Statistica Neerlandica (2012) Vol. 66, nr. 1, pp. 2–7
doi:10.1111/j.1467-9574.2011.00505.x

Methodological challenges of register-based research

Bart F. M. Bakker[*]

Statistics Netherlands, VU University Amsterdam, P.O. Box 24500, 2490 HA, The Hague, The Netherlands

Piet J. H. Daas[†]

Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands

1 Editorial introduction

Administrative data have become more and more important for both official statistics and academic research. However, the, more administrative registers are used, the more methodological issues arise (Bakker, 2009; Daas, Kuijvenhoven and Zeelenberg, 2010a). This was apparent at the 2011 congress of the International Statistical Institute, when more than ten sessions focused explicitly on the quality, linking and analyses of administrative data and many others discussed studies based on administrative data. Moreover, the increasing use of administrative data is one of the core transitions in research projects funded in Europe. In both the European Statistical System networks (ESSnets), launched by Eurostat, and the Framework Programmes, funded by the European Commission, the primary focus is no longer solely on data collected by questionnaires. Both Eurostat and the European Commission promote the development of new theory and applications in the area of the use of administrative data. Examples of recent initiatives are the ESSnet on 'Data Integration', the ESSnet 'On the Use of Administrative and Accountants Data' and the seventh framework Programme BLUE Enterprise and Trade Statistics. Each of these projects has a strong component that focuses on key methodological issues in the use of administrative data and registers for research purposes.

However, the theory of register-based research is scarcely out of the egg. After the first mention of this problem in the mid-1990s (Hoffmann, 1995; Statistics Denmark, 1995), the rudiments of such a theory were merely sketched in the first decade of this century (UNECE, 2007; Wallgren and Wallgren, 2007). It will take a serious period of time to develop it further and take it to an acceptable level; just as it took time to develop and accept the use of probabilistic ideas in to the

*bbkr@cbs.nl/b.f.m.bakker@vu.nl
†Pjh.daas@cbs.nl

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

sampling survey domain (Kruskal and Mosteller, 1980). The state of the art in the domain of registers is at this moment more a set of best practices from countries that have used administrative data over a longer period of time; like the Nordic countries and the Netherlands. One of the first obstacles to be overcome is that there is no consensus, even on the basic terminology. What are administrative data? Are terms like register, registry and administrative data interchangeable, or should we use only the term secondary data? Is a survey always a sample survey, or can this term also be used for register-based data collection? The main reason for the lack of consensus in this respect is that although theories and practices urged the use of terminology, it was developed in relative isolation.

In our opinion a platform, such as a journal, is needed that focuses solely on register-based research. This would greatly stimulate the creation of a common terminology which is a key to the ultimate goal: the development of a theoretical framework. Apart from this, a dedicated journal can also function as a scientific forum. To foster the harmonization process, papers in which comments are given from different viewpoints and from different scientific disciplines should be welcomed. Such a journal would greatly assist the developments of new theories in all research areas that use administrative data on a regular basis as a source of information. Examples of such studies can be found in statistical, social, criminological, medical, epidemiological and economical research areas, to name a few. To support the application of the theories developed, empirical studies should form an important part of such a journal. As a start this special issue of *Statistica Neerlandica* brings together six articles on the use of administrative data for statistics.

2 Methodological issues

The more administrative data are used, the more methodological issues arise (Bakker, 2009; Daas *et al.*, 2010a). One of the most important issues is related to quality. As the researcher was not involved in data collection and in metadata definition, to name just two essential issues, the information in the data source is at risk of being misinterpreted. As a result of this, administrative data need to be carefully evaluated prior to use (Daas, Ossen and Tennekes, 2010b). Errors in registers are also a major concern to users. Similar to leading publications on errors in surveys (Groves *et al.*, 2004), representation and measurement errors can be distinguished in register-based research. Representation errors occur if the measured population elements differ from the target population, while measurement errors occur if the attributes of a population are measured imperfectly. In this section, we provide an overview of the errors identified by Bakker (2010) for registers in a first attempt to sketch a theoretical framework.

On the representation side, we are looking for coverage problems that cause biased results. For instance, the population of the Population Register includes all inhabitants living in the Netherlands for at least 4 months. However, the administrative register does not include the 'illegal' population, even though it is part of the target population (van der Heijden *et al.*, 2006). This results in under-coverage of the target population. Administrative delay in registers can lead to under-coverage (e.g. birth and immigration) and over-coverage (e.g. death and emigration).

Administrative records in different registers can be combined by linking. In most cases the linking key used is a personal identification number (PIN), or – if the register lacks such a number – a combination of variables, for example, birth date, sex and address. Two types of linking errors can occur: missed links and mislinks (Fellegi and Sunter, 1969; Arts, Bakker and van Lith, 2000). Missed links are cases where the matching record is not found. The resulting errors are similar to those caused by non-response in surveys: links that are non-randomly missed will bias the outcomes. Mislinks occur if records of two non-matching (different) elements are combined.

On the measurement side, the fundamental problem is that statistical concepts do not fully match the administrative ones which have been defined with other purposes in mind (Daas and Ossen, 2011). Moreover, the researchers have little or no influence on the definitions and the production processes of the administrative data holder. As a result, little is known *a priori* on the quality of administrative data (Grünewald and Körner, 2005; Daas *et al.*, 2010b).

The size of errors in administrative registers also depends on the control processes that the administrative data holder executes. Of course, all kinds of checks in the administrative process can correct for errors made earlier on, such as the interview during which the data were collected. For example, an employment officer may demand to see documents (e.g. pay-slips, diplomas or certificates) to verify the information a job-seeker has given to prevent possible errors. Similarly, a tax employee may decide to use information stored in another administrative register to reveal inconsistencies in the source studied. Any remaining irregularities are mostly corrected in the administrative register in consultation with the reporting instance or person concerned. In some cases, the recorded data are audited by accountants or other inspectors. These administrative protocols are formulated to maximize the quality of the measurement of the variables that are important for the purpose of the register keeper. Therefore, one may assume that the quality of these data is better than that of variables of lesser interest. An illustrative example is the quality of the variable 'period of employment' registered by the income tax authorities: as the tax sum in the Netherlands does not depend on this information, the tax authorities do not pay much attention to the correctness of the value for this variable. Therefore, these data are at risk of poor quality.

The size of errors also depends on the interest of the registered persons. If it is in the interest of a registered person to be registered – incorrectly – in a specific way, the probability of this type of misregistration occurring increases. If one is interested in the data quality of a specific administrative register, it is important to specify the interests of both register keeper and registered. This information can be useful to formulate hypotheses for a potential bias in the data.

3 In this issue

This special issue of *Statistica Neerlandica* brings together six articles on the use of administrative data. It can be conceived as a pilot issue for a dedicated journal on register-based research. The contributions vary in scope and subject: one article sketches a theoretical framework, one is on representation problems, two articles examine measurement problems and two comment on one or more of the others.

Zhang starts with a theoretical article discussing data integration on a conceptual level. In particular, he presents a two-phase life-cycle model for integrated statistical microdata. This should provide a framework for the various potential error sources and is an extension of the error sources framework discussed before. The author's aim is to provide a framework that allows one to disentangle and clarify the various potential types of errors and their origin, and to detect the relation to the different production processes of integrated statistical microdata. The two-phase model includes the life-cycle model of survey data (Groves *et al.,* 2004) as a special case. The first phase focuses on the data in each single source while the second phase concerns the possible integration of data from different sources. The latter often involves a transformation of the initial input data. Furthermore, some concepts and topics for quality assessment beyond the ideal of error-free data are outlined. A shared understanding of these issues will hopefully help to collocate and co-ordinate efforts in future research and development.

Bakker presents a method to estimate the size of the validity of administrative variables. One of the problems that might occur in administrative data is that they are biased and of low validity. Although this problem is often addressed qualitatively, the validity is seldom measured quantitatively. By applying classical test theory and structural equation models to linked survey and register data, he is able to measure the same concepts and estimate the construct validity of both survey and administrative variables. He presents an empirical example in which the construct validity of age, gender, educational attainment and wages is simultaneously determined. The analyses reveal that registered educational attainment and wages show some bias, but not higher than the bias found in the survey. The author concludes that the method is suitable for this purpose on the condition that the model used is grounded in theory, and the strength of the relationships between the variables in the model is more or less known.

The article by Berka *et al.* investigates the usefulness of the Dempster–Shafer theory to assess the quality of databases with multiple underlying sources for a single attribute, with particular focus on the register-based Austrian Census. This 'fuzzy' approach allows them to simultaneously take into account the uncertainty associated with support and conflict between underlying registers and the quality of the underlying registers when summarized in the form of several indicators. The application results in both quality measures and plausibility intervals for the value of an attribute as derived from the underlying sources on the basis of a rule set. This was done for the values of the Census-attribute sex which was based on four underlying registers. They found that the support for the values for sex in the Census is relatively high. From this, it follows that the rule set is rather good.

Kim and Chambers look into a problem that occurs when multiple sources for the same target population are linked with probability-based methods. Standard methods of regression analysis can lead to serious bias if they do not take the effect of linkage error into account. Until now, current approaches to modifying standard methods of regression analysis assume that all records on the two sources combined are linked. The paper extends these ideas to accommodate a situation where more than two sources are probabilistically linked and that linkage is incomplete. Their results indicate that the developed estimation methods are successful in eliminating the bias induced by linkage errors, provided we know, or are able to unbiasedly estimate, the appropriate linkage probabilities. They also correct the biases introduced by both sampling and non-linkage via the introduction of appropriate weights, assuming that these processes are non-informative and are independent of one another. However, it is also clear that these bias correction methods generally lead to larger variances.

As is clear from the papers introduced before, a considerable amount of methodological work has already been performed in the area of register-based research. We expect this – and the number of researchers working in this fascinating area – to increase in the coming years. It would be stimulating to all involved to have a joint platform on which they could inform one another of the results obtained, discuss the progress made and exchange ideas. Last, but certainly not the least, we would like to thank all the authors and reviewers for their contribution to this issue. We hope you enjoy reading the papers in this special issue as much as we did.

References

Arts, K., B. F. M. Bakker and E. van Lith (2000), Linking administrative registers and household surveys, in: P. Al and B. F. M. Bakker (eds), *Re-engineering social statistics by micro-integration of different sources*, Netherlands Official Statistics **15**, 16–22.

Bakker, B. F. M. (2009), *Trek alle registers open!*, Vrije Universiteit, Amsterdam.

Bakker, B. F. M. (2010), *Micro-integration. State of the Art*, United Nations, New York.

Daas, P. J. H. and S. J. L. Ossen (2011). Metadata quality evaluation of secondary data sources, *International Journal for Quality Research* **5**, 57–66.

Daas, P. J. H., L. Kuijvenhoven, and K. Zeelenberg (2010a), Registers: Onderzoeksagenda voor de toekomst, in: B. F. M. Bakker and L. Kuijvenhoven (eds), *Registers in sociaalwetenschappelijk onderzoek, Mogelijkheden en valkuilen*, Vrije Universiteit Amsterdam/Centraal Bureau voor de Statistiek, Den Haag, pp. 173–186.

Daas, P. J. H., S. J. L. Ossen, and M. Tennekes (2010b), *The determination of administrative data quality: recent results and new developments*. Paper for the European Conference on Quality in Official Statistics 2010, Helsinki, Finland.

Fellegi, I. and A. Sunter (1969), A theory of record linkage, *Journal of the American Statistical Association* **64**, 1183–1210.

Groves, R. M., F. J. Fowler JR.,, M. P. Couper, J. M. Lepkowski, E. Singer and R. Tourangeau (2004), *Survey methodology*, Wiley Interscience, New York.

Grünewald, W. and T. Körner (2005), Quality on its way to maturity: results of the European conference on quality and methodology in official statistics (Q2004). *Journal of Official Statistics* **21**, 747–759.

Hoffman, E. (1995) We must use administrative data for official statistics – but how should we use them?, *Statistical Journal of the United Nations ECE* **12**, 41–48.

Kruskal, W. and F. Mosteller (1980), Representative sampling, IV: The history of the Concept in Statistics, 1895–1939, *International Statistical Review* **48**, 169–195.

Statistics Denmark (1995), *Statistics on persons in Denmark, a register-based statistical system*, Eurostat Publication, Brussels/Luxembourg.

UNECE (2007), *Register-based statistics in the Nordic countries – review of best practices with focus on population and social statistics*, United Nations Publication, Genova.

van der Heijden, P. G. M., G. van Gils, M. Cruijff and D. Hessen (2006), *Een schatting van het aantal in Nederland verblijvende illegale vreemdelingen in 2005*, IOPS Universiteit Utrecht, Utrecht.

Wallgren, A. and B. Wallgren (2007), *Register-based statistics: administrative data for statistical purposes*, John Wiley & Sons, Chichester, UK.

Received: 14 September 2011. Revised: 26 September 2011.