To the Scientific Council of the Faculty of Informatics and Electronic Economy, Poznań University of Economics and Business

Review of Doctoral dissertation “Internet data sources for real estate market statistics”, authored by Maciej Beręsewicz

The doctoral dissertation presents results on methodological research concerning new Internet-based data sources for official statistics. The Internet as raw data source for official statistics can involve improvement in cost efficiency of statistics production and better timeliness in publication of statistics. However, theory for this new data source is premature and its real use in official statistics is in its infancy. Therefore, scientific research on theoretical and methodological aspects on the Internet as data source, and experimentation on the various uses of these data in official statistics, is of great importance. Statistical communities are aware of this fact and academic research and empirical testing and experimentation are increasingly introduced under various data infrastructures. The work by the author of the doctoral dissertation is timely and provides an important contribution to scientific research in the area.

The doctoral dissertation is prepared in economics discipline. In this review I concentrate on statistical issues, in particular: (1) quality of the scientific problem and related methodological research problems, (2) knowledge on appropriate statistical methods in the field, and (3) ability to sound empirical application of the selected methods. It should be noted that the coverage of statistical methodology needed is broad including survey sampling and survey methodology, econometric methods, mathematical statistics, data integration methods, data management techniques, data cleaning methods, statistical estimation, modelling and inference, as well as and statistical computing and visualization.

The dissertation contains an introductory section, five numbered chapters, a conclusive section and a bibliography section. A set of tables, figures and diagrams as well as definitions of important concepts and selected R codes are inserted in an appendix section. The dissertation is extensive (249 numbered pages) and in addition to theoretical and methodological considerations the document contains a lot of detailed technical materials and data source specific findings, indicating the author's in-depth look into the problem.

In the introductory section, the author states and motivates the scientific problem, presents the goals of research and formulates the hypotheses. Real estate market is chosen as the target subject matter area in economics. The evaluation of the Internet as data source for real estate market statistics is stated as the scientific problem. Three methodology specific research objectives are: (1) Identification of non-sampling error in Internet data sources about the secondary real estate market, (2) Assessment of the representativeness of Internet data sources for the secondary real estate market, and (3) Assessment of bias in Internet data sources for the secondary real estate market. Real estate market statistics is a reasonable choice because of their economic relevance, and suitable real data from Internet data sources are available.

Based on the scientific problem the author states his hypotheses. The main hypothesis is as follows: Internet data sources enable acceptable estimation of real estate market characteristics. Four more specific hypotheses are stated: (1) The approach proposed by the author to

measure representativeness can be effectively used to assess the representativeness of Internet data sources about the secondary real estate market, (2) Internet data sources are biased and this bias varies between sources and domains, (3) Self-selection in Internet data sources about the real estate market is informative (depends on the target variable), (4) Internet data sources can be used to estimate the offer price per $m^2$ in the secondary real estate market with acceptable error measured by absolute relative bias. The specific hypotheses are appropriately derived from the main hypothesis and they appear testable.

As a whole, the scientific problem and hypothesis setting are well structured. The author examines Internet data sources in the framework of survey statistics theory and methodology, providing a fresh and useful view to the problem. The view chosen offers the use of such concepts as target population, multiple frames, representativeness and coverage, sample, self-selection mechanism and selection bias. The author incorporates these conceptual tools in the study design for an empirical verification of the stated hypotheses.

The study design consists of (1) theoretical and conceptual analysis of the various data source types in the framework of official statistics and the identification of a new potentially useful data source type (called Internet data source), (2) introduction of methodological and technical tools and measures to identify and analyze the sources of non-sampling errors and biases caused by these errors in generic Internet data source, and (3) using selected statistical methods and available data sources, conduct empirical analysis of important non-sampling errors and biases in selected real Internet data sources.

Availability of real data from suitable Internet data sources and respective auxiliary (reference) data constitute prerequisites for implementing the study design. Three major online advertising services, which are operating in real estate market in Poland, are chosen. Representativeness of the Internet data sources and bias properties of statistics obtained from them was assessed by using benchmark data from sample surveys and register data sources. The study was restricted to twelve cities of Poland.

Because *Internet data source* (IDS) is not an established concept in statistical literature, a starting point for the development of theoretically sound and applicable methodology requires a clear definition of this data source type such that the unique properties relative to the existing data source types are clearly indicated. A suitable approach is to classify the various data types based on how the data are created. A definition of *Internet data source* is as follows: an Internet data source is a self-selected (non-probabilistic) sample, which is created through the Internet and maintained by entities external to NSIs and administrative regulations (p. 5). This definition is more comprehensive than that presented earlier by the author (Beręsewicz, Australian Journal of Statistics 2015) because the important property of non-probabilistic character of Internet data source is emphasized. To make Internet data sources manageable for statistical purposes, it also has been necessary to define the type and scope of elements and variables in specific Internet data sources. It appears that Internet data sources resemble properties (and problems) similar to those of traditional data sources based on administrative registers. It should be noted that the definition of Internet data source includes a reference to official statistics, which may restrict the scope of the concept of Internet data source beyond official statistics framework.

A hierarchy of relevant populations underlying Internet data source is defined in Diagram 2.1. It appears that a specific Internet data source (or a set of overlapping Internet data sources) does not necessarily cover entirely the intended target population but constitutes a subgroup of it. To emphasize the conceptual difference w.r.t. *probability sample*, the observed subgroup is called *self-selection sample*. The self-selection property of Internet data source introduces bias in estimates called *selection bias*.

The concept of *bias* deserves special attention because of its central role in the dissertation. Moreover, in literature the term is often used with different meanings, depending on the context. In *design-based* statistical inference for well-defined finite target population, bias refers to a theoretical property of an estimator of finite population parameter. An estimator is called *design unbiased* if in repeated sampling with the same probability sampling design from a finite population, the expectation of the estimator coincides with the true parameter. Many popular estimators of population (or sub-population) totals are constructed design unbiased (the Horvitz-Thompson estimator is a good example).

My interpretation on not referring by the author to design bias of estimators is as follows. The theoretically convenient property of design unbiasedness of design-based estimators is jeopardized or even demolished under informative unit nonresponse (if the response mechanism correlates with the variable of interest) causing unknown selection bias. As a consequence, the originally design unbiased estimator produces biased estimation for the target parameter. This also occurs for *model-based* estimators that are design biased by the construction principle (in this case we should address both the design bias due to the estimator and the selection bias due to informative unit nonresponse). Thus, selection bias is not a property of the original probability sampling procedure or the estimator itself but is realized in estimates via the selection mechanism of the procedure that generates the observed data. In Internet data sources, which are non-probability samples by definition, design bias looses relevance but the selection bias becomes dominant. Therefore, the author restricts discussion to biases occurring due to the self-selection mechanism. The author recognizes that both design-based and model-based estimates from probability sampling and estimates from Internet data sources can suffer from an unknown selection bias. Therefore, similar statistical methods are applicable for both data source types.

From a total of 16 quality properties of traditional data sources (Census, Survey, Administrative) and new Internet data sources presented in Table 1.1, the author identifies representativeness and selection bias as the main indicators of non-sampling errors in Internet data sources. Identification of suitable statistical methods is a critical requirement to the analysis of these errors. Because of the non-probabilistic property of Internet data source, the author correctly states that some of the traditional methods of official statistics, such as the Horvitz-Thompson estimator, are not readily applicable and some others (e.g. data cleaning and edition, imputation, calibration) constitute a challenge (Table 1.2).

The author notes that a fruitful avenue is offered by methods developed for survey data contaminated by self-selection. Examples are methods for *self-selection web surveys* and *nonresponse adjustment* methods for probability samples that suffer from missingness. The method of *propensity scores* is often used to adjust for bias caused by self-selection in web

surveys and by non-response in probability samples, and the method is much studied in survey methodology literature (e.g. work by Jelke Bethlehem and others). Important link is recognized with *calibration methods* (e.g. work by Jean-Claude Deville, Carl-Erik Särndal and others) widely used in official statistics production. Current theoretical and methodological developments in *register-based statistics* (e.g. work by Li-Chun Zhang and others) appear useful. The extension by the author of the methodology of Zhang and collaborators provides the "hard core" statistical contribution in the dissertation.

To be effective for reduction of selection bias, it is essential for all these methods that powerful auxiliary (reference) data on the target population (or a probability sample of it), or from other reliable statistical data sources, are available. Administrative and statistical registers and sample surveys are good examples of such data sources. The author is well aware of this when addressing in Chapters 2 and 3 the relevant theoretical and technical considerations necessary in constructing an integrated statistical system for real estate market consisting of micro data (or, aggregated data) from registers, sample surveys and Internet data sources.

In Chapter 3 the author develops methodologies to assess representativeness and selection bias in Internet data sources. This task required a thorough review of the relevant existing methodologies and an assessment of their adaptation for Internet data sources. Topics considered include an excursion to Rubin's theory on missingness types and the use of the typology and modelling methods for accounting for informative self-selection of estimates, a brief discussion on imputation techniques and calibration methods, and an introduction to estimators that are aimed to reduce estimation bias. It is inferred that for non-probabilistic data, such as Internet data source, a missing not at random (MNAR) type missingness structure can be expected, meaning that the phenomenon of interest correlates with the selection mechanism causing selection bias. The author concludes that in order to estimate response propensities or apply weighting (calibration) procedures, it is necessary to examine the underlying selection mechanism and this mechanism can be detected by measuring representativeness.

After first defining and discussing representativeness in the context of Internet data sources, the author turns to present his two-step procedure to measure representativeness of generic Internet data source. Flow diagram 3.1 presents a very helpful summary of the procedure. Potential statistical methods are discussed and evaluated carefully for their ability to assess representativeness. Scope of methods is large including for example record linking and data integration techniques, R-indicators, capture-recapture methods, propensity scores method, various test statistics, and certain econometric methods. The methods differ mainly with respect to data requirements, underlying assumptions and technical complexity.

The availability and type of data from Internet and reference data sources appears critical for successful application of the two-step procedure. Obviously, access to micro data offers the best possibilities for assessing representativeness and selection mechanism. This option is possible in advanced data infrastructures. However, a mixture of micro data and aggregate (domain) data is often the main option under the actual data infrastructure. Careful analysis of properties of available data constitutes the first task and this requires a lot of time and effort. It can be concluded that the two-step procedure can offer a comprehensive framework to assess representativeness in a generic Internet data source, giving support to hypothesis (1).

Empirical assessment of representativeness of the selected real Internet data sources is carried out in Chapter 4 by applying the two-step procedure. Even if the situation is not straightforward, the author is able to confirm the main hypothesis, that is, the Internet is useful for providing statistics in the case considered. Then, the author addresses the measurement of representativeness of the selected Internet data sources by using sample survey data (available at domain level) and register data (available at transaction level) as reference data sets. The properties of data sources were carefully investigated. Bias and *absolute relative bias* (ARB) were used as empirical measures of representativeness. Bias was defined as the difference of category-specific proportion estimates from Internet data source and reference data source, for given domain and time period. ARB is useful because it measures the relative magnitude of overall bias. The design consists of three Internet data sources, 12 domains (cities) and four time quarters.

The first analysis indicated systematic selection bias in the three Internet data sources, supporting hypothesis (2). Next, the goal was to assess whether the bias is of MNAR type. Due to limitations in access of Internet data source and reliable register-based reference data at the micro level, the analysis was restricted to a single Internet data source and a single domain. Comprehensive analysis, which is illustrated by extensive graphical displays, indicated that the response model for the Internet data source for Poznań is of MNAR type. The results give partial support to hypotheses (3). The conclusion leads to the next phase of the empirical analysis: the estimation of selection bias in order to examine if the MNAR property is also present for the other cities and Internet data sources. This task is the target of the final chapter.

In Chapter 5, the bias of estimates was estimated for 12 cities and three selected Internet data sources. From statistical inference perspective, this chapter is clearly the most advanced. The aim of the author was to adapt the method of Fosen and Zhang (2011) and Zhang (2012) for bias estimation in a selected Internet data source. These authors used a small area estimation approach called Fay-Herriot area level mixed model. The model included a single random effect in addition to a fixed effect and the residual term. Also in the dissertation, area-level linear mixed modeling was taken as the modeling approach. However, the setting is more complex involving several Internet data sources and a temporal dimension and thus, several random effects were needed. In the analysis design, the study variable is offer price per $m^2$ and the statistic is the average price per $m^2$. The target was to estimate the bias of the average price per $m^2$. ARB was defined in a similar way as in Chapter 4. Three Internet data sources were selected covering 12 domains (cities) and 12 time periods. Thus, the analysis data set comprises of 432 records. The goal was to model and estimate the covariance structure of the data and make statistical inference on the structure.

After introducing technical notation for bias, mean squared error and coefficient of variation, a linear mixed model was derived for the estimation of bias. In addition to the fixed (intercept) term, random (intercept) effects were defined for different hierarchical levels of the data: spatial i.e. domains (cities), temporal (AR1 for domains), and clusters of real estates (Internet data sources). A random interaction effect (domain and Internet data source) was also included. A small number of Internet data sources might provide a potential technical problem in the mixed model. However, the corresponding random component is important if generalizing results beyond the sampled Internet data sources. The model is well justified and can be as-

sessed suitable for the analysis problem. The modelling setting involves complex covariance structures and several assumptions, approximations and reservations underlying the model were necessary in order to make the analysis setting manageable and computable. A comprehensive model diagnostics was executed to assess the model quality and the importance of random effects in the model.

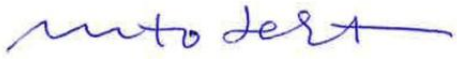
Absolute relative bias appears to be a reasonable measure of error. An assessment of the accuracy of bias and ARB estimates would be more demanding because an approximation to the MSE of estimates must be first calculated. Variance estimates of bias estimates are reported in detail (Appendix tables). Accuracy estimation is left for further research. Empirical results give support to hypotheses (3) and (4).

The author discusses carefully the limitations of the model and presents topics for further research. Additional limitations of the Fay-Herriot model are presented in literature (e.g. Guadarrama, Molina and Rao 2016, Statistics in Transition). For example, the model assumes known sampling variances. In practice, they are rarely known and their estimation is not straightforward (here, a "gold standard" was used). The EBLUP approach for Fay-Herriot model is essentially model-based and therefore, the estimates may suffer from design bias. This could be studied for example with simulation experiments. Additional possible directions for future research are presented in the summary section.

As a summary, the candidate has successfully introduced a relevant scientific problem and presented an original solution to it. He has shown excellent knowledge on theoretical and methodological aspects of statistical methods in the field. He has conducted empirical analysis innovatively and with creativity. His work indicates very good mathematical and technical skills and statistical imagination. The PhD candidate has demonstrated ability to carry out scientific research independently.

My conclusion is that the submitted dissertation meets the requirements of a doctoral thesis, as stated in the Act on Academic Degrees and Titles and on Degrees and Title in Art. Therefore I request that the Faculty Council should admit the PhD candidate to a public defence of the thesis. Because of the high level of the doctoral dissertation, I recommend that the dissertation will be awarded an appropriate prize from the Faculty Council.

Helsinki, 13 May 2016



Risto Lehtonen
Professor
University of Helsinki