Internet data sources for real estate market statistics

An abstract of the doctoral thesis

Maciej Beręsewicz, M.Sc.
Department of Statistics
Poznan University of Economics and Business

Poznań 2016

Information in the system of official statistics is collected through statistical surveys, both sample-based and full enumeration surveys; specific stages of such surveys are predetermined and well described in the literature (cf. Särndal et al., 2003; Bethlehem, 2009; Lohr, 2009; Groves et al., 2011). One classic example of a statistical data source are censuses: they are expensive and are usually conducted once a decade; the resulting information is often published with a considerable delay with respect to the date of data collection. In contrast, sample surveys are designed to provide estimates of selected characteristics of the target population based on information obtained only from a sampled fraction of the population. The scope of sample surveys is wider than that of censuses and covers topics such as economic activity of the population, life quality or household spending.

However, given budget cuts in statistical programmes and a growing fraction of refusals to participate in sample surveys, statisticians are increasingly becoming aware of the need to exploit other available data sources, including administrative registers, Internet data sources (IDSs) or, more generally, "big data". In the literature this approach is referred to as a paradigm shift in official statistics, marked by a decreasing reliance on traditional data sources in favour of using and linking available sources.

Administrative registers have been created by laws and regulations without consulting statisticians (Wallgren and Wallgren, 2014). Wide use of registers since 1980s in Scandinavian countries (e.g. Finland, Norway, Denmark) and in the Netherlands has testified to enormous possibilities these data sources hold to meet the needs of official statistics. During the Polish National Census of Population and Housing 2011 (NSP 2011), in addition to data from a 20% sample of the population, a number of administrative registers were used. One should bear in mind that administrative registers are considerably different from censuses and sample surveys, mainly

in that they have been created for purposes other than statistical. Besides, registers contain a scope of information that has been specifically defined to meet the needs of public administration. For example, The Register of Real estate Prices and Values contains information about apartments offered for sale. For this reason statisticians have turned to the Internet and "big data", which can provide more complete information in combination with existing data sources.

Internet and "big data" sources are described in the literature devoted to information science, information technology, e-commerce or sociology (Abramowicz et al., 2002; Miller, 2011; Lazer et al., 2014). Few statistical studies deal with methodological problems related to the use of IDSs. Even if these problems are actually addressed, however, IDSs are not treated comprehensively as statistical data sources, or as a special kind of information source. Statistical information should be characterized, among other things, by relevance, precision (of estimation), timeliness, accessibility and transparency, comparability (in time and space) and coherence. One particularly important criterion of statistical data sources is their representativeness, which enables estimation of target population characteristics. Consequently, before IDSs (e.g. web portals) and administrative registers can be used for statistical purposes, they must be evaluated to identify opportunities and threats involved in using them, particularly in terms of the theory of estimation.

One should also be aware of definitional problems concerning IDSs or big data in the context of existing sources of statistical information. Hence, the following definition has been adopted in the dissertation: an Internet data source is a self-selected (non-probabilistic) sample that is created through the Internet and maintained by entities external to NSIs and administrative regulations. Additional, the author has limited the definition of the Internet to the world wide web.

The dissertation is a methodological and empirical study. The proposed approaches have been applied to data from the secondary real estate market, which is only narrowly surveyed by official statistics. The thesis is also based on the assumption that the secondary real estate market is a hard-to-reach population, for which there is no available sampling frame and sampling units are hard to identify.

The main goal of the dissertation is to evaluate the Internet as a data source for real estate market statistics. To achieve this goal the author has set the following specific goals:

G1 to identify sources of non-sampling errors in Internet data sources in the secondary real estate market;

G2 to assess the representativeness of Internet data sources about the secondary real estate market.

G3 to assess bias in Internet data sources about the secondary real estate market.

To achieve these goals, the author has put forward the following research hypotheses. The main hypothesis to be verified in the dissertation is: Internet data sources enable acceptable estimation of real estate market characteristics. The following specific hypotheses have been formulated:

H1 An original approach proposed in the dissertation can be used to assess representativeness of Internet data sources in the secondary real estate market.

H2 Internet data sources contain systematic bias, which varies across sources and domains.

H3 Self-selection error in Internet data sources in the real estate market is informative (i.e. depends on the target variable y).

H4 Internet data sources enable estimation of the offer price in the secondary real estate market with acceptable precision measured by absolute relative bias.

The selection of data sources, domains and the reference period was dictated by data availability. The empirical study was limited to 12 cities (Białystok, Gdańsk, Katowice, Kraków, Łódź, Lublin, Olsztyn, Opole, Poznań, Szczecin, Warszawa, Wrocław) and the period from Q1 of 2012 to Q4 of 2014. The domain of interest for the estimation of the bias of the mean offer price per m2 was the city. Three web portals were chosen as data sources: Dom.Gratka.pl, OtoDom.pl and Nieruchomości-Online.pl. The Central Statistical Office and the National Bank of Poland conduct a survey of residential and commercial property prices, which collects information about the secondary real estate market. Results of the NBP/CSO survey were used as reference statistics for comparison with IDS-based estimates. Another source of information used in the evaluation of the representativeness of IDSs was the Information Society survey conducted by the CSO. In addition, the author compared the results with real estate data for Poznań, obtained from the Register of Real Estate Prices and Values.

The dissertation presents an extension of a small area estimator based on a linear mixed model proposed by Fosen and Zhang (2011) and Zhang (2012a). The proposed model was adapted to estimate and decompose the bias of the mean offer price per m2 estimated from IDS data from 12 Polish cities. The model takes into account four types of random effects: the effect of correlated data sources, the domain (city) effect, the effect of interactions between the domain and the data source and the effect of autocorrelation of bias in time.

The first chapter is devoted to the literature review and a summary of recent articles on new data sources for official statistics. It is accompanied by a historical overview of the use of statistical and non-statistical data sources. The author then goes on to propose a definition

of Internet data sources. The chapter ends with a summary of challenges and risks related to the use of IDSs for official statistics.

The second chapter focuses on IDSs from the point of view of statistical data sources. The author notes the ambiguity involved in applying the definitions of population and unit with respect to IDSs and the real estate market. He compares characteristics of IDSs with censuses, sample surveys and administrative registers. This is followed by a description of selected data sources about the real estate market in Poland. The author puts forward an original approach to the integration of statistical data sources, which takes into account IDSs. The chapter ends with a taxonomy of errors that can be found in IDSs developed on the basis of Zhang (2012b), and illustrated with examples from the real estate market.

The third chapter lays out the underlying theory for the measurement of representativeness of IDSs. Possible sources of bias in IDSs are presented, which are related to the non-random character. The author presents the notion of representativeness according to Kruskal and Mosteller (1979a,b,c) and applies it to IDSs. On this basis he proposes an original procedure to measure representativeness, which takes into account the aggregation level of available data (individual and aggregate data) and their source: statistical (survey, census, reporting) and non-statistical (mainly administrative registers). Several possible measures of representativeness are considered, depending on data availability.

The fourth and fifth chapters report results of the empirical study. The fourth chapter describes the application of the two-step procedure of measuring representativeness proposed in the third chapter. The first step consisted in assessing the rationale behind the use of the Internet for study of the real estate market in Poland. This step was based on data from the ICT survey conducted by the CSO. The second step involved the assessment of distributions of floor area and the number of rooms in properties offered for sale in the secondary market in 12 cities in Poland. The results were compared with reference data from the NBP/CSO survey. Another comparison made in the study looked at distributions of prices of apartments offered in the secondary market and transaction prices recorded in the Register of Real Estate Prices and Values.

In the fourth chapter the author identifies differences between IDSs, the survey conducted by NBP and CSO and the Register of Real Estate Prices and Values. Results of this analysis indicate that two groups of properties are not represented online: (1) low-priced properties with small floor area, and (2) expensive properties, with big floor area. The following potential explanations for this situations have been identified: (1) some properties are auctioned off by court bailiffs for a price lower than the market price; (2) properties sold between family members or relatives;

(3) transactions made without the use of advertisement services; (4) brokers do not advertise all their properties for sale online; (5) the cheapest and the most expensive residential properties are presented to a specific, narrow group of buyers. Compared to the NBP / CSO survey, IDS-based estimates exhibit significant differences in the distribution of one-room apartments with floor area of up to 40 m2 and large apartments (over 80 m2) put up for sale in Poznań.

The fifth chapter describes a small area estimator based on a linear mixed model proposed by Fosen and Zhang (2011) and Zhang (2012a), which is extended in order to enable the estimation and decomposition of bias of the mean offer price per m2. The proposed model assumes the existence of a "gold standard", namely unbiased estimates of the variable of interest, usually obtained from sample surveys or censuses. The "gold standard" adopted in this study was the survey conducted by the NBP and CSO. The model takes into account 4 random effects and enable bias detection in IDS-based estimates of the mean offer price per m2.

The inclusion of the random effects for the domain, for the data source and the interaction between them considerably improves model performance. Besides, the results exhibit very strong autocorrelation of bias over time. The biggest difference in the mean offer price per m2 was observed in the least popular data source (Nieruchomosci-Online.pl), whereas the smallest differences were found in the most popular source. The intercept of the model was found to be insignificant, which may indicate the lack of statistically significant differences between IDSs and the NBP / CSO survey. The relative error of IDS-based estimates of the mean offer price per m2 ranges from 0.2% to a little over 4%.

In addition, the estimated model indicates that bias varies depending on the city and is correlated with the mean offer price per m2 in a given city. This implies that the selection mechanism is informative. Bias is higher for properties located in cities where property prices are high (e.g. Kraków, Warszawa and Gdańsk), and is lower for properties located in cities where prices are lower (e.g. Białystok, Katowice and Szczecin).

The dissertation ends with a summary and a discussion of possible directions for further research.

Bibliography (selected)

Abramowicz, W., Kalczynski, P. J., and Wecel, K. (2002). *Filtering the Web to feed data warehouses*. Springer Science & Business Media.

Bethlehem, J. (2009). *Applied survey methods: A statistical perspective*. John Wiley & Sons.

Fosen, J. and Zhang, L.-C. (2011). The approach to quality evaluation of the micro-integrated employment statistics. ESSnet Data Integration.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey methodology*, volume 561. John Wiley & Sons.

Kruskal, W. and Mosteller, F. (1979a). Representative sampling I: Non-scientific literature. *International Statistical Review*, 47(1):13–24.

Kruskal, W. and Mosteller, F. (1979b). Representative sampling II: Scientific literature excluding statistics. *International Statistical Review*, 47(2):111–123.

Kruskal, W. and Mosteller, F. (1979c). Representative sampling III: The current statistical literature. *International Statistical Review*, 47(3):245–265.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, 343(14).

Lohr, S. (2009). *Sampling: design and analysis*. Cengage Learning.

Miller, G. (2011). Social scientists wade into the tweet stream. *Science*, 333(6051):1814–1815.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

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

Zhang, L.-C. (2012a). On the accuracy of register-based census employment statistics. European Conference on Quality in Official Statistics.

Zhang, L.-C. (2012b). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1):41–63.