structure can be helpful for creating linguistic variables over a set of dimensions and measures. The possibility of managing fuzzy data by the SDMX standard is touched upon by Hudec and Praženka (2016).

4.2.4. Applying Linguistic Summaries to New Data Sources

National Statistical Institutes (NSI) are also focusing their activities on alternative sources, including social networks (e.g., Torres van Grinsven and Snijkers 2015), web scraping (e.g., Barcaroli et al. 2015), mobile positioning data (e.g., Altin et al. 2015) and the like. Because the validities (6) and (8), as well as the quality measure (9) depend only on the intensities of belonging to fuzzy sets, it means that we can straightforwardly summarize from other data types. The only difference is in computing matching degrees of imprecise numbers (known as fuzzy numbers), (weighted) categorical data and sentence fragments to fuzzy sets in summarizer and restriction. For weighted categorical data (e.g., *negative (0.7) and neutral (0.3) opinion*) and fuzzy data (e.g,. *value is most likely 120 but for sure not lower than 100 and not higher than 150*) instead of calculating matching degrees of crisp numbers to the fuzzy sets, the possibility and necessity measures are applied (Galindo et al. 2006). For data expressed by short sentences or sentence fragments (e.g., *productivity is remarkably low*) matching degrees to the fuzzy concepts can be calculated by application of methods suggested by Duraj et al. (2015) and Niewiadomski (2002).

4.2.5. Enhanced Dissemination as an Incentive for Data Providers

Although data collection and dissemination are at two opposite ends of the statistical data production process, they influence each other. Adolfsson et al. (2010) estimated that 30% of total data collection costs is allocated to data editing (imputation). Ross (2009) observed the paradox that users of official statistics are becoming more demanding with regard to data, but are less willing to provide their own data to NSIs. This problem results from the fact that respondents cooperate in many official surveys, but on the other hand, they often are not able to easily find and interpret relevant information on NSI data portals (Bavdaž 2011). One possible solution is in flexible and tailored data dissemination (Hudec and Torres van Grinsven 2013). As further motivation, we could offer sophisticated methods for linguistically interpreted summaries (means, deviations, time series, etc.) to businesses that cooperate timely in surveys. The practical feasibility of achieving this (while maintaining the principle of impartiality) is a topic for future research.

5. Conclusions

One of the missions of NSIs is the dissemination of statistical data to a large variety of users, ranging from experts to the general public (including disabled people). Statistical agencies should offer flexibility in dissemination to avoid jeopardizing their mission (Bavdaž 2011). This may require rules for using natural human languages to describe key measures (Schield 2011) and to make statistics easily understandable and usable by the general public (Bier and Nymand-Andersen 2011). Thus, NSIs should apply different strategies in order to meet the expectations of diverse user categories. This article tackles innovative dissemination by short quantified sentences of natural language, which is

definitely a promising method to reach these goals. In particular, for some categories of users with disabilities, textual interpretation or interpretation by voice (rendered possible by LSs) would be more suitable than what is offered by current dissemination methods.

The potential of LSs is demonstrated on test interfaces on the real-world data. In order to reduce both complexity and interaction requirements on users, we have suggested approaches for constructing fuzzy sets and for measuring quality that minimally burden users. Further, our research has documented perspectives, obstacles and problems leading to future research directions. The important activity is in real-world testing with users to develop broadly accepted designs for full-featured and easy-to-use interfaces. These tasks should be solved in cooperation between NSI data dissemination units and scientists working in the aforementioned fields.

Finally, we emphasize that our approach based on LSs should not be considered as a rival to existing ones, but rather as a complementary dissemination practice to well-established ones.

6. Appendix A. Theoretical Concepts Related to Fuzzy Set Theory and Linguistic Summaries

This appendix provides an insight into fuzzy set theory, fuzzy "and operator" and quality measures of summaries.

6.1. Fuzzy Sets

The linguistic terms *low*, *medium* and *high* (Figure 1) can be formalized by an L fuzzy set, a trapezoidal fuzzy set and a linear gamma fuzzy set, respectively, as illustrated in Figure A.1.

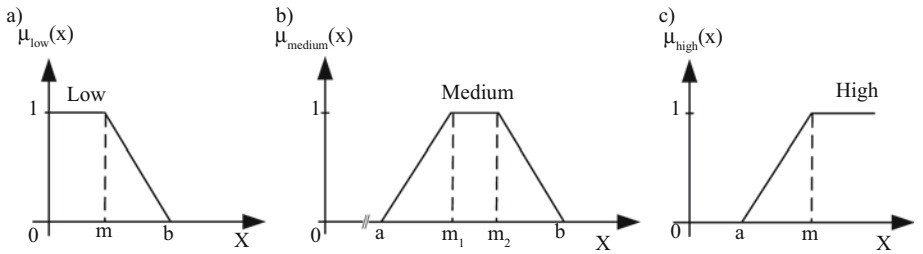

Fig. A.1. Fuzzy sets: a) L fuzzy set, b) trapezoidal, c) linear gamma.

Fuzzy sets are context dependent, for example, they may have different parameters for various given concepts. For instance, the set "*short distance*" has a different meaning – expressed by parameters m and b in Figure A.1 – for a small and densely populated country, and for a large but sparsely inhabited country. Two important concepts are the core and the support of fuzzy sets. The core of a fuzzy set contains all elements that fully belong to the set. The core of the fuzzy set *medium* contains all elements in interval $[m_1, m_2]$. The support of the fuzzy set contains all elements that belong to the set with degree greater than 0, that is, the support of fuzzy set *medium* is interval [a, b].

For instance, assume that someone wishes to know whether certain municipalities belong to the set "*high pollution*" (*HP*). The set *HP* is expressed as a fuzzy set shown in

Figure A.2a, and as a crisp set in Figure A.2b, where φ is a characteristic function (a bi-valued function expressing membership of a crisp set). In Figure A.2a, the values 50 mg and 55 mg delimit the area where belonging to the set is a matter of degree. If we apply classical set theory, two similar values may be treated differently. For example: a municipality, in which a value of 54.73 mg was recorded, does not belong to the crisp set *HP*, whereas a municipality having a recorded value of 55 mg does belong to it. In the case of a fuzzy set, a municipality polluted with 54.73 mg participates in the set *HP* with a slightly lower degree than 1. The possible measurement error for values around 55 mg may cause assignment to the wrong crisp set.

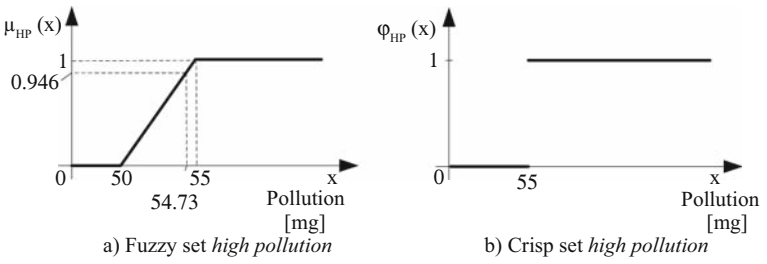

Fig. A.2. Concept "high pollution" expressed as fuzzy set (a) and crisp set (b).

On the other hand, when a categorization relies on precise or sharp rules, we should use crisp sets. For instance, the category Small and Medium-sized Enterprise (SME) is divided into three subsets (by number of employees): micro enterprises – fewer than 10 persons employed; small enterprises – from 10 to 49 persons employed; medium-sized enterprises – from 50 to 249 persons employed (e.g., EU Guide 2015). In this case, these sets have sharp boundaries (or $\alpha = 0$ in Figure 1).We can still use these sets in LSs, for example, to assess whether "*few micro enterprises in tourism have low turnover*".

6.2. *Triangular Norms*

The "and operator" is expressed by triangular norms, which were initially developed for statistical metric spaces and later modified and applied for the fuzzy "and operator" (Schweizer and Sklar 1983).

When a restriction R and/or summarizer S consisting of several atomic predicates aggregated by the "and operator", triangular norms (t-norms) should be used. Two well-known t-norms, both of which are discussed in Klement et al. (2005), are the minimum t-norm

$$\mu_P(x) = \underbrace{\min}_{i=1\ldots n} \mu_{Pi}(x) \tag{A.1}$$

and the product t-norm expressed as

$$\mu_P(x) = \prod_{i=1}^{n} \mu_{Pi}(x) \tag{A.2}$$

where P stands for the compound predicate. All t-norms meet all axiomatic properties of "and operator", but differ in satisfying algebraic properties (Klement et al. 2005) to cover a variety of tasks.

6.3. Quality Measures of Summaries

The basic quality criterion (validity or truth value as defined in (6) and (8)) is the most important one, but it does not cover all aspects of quality (Kacprzyk and Yager 2001). Let us focus on LSs with restriction (7) and (8). It is possible that the validity equal to 1 explains the summary from the outliers (Hudec 2017). In order to avoid this problem, several quality measures have been suggested.

Hirota and Pedrycz (1999) have introduced five features for measuring quality of mined and aggregated information: validity, novelty, usefulness, simplicity and generality. Wu et al. (2010) have proposed equations for calculating these measures for linguistic summaries with restriction. In that approach, validity corresponds to (8). The generality measure is expressed by sufficient coverage that indicates whether a summary is supported by a sufficient subset of the data. First, the coverage ratio is calculated as (Wu et al. 2010)

$$i_c = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{A.3}$$

where n is the number of records and $p_i = \begin{cases} 1 & \mu_S(x_i) > 0 \wedge \mu_R(x_i) > 0 \\ 0 & \text{otherwise} \end{cases}$

Because a summary of the structure (8) covers a subset of the whole database, i_c is considerably smaller than 1. Thus, the following mapping $[0, 1] \rightarrow [0, 1]$ converts this ratio into the degree of sufficient coverage (Wu et al. 2010)

$$C = f(i_c) = \begin{cases} 0 & i_c \leq r_1 \\ 2((i_c - r_1)/(r_2 - r_1))^2 & r_1 \leq i_c < (r_1 + r_2)/2 \\ 1 - 2((r_2 - i_c)/(r_2 - r_1))^2 & (r_1 + r_2)/2 \leq i_c < r_2 \\ 1 & i_c \geq r_2 \end{cases} \tag{A.4}$$

where the suggested values for parameters r_1 and r_2 are 0.02 and 0.15, respectively.

The degree of usefulness is computed as a minimum of validity and coverage (i.e., $U = \min(v_{LSr}, C)$). The degree of outlyingness O, referring to novelty (unexpected summaries are very valuable for users if they cover the regular behavior in the data, not in outliers), is an aggregation of validity and coverage as: "the validity degree v is very small or very high and the sufficient coverage C must be very small" (Wu et al. 2010, 14). To keep the best value of each measure equal to 1, instead of the outlier measure, we should use its negation $(1 - O)$. Finally, the simplicity measure expresses the length of a sentence as (Wu et al. 2010)

$$SL = 2^{2-|S\cup R|} \tag{A.5}$$

where $|S \cup R|$ is the cardinality of union between R and S. When R and S contains one attribute each, the simplicity measure gets the value 1. All aforementioned measures get values from the unit interval, which makes their aggregation easier, but some measures are functionally dependent (Hudec 2017).

Kacprzyk and Strykowski (1999) have introduced the following quality measures: truth value or validity, degree of precision, degree of coverage, degree of appropriateness, and

length of summary. These measures are mainly focused on the basic structure of LSs, (5). The truth value (T_1) basically corresponds to validity (6). The degree of fuzziness is high for summaries based on very vague attributes in S. The wider the support of fuzzy set, the higher the value of fuzziness, that is

$$d_{fz}(S_j) = (|\{x \in A_j : \mu_{Sj}(x) > 0\}|)/(|A_j|) \tag{A.6}$$

where S_j is predicate on attribute A_j in summarizer S. This quality measure, the degree of precision, is defined as

$$T_2 = 1 - \sqrt{\prod_{j=1}^{s} d_{fz}(S_j)} \tag{A.7}$$

where s is the number of atomic predicates in summariser S. Values close to 1 are associated with summaries of low fuzziness.

The degree of coverage (T_3) basically corresponds to (A.3) and (A.4). The degree of appropriateness is a measure functionally dependent on T_3

$$T_4 = \left| \prod_{j=1}^{s} k_j - T_3 \right| \tag{A.8}$$

where $k_j = \left(\sum_{i=1}^{n} h_i\right)/n$, and h_i is defined to be equal to 1 when the ith record satisfies membership function μ for S_j, and 0 otherwise. The role of this measure is to exclude trivial summaries of high validity.

The length of the summary corresponds to (A.5), but it is adjusted to the basic structure of LSs by

$$T_5 = 2 \cdot 0.5^{|S|} \tag{A.9}$$

This measure gets value 1 when the cardinality of S is equal to 1, that is, S consists of one atomic predicate.

The problem of applying (A.6) to summaries on the Municipal Statistics Database is that for many attributes, the data distribution is unbalanced and therefore a low value of T_2 does not necessarily imply low quality. In addition, users may have particular reasons to express requirements by "wide" fuzzy sets. Regarding the summary length, we should use (A.5) for LSs with restriction and (A.9) for the basic structure.

Another problem is the aggregation of quality measures. Kacprzyk and Yager (2001) suggest the weighted average

$$T = \sum_{i=1}^{5} w_i T_i \tag{A.10}$$

where $\sum_{i=1}^{5} w_i = 1$.

For example, this way is suitable for decision support (e.g., in the medical domain), where decision makers assign values to w_i either individually or by consensus. On the other hand, this way is not applicable for disseminating statistical data to the general public, because assigning weights imposes a burden on users.

George and Srikanth (1996) have developed a genetic algorithm for fitness function to compute the best summary. Having the simplicity and robust solution for statistical dissemination in mind, this way is not elaborated further.

6.4. A Brief Review of Using Fuzzy Sets in Queries

The first practical implementations of flexible queries were FQUERY introduced by Kacprzyk and Zadrożny (1995) and SQLf introduced by Bosc and Pivert (1995). These approaches faced the problems of covering complex aggregation operators. Quantified query conditions, that is, selecting entities that meet the majority of atomic conditions, were introduced by Kacprzyk and Ziółkowski (1986). An illustrative example is to find municipalities where most of the conditions "*altitude above sea level is around 700 m and population density is small and municipality size is medium and pollution is low and opinion about municipality is positive*" are satisfied. The empty answer problem is an issue when a higher number of atomic conditions is merged by the "and operator". Quantified query conditions based on LSs mitigate this problem by retrieving not only entities that meet all atomic conditions, but also entities that meet the majority of these conditions.

The first querying tool for summarizing the data was SummarySQL (Rasmussen and Yager 1997) followed by SAINTETIQ (Raschia and Mouaddib 2002) and the extension of FQUERY (Kacprzyk and Zadrożny 2005). Achievements related to the official statistics data dissemination community were mainly focused on the fuzzy queries (Hudec 2013).

7. Appendix B. Overview of Symbols Used

a	Left border of fuzzy set support
A	Attribute, topic
α	Length of the uncertain area in fuzzy set
b	Right border of fuzzy set support
C	Coverage
d_{fz}	Degree of fuzziness
f	Function
F	Fuzzy set
φ	Characteristic function of crisp set
G	Syntactic rule for *LV*
h	Parameters used to calculate T_4
H	Semantic rule for *LV*
i_c	Coverage ratio
k	Parameter used to calculate T_4
K	Number of classes
L	Name of linguistic variable
LS	Linguistic Summary
LV	Linguistic Variable
m	Modal value of fuzzy set
m_1	Left border of fuzzy set core

m_2	Right border of fuzzy set core
M	Mean value (average)
μ	Membership function of fuzzy set
$\mu_F(x)$	Membership degree of element x to fuzzy set F
n	Number of records
O	Outlier degree
p	Parameters used to calculate i_c
P	Predicate
Q	Fuzzy quantifier
Q_c	Quality measure aggregating validity and coverage
r	Parameters used to calculate coverage C from its ratio i_c
R	Restriction
s	Number of anomic predicates in summarizer S
S	Summarizer
SL	Simplicity measure
t	t-norm
T	Set of labels (linguistic terms)
T_1	Truth value
T_2	Degree of precision
T_3	Degree of covering
T_4	Degree of appropriateness
T_5	Length of summary
θ	Threshold value
Q_T	Fuzzy quantifier applied to time attribute
U	Usefulness measure
v	Validity or truth value of summary
w	Weight
wf	Widening factor
X	Universe of disclosure
x	Element of universal set
y	Proportion
y_{LSb}	Proportion in basic structure of a summary
y_{LSr}	Proportion in summary with restriction

8. References

Adolfsson, C., G. Arvidson, P. Gidlund, A. Norberg, and L. Nordberg. 2010. "Development and Implementation of Selective Data Editing at Statistics Sweden." In Proceedings of the European Conference on Quality in Official Statistics, May 4, 2010. Helsinki Available at: https://q2010.stat.fi/media//presentations/Norberg_et_all__Statistics_Sweden_slutversion.pdf (accessed April 2017).

Almeida, R.J., M-J. Lesot, B. Bouchon-Meunier, U. Kaymak, and G. Moyse. 2013. "Linguistic Summaries of Categorical Time Series Septic Shock Patient Data." In Proceedings of the 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2013), July 7–10, 2013. 1–8. Hyderabad.

Altin, L., M. Tiru, E. Saluveer, and A. Puura. 2015. "Using Passive Mobile Positioning Data in Tourism and Population Statistics." In Proceedings of the New Techniques and Technologies in Statistics (NTTS 2015), March 10–12, 2015. Brussels. Available at: https://ec.europa.eu/eurostat/cros/system/files/Altin-etal_abstract_ntts_2301LA_0.pdf (accessed January 2017).

Arguelles, L. and G. Triviño. 2013. "I-struve: Automatic Linguistic Descriptions of Visual Double Stars." *Engineering Applications of Artificial Intelligence* 26: 2083–2092. Doi: http://dx.doi.org/10.1016/j.engappai.2013.05.005.

Barcaroli, G., M. Scannapieco, D. Summa, and M. Scarnò. 2015. "Using Internet as a Data Source for Official Statistics: a Comparative Analysis of Web Scraping Technologies." In Proceedings of the New Techniques and Technologies in Statistics (NTTS 2015), March 10–12, 2015. Brussels. Available at: https://ec.europa.eu/eurostat/cros/system/files/Barcaroli-etal_WebScraping_Final_unblinded.pdf (accessed February 2017).

Bavdaž, M. (editor). 2011. *Final Report Integrating Findings on Business Perspectives Related to NSIs Statistics*. Brussels: European Commission. (Deliverable 3.2 from FP7 project BLUE-Enterprise and Trade Statistics). Blue-Ets Project: SSH-CT-2010-244767.

Bier, V. and P. Nymand-Andersen. 2011. "Communicating Statistics to Frequent Users – One Size Fits All?" In Proceedings of the Committee for the Coordination of Statistical Activities (CCSA Special Session), September 8, 2011. Luxembourg.

Boran, F.E., D. Akay, and R.R. Yager. 2016. "An Overview of Methods for Linguistic Summarization with Fuzzy Sets." *Expert Systems with Applications* 61: 356–377. Doi: http://dx.doi.org/10.1016/j.eswa.2016.05.044.

Bosc, P. and O. Pivert. 1995. "SQLf: a Relational Database Language for Fuzzy Querying." *IEEE Transactions on Fuzzy Systems* 3: 1–17. Doi: http://dx.doi.org/10.1109/91.366566.

Coddington, M. 2015. "Clarifying Journalism's Quantitative Turn." *Digital Journalism* 3: 331–348. Doi: http://dx.doi.org/10.1080/21670811.2014.976400.

Disability Rights Commission. 2004. *The Web Access and Inclusion for Disabled People – A Formal Investigation conducted by the Disability Rights Commission*. London: TSO. Available at: https://www.city.ac.uk/__data/assets/pdf_file/0004/72670/DRC_Report.pdf (accessed, May 2018).

Duraj, A., P.S. Szczepaniak, and J. Ochelska-Mierzejewska. 2015. "Detection of Outlier Information Using Linguistic Summarization." In Proceedings of the 11th International Conference Flexible Query Answering Systems (FQAS 2015), October 26–28, 2015. 101–113. Cracow.

EU Guide. 2015. *User guide to the SME Definition*. Luxembourg: Publications Office of the European Union. Available at: http://ec.europa.eu/growth/tools-databases/newsroom/cf/itemdetail.cfm?item_id=8274&lang=en (accessed November, 2016).

Galindo, J., A. Urrutia, and M. Piattini. 2006. *Fuzzy Databases—Modeling. Design and Implementation*. Hershey: Idea Group Publishing.

George, R. and R. Srikanth. 1996. "Data Summarization Using Genetic Algorithms and Fuzzy Logic." In *Genetic Algorithms and Soft Computing*, edited by F. Herrera and J.L. Verdegay, 599–611. Heidelberg: Physica–Verlag.

Glöckner, I. 2006. *Fuzzy Quantifiers – A Computational Theory*. Berlin Heidelberg: Springer-Verlag.

GSIM. 2013. *Generic Statistical Information Model (GSIM): Specification*. Geneva: United Nations Economic Commission for Europe (UNECE). Available at: http://www1.unece.org/stat/platform/display/gsim/GSIM+Specification (accessed February 2017).

Goebel, R., A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger. 2018. "Explainable AI: The New 42?" In *Machine Learning and Knowledge Extraction, Springer Lecture Notes in Computer Science LNCS 11015*, edited by A. Holzinger, P. Kieseberg, A. Tjoa, and E. Weippl, 295–303. Cham: Springer.

Graefe, A. 2016. *Guide to Automated Journalism*. New York: Tow Center for Digital Journalism. Available at: https://www.cjr.org/tow_center_reports/guide_to_automated_journalism.php (accessed April 2018).

Heimgärtner, R., A. Holzinger, and R. Adams. 2008. "From Cultural to Individual Adaptive End-User Interfaces: Helping People with Special Needs." In Proceedings of the 11th International Conference on Computers Helping People with Special Needs (ICCHP 2008), July 9–11, 2008. 82–89. Linz.

Hirota, K. and W. Pedrycz. 1999. "Fuzzy Computing for Data Mining." *Proceedings of IEEE* 87: 1575–1600. Doi: http://dx.doi.org/10.1109/5.784240.

Holzinger, A. 2002. "User-Centered Interface Design for Disabled and Elderly People: First Experiences with Designing a Patient Communication System (PACOSY)." In Proceedings of the 8th International Conference on Computer Helping People with Special Needs (ICCHP 2002), July 15–20, 2002. 33–40. Linz.

Holzinger, A., B. Malle, P. Kieseberg, P.M. Roth, H. Müller, R. Reihs, and K. Zatloukal. 2017. "Machine Learning and Knowledge Extraction in Digital Pathology needs an integrative approach." In *Towards Integrative Machine Learning and Knowledge Extraction*, edited by A. Holzinger, R. Goebel, M. Ferri, and V. Palade, 13–50. Cham: Springer.

Hudec, M. 2013. "Improvement of Data Collection and Dissemination by Fuzzy Logic." In Proceedings of the Joint UNECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS), April 22–24, 2013. Paris and Bangkok. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_3_Slovakia.pdf (accessed January 2017).

Hudec, M. 2016. *Fuzziness in Information Systems – How to Deal with Crisp and Fuzzy Data in Selection, Classification, and Summarization*. Cham: Springer.

Hudec, M. 2017. "Merging Validity and Coverage for Measuring Quality of Data Summaries." In *Information Technology and Computational Physics*, edited by P. Kulczycki, L.T. Kóczy, R. Mesiar, and J. Kacprzyk, 71–85. Cham: Springer.

Hudec, M. and D. Praženka. 2016. "Collecting and Managing Fuzzy Data in Statistical Relational Databases." *Statistical Journal of the IAOS* 32: 245–255. Doi: http://dx.doi.org/10.3233/SJI-160956.

Hudec, M. and V. Torres Van Grinsven. 2013. "Business' Participants Motivation in Official Surveys by Fuzzy Logic." In Proceedings of the 1st Eurasian Multidisciplinary Forum (EMF 2013), October 24–26, 2013. 42–52. Tbilisi.

Kacprzyk, J. and P. Strykowski. 1999. "Linguistic Data Summaries for Intelligent Decision Support." In Proceedings of the fourth European Workshop on Fuzzy Decision

Analysis and Recognition Technology for Management, Planning and Optimization (EFDAN 1999), June 14–15, 1999. 3–12. Dortmund.

Kacprzyk, J., A. Wilbik, and S. Zadrożny. 2006. "Linguistic Summarization of Trends: A Fuzzy Logic Based Approach." In Proceedings of the 11th Information Processing and Management of Uncertainty in Knowledge Based Systems (IPMU 2006), July 2–7, 2006. 2166–2172. Paris.

Kacprzyk, J. and R.R. Yager. 2001. "Linguistic Summaries of Data Using Fuzzy Logic." *International Journal of General Systems* 30: 133–154. Doi: http://dx.doi.org/10.1080/03081070108960702.

Kacprzyk, J. and S. Zadrożny. 1995. "FQUERY for Access: Fuzzy Querying for Windows-Based DBMS." In *Fuzziness in Database Management Systems*, edited by P. Bosc and J. Kacprzyk, 415–433. Heidelberg: Physica-Verlag.

Kacprzyk, J. and S. Zadrożny. 2005. "Linguistic Database Summaries and Their Protoforms: Towards Natural Language Based Knowledge Discovery Tools." *Information Sciences* 173: 281–304. Doi: http://dx.doi.org/10.1016/j.ins.2005.03.002.

Kacprzyk, J. and A. Ziółkowski. 1986. "Database Queries with Fuzzy Linguistic Quantifiers." *IEEE Transactions Systems, Man and Cybernetics SMC-16* 3: 474–479. Doi: http://dx.doi.org/10.1109/tsmc.1986.4308982.

Klement, E.P., R. Mesiar, and E. Pap. 2005. "Triangular Norms: Basic Notions and Properties." In *Logical, Algebraic, Analytic, and Probabilistic Aspects of triangular Norms*, edited by E.P. Klement and R. Mesiar, 17–60. Amsterdam: Elsevier.

Lesot, M-J., G. Moyse, and B. Bouchon-Meunier. 2016. "Interpretability of Fuzzy Linguistic Summaries." *Fuzzy Sets and Systems* 292: 307–317. Doi: http://dx.doi.org/10.1016/j.fss.2014.10.019.

Liu, B. 2011. "Uncertain Logic for Modeling Human Language." *Journal of Uncertain Systems* 5: 3–20. Available at: www.jus.org.uk (accessed September 2012).

Meyer, A. and H.J. Zimmermann. 2011. "Applications of Fuzzy Technology in Business Intelligence." *International Journal of Computers, Communications & Control* VI(3): 428–441. Doi: http://dx.doi.org/10.15837/ijccc.2011.3.2128.

Moyse, G., M-J. Lesot, and B. Bouchon-Meunier. 2013. "Mathematical Morphology Tools to Evaluate Periodic Linguistic Summaries." In *Flexible Query Answering Systems*, edited by H.L. Larsen, 257–268. Berlin Heidelberg: Springer-Verlag.

Niewiadomski, A. 2002. "Appliance of Fuzzy Relations for Text Documents Comparing." In Proceedings of the 6th Conference on Neural Networks and Soft Computing (ICNNSC' 2002), June 11–15, 2002. Zakopane.

Niewiadomski, A., J. Ochelska, and P.S. Szczepaniak. 2006. "Interval-Valued Linguistic Summaries of Databases." *Control and Cybernetics* 35: 415–443. Available at: http://matwbn.icm.edu.pl/ksiazki/cc/cc35/cc35212.pdf (accessed June 2016).

Raschia, G. and N. Mouaddib. 2002. "SAINTETIQ: A Fuzzy Set-Based Approach to Database Summarization." *Fuzzy Sets and Systems* 129: 137–162. Doi: https://doi.org/10.1016/S0165-0114(01)00197-X.

Rasmussen, D. and R.R. Yager. 1997. "Summary SQL – A Fuzzy Tool for Data Mining." *Intelligent Data Analysis* 1: 49–58. Doi: http://dx.doi.org/10.1016/S1088-467X(98)00009-2.

Ross, M.P. 2009. "Official Statistics in Malta – Implications of Membership of the European Statistical System for a Small Country/NSI." In Proceedings of the 95th DGINS Conference, October 1, 2009. Malta. Available at: https://ec.europa.eu/eurostat/documents/1001617/4339944/MPR-opening-address-00909.pdf/7c298770-0869-415c-9833-d702e8b3ce9e (accessed October, 2016).

Scanu, M. and C. Casagrande. 2016. "The Generic Statistical Information Model (GSIM): State of Application of the Standard." In Workshop on Implementing Standards for Statistical Modernisation, 21–23 September 2016. Geneva. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2016/mtg4/Paper_17_Italy_-_The_Generic_Statistical_Information_Model__GSIM__and_the_Sistema_Unitario.pdf (accessed March 2017).

SDMX. 2012. *SDMX 2.1 User Guide, SDMX 2.1 Documentation*. SDMX Consortium. Available at: https://sdmx.org/?page_id=1119 (Accessed January 2017).

Schweizer, B. and A. Sklar. 1983. *Probabilistic Metric Spaces*. Amsterdam: North-Holland.

Schield, M. 2011. "Statistical Literacy: A New Mission for Data Producers." *Statistical Journal of the IAOS* 27: 173–183. Doi: http://dx.doi.org/10.3233/SJI-2011-0732.

Smits, G., O. Pivert, and T. Girault. 2013. "ReqFlex: Fuzzy Queries for Everyone." In Proceedings of the 39th International Conference on Very Large Data Bases, 26–30 August, Trento.

Torres van Grinsven, V. and G. Snijkers. 2015. "Sentiments and Perceptions of Business Respondents on Social Media: An Exploratory Analysis." *Journal of Official Statistics* 31: 283–304. Doi: http://dx.doi.org/10.1515/jos-2015-0018.

Wu, D., J.M. Mendel, and J. Joo. 2010. "Linguistic Summarization Using If-Then Rules." In Proceedings of the 2010 IEEE International Conference on Fuzzy Systems, July 18–23, 2010. 1–8. Barcelona.

Yager, R.R. 1982. "A New Approach to the Summarization of Data." *Information Sciences* 28: 69–86. Doi: http://dx.doi.org/10.1016/0020-0255(82)90033-0.

Yager, R.R. 1984. "General Multiple-Objective Decision Functions and Linguistically Quantified Statements." *International Journal of Man-Machine Studies* 21: 389–400. Doi: http://dx.doi.org/10.1016/S0020-7373(84)80066-8.

Yager, R.R. 1988. "On Ordered Weighted Averaging Operators in Multicritera Decision Making." *IEEE Transactions on Systems, Man and Cybernetics,* SMC-18: 183–190. Doi: http://dx.doi.org/10.1080/03081070108960702.

Yager, R.R., M. Ford, and A.J. Canas. 1990. "An Approach to the Linguistic Summarization of Data." In Proceedings of the 3rd International Conference of Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU 1990), July 2–6, 1990. 456–468. Paris.

Zadeh, L.A. 1965. "Fuzzy Sets." *Information and Control* 8: 338–353. Doi: http://dx.doi.org/10.1016/S0019-9958(65)90241-X.

Zadeh, L.A. 1975. "The Concept of a Linguistic Variable and Its Application to Approximate Reasoning: Part I." *Information Sciences* 8: 199–249. Doi: http://dx.doi.org/10.1016/0020-0255(75)90036-5.

Zadeh, L.A. 1983. "A Computational Approach to Fuzzy Quantifiers in Natural Languages." *Computers & Mathematics with Applications* 9: 149–184. Doi: http://dx.doi.org/10.1016/0898-1221(83)90013-5.

Zadeh, L.A. 2001. "From Computing With Numbers to Computing With Words—From Manipulation of Measurements to Manipulation of Perceptions." In *Computing with Words*, edited by P. Wang, 35–68. New York: Wiley.

Zottoli, M., S. Laurita, and F. Monteleone. 2017. "Contestina: A Visibly Understandable Path toward More Effective Data Dissemination." In Proceedings of the New Techniques and Technologies in Statistics (NTTS 2017), March 14–16, 2017. Brussels. Available at: https://www.conference-service.com/NTTS2017/documents/agenda/data/abstracts/abstract_151.html (accessed May 2017).

Received June 2017
Revised September 2018
Accepted October 2018

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 1011–1015, http://dx.doi.org/10.2478/JOS-2018-0049

Editorial Collaborators

The editors wish to thank the following referees who have generously given their time and skills to the Journal of Official Statistics during the period October 1, 2017 – September 30, 2018. An asterisk indicates that the referee served more than once during the period.

Abraham, Katharine, JPSM, College Park, Maryland, U.S.A.
Adiguzel, Feray, Erasmus University Rotterdam, Rotterdam, the Netherlands
Baffour, Bernard, Australian National University, Canberra, Australia
Baker, Kellan, Johns Hopkins University, Baltimore, Maryland, U.S.A
Balk, Bert, Rotterdam School of Management, Erasmus University, Rotterdam, the Netherlands
Barcaroli, Giulio, Italian National Institute of Statistics, Rome, Italy
Basel, Wesley, US Census Bureau, Washington, D.C., U.S.A.
Bashir, Shakila, Forman Christian Collage, Lahore, Punjab, Pakistan
Bates, Nancy, US Census Bureau, Washington, D.C., U.S.A.*
Bautista, René, NORC at the University of Chicago, Chicago, Illinois, U.S.A.
Bavdaž, Mojca, University of Ljubljana, Ljubljana, Slovenia
Belli, Robert, University of Nebraska, Lincoln, Nebraska, U.S.A.
Beręsewicz, Maciej, Poznań University of Economics and Business, Wielkopolska, Poland*
Beresovsky, Vladislav, CDC, Atlanta, Georgia, U.S.A.
Berzofsky, Marcus, RTI International, Research Triangle Park, North Carolina, U.S.A.
Białek, Jacek, University of Lodz, Lodz, Poland
Bilgen, Ipek, NORC at the University of Chicago, Chicago, Illinois, U.S.A.
Bivand, Roger, NHH Norwegian School of Economics, Bergen, Norway
Blackwell, Louisa, Office for National Statistics, Hampshire, UK
Blumberg, Stephen, CDC, Atlanta, Georgia, U.S.A.
Bonnéry, Daniel, University of Maryland at College Park, College park, Maryland, U.S.A.
Borsi, Lisa, Trier University, Trier, Germany
Brenner, Philip, University of Massachusetts Boston, Boston, Massachusetts, U.S.A.
Broome, Jessica, University of Michigan, Ann Arbor, Michigan, U.S.A.
Brown, Gary, Office for National Statistics, Newport, UK
Bruil, Arjan, Statistics Netherlands, The Hague, the Netherlands
Buelens, Bart, Statistics Netherlands, Heerlen, the Netherlands
Buono, Dario, Eurostat, Mamer, Luxembourg*
Bycroft, Christine, Statistics New Zealand (Stats NZ), Wellington, New Zealand
Böhning, Dankmar, University of Southampton, Southampton, UK
Callegaro, Mario, Google, London, UK
Cantor, David, Westat, Rockville, Maryland, U.S.A.

Glick, Jennifer, Johns Hopkins University, Baltimore, Maryland, U.S.A.
Goodhart, Charles, London School of Economics, London, UK
Graham, Patrick, Statistics New Zealand, Christchurch, New Zealand
Grossenbacher, Armin, Dissemination, Köniz, Switzerland
Gubman, Yury, Central Bureau of Statistics, Jerusalem, Israel
Hand, David, Imperial College, London, UK
Haque, Shovanur, Queensland University of Technology, Brisbane, Australia
Hasslett, Stephen, Massey University, Palmerston North, Manawatu, New Zealand
He, Yulei, National Center for Health Statistics, CDC, Hyattsville, Maryland, U.S.A.
Heldal, Johan, Statistics Norway, Oslo, Norway
Hill, Robert, University of Graz, Graz, Austria*
Holbrook, Allyson, University of Chicago, Chicago, Illinois, U.S.A.
Huang, Ning, Statistics Canada, Ottawa, Ontario, Canada
Höhne, Jan Karem, University of Göttingen, Göttingen, Germany*
Jang, Don, NORC at the University of Chicago, Bethesda, Maryland, U.S.A.*
Jans, Matt, University of California, Los Angeles, California, U.S.A.
Jansen, Ronald, UN Statistics Division, New York, U.S.A.
Janssen, Eric, OFDT, Saint-Denis, France
Johansson, Anton, Statistics Sweden, Örebro, Sweden
Joyce, Patrick, U.S. Census Bureau, Washington, D.C., U.S.A.
Joye, Dominique, University of Lausanne, Lausanne, Switzerland
Kamanda, Amie, Office for National Statistics, Fareham, UK
Kaplan, Robin, Bureau of Labor Statistics, Washington, D.C., U.S.A.
Kaputa, Stephen, U.S. Census Bureau, Washington, D.C., U.S.A.
Khan, Diba, CDC, Atlanta, Georgia, U.S.A.
Kim, Jong-Min, University of Minnesota, Morris, Minnesota, U.S.A.
Kirby, Graham, University of St. Andrews, St. Andrews, Fife, UK
Klee, Mark, U.S. Census Bureau, Washington, D.C., U.S.A.
Klein, Martin, U.S. Census Bureau, Washington, D.C., U.S.A.
Kowarik, Alexander, Statistics Austria, Vienna, Austria
Laaksonen, Seppo, University of Helsinki, Helsinki, Finland
Lee, Kimya, U.S. Office of Personnel Management, Washington, D.C., U.S.A.
Lee, Sunghee, University of Michigan, Ann Arbor, Michigan, U.S.A.*
Levell, Peter, Institute for Fiscal Studies, London, UK
Lewis, Daniel, Office for National Statistics, Newport, UK
Lewis, Taylor, U.S. Office of Personnel Management, Washington D.C., U.S.A.
Li, Feng, Stockholm University, Stockholm, Sweden
Liiv, Innar, Tallinn University of Technology, Tallinn, Estonia
Liseo, Brunero, University of Rome, Rome, Italy*
Lisic, Jonathan, USDA, Washington, D.C., U.S.A.
Liu, Benmei, National Institutes of Health, Rockville, Maryland, U.S.A.*
Liu, Mingnan, Survey Monkey, Palo Alto, California, U.S.A.
Loosveldt, Geert, KU Leuven, Leuven, Belgium
Lundquist, Peter, Statistics Sweden, Stockholm, Sweden
Macdonald, Mitch. Simon Fraser University, Burnaby, BC, Canada

Scanu, Mauro, ISTAT, Rome, Italy
Scarnò, Marco, Cineca SCAI, Rome, Italy
Schechter, Susan, NORC at the University of Chicago, Bethesda, Maryland, U.S.A.
Scheffer, Fredrik, Statistics Sweden, Stockholm, Sweden*
Schierholz, Malte, the Research Institute of the Federal Employment Agency, Nuremberg, Germany*
Schnell, Rainier, University of Duisburg-Essen, Duisburg, Germany
Schober, Michael, New School for Social Research Psychology, New York, U.S.A.
Scholtus, Sander, Statistics Netherlands, The Hague, the Netherlands*
Schouten, Barry, Statistics Netherlands, The Hague, the Netherlands*
Sigh, Sarjinder, Texas AM University-Kingsville, Kingsville, Texas, U.S.A.
Singleton, Ann, University of Bristol, Bristol, UK
Slud, Eric, University of Maryland, College Park, Maryland, U.S.A.*
Smeets, Marc, Statistics Netherlands, Heerlen, the Netherlands
Smith, Duncan, University of Manchester, Manchester, UK
Suesse, Thomas, University of Wollongong, Wollongong, Australia
Thalji, Lisa, RTI International, Research Triangle Park, North Carolina, U.S.A.*
Tongur, Can, Statistics Sweden, Stockholm, Sweden
Toth, Daniell, U.S. Bureau of Labor Statistics, Washington, D.C., U.S.A.
Toulemon, Laurent, French Institute for Demographic Studies, Paris, France
Tourangeau, Roger, Westat, Rockville, Maryland, U.S.A.
Tuoto, Tiziana, ISTAT, Rome, Italy*
Tønnessen, Marianne, Statistics Norway, Oslo, Norway
Übi, Jaan, the University of Tartu, Tartu, Estonia*
Van der Loo, Mark, Statistics Netherlands, The Hague, The Netherlands*
Wagner, James, University of Michigan, Ann Arbor, Michigan, U.S.A.*
Wallgren, Anders and Britt, BA Statistiksystem AB, Vintrosa, Sweden
Wang, Kevin, RTI, Research Triangle Park, NC, U.S.A.
Watson, Nichole, University of Melbourne, Melbourne, Victoria, Australia
Wilhelm, Matthieu, University of Neuchâtel, Neuchâtel, Switzerland
Willis, Gordon, National Institutes of Health, Rockville, Maryland, U.S.A.
Wilson, Bianca, University of California, Los Angeles, California, U.S.A.
Wilson, Tom, Charles Darwin University, Darwin, Australia*
Winglee, Marianne, Westat, Rockville, Maryland, U.S.A.
Winkler, William, U.S. Census Bureau, Washington, D.C., U.S.A.
Virgillito, Antonino, ISTAT, Rome, Italy
Wiśniowski, Arkadiusz, University of Southampton, Southampton, UK
Wissoker, Douglas, Urban Institute, Washington, D.C., U.S.A.
Xie, Yingfu, Statistics Sweden, Stockholm, Sweden
Zabala, Felipa, Statistics New Zealand, Wellington, New Zealand*
Zadrozny, Peter, U.S. Bureau of Labor Statistics, Washington, D.C., U.S.A.
Zeelenberg, Kees, Statistics Netherlands, The Hague, the Netherlands
Zhang, Guangyu, National Center for Health Statistics, Washington D.C., U.S.A.
Zimmermann, Thomas, University of Trier, Trier, Germany
Öberg, Sebastian, Stockholm, Sweden