Maciej Beręsewicz

Estimating the size of the secondary real estate market based on internet data sources

Folia Oeconomica Stetinensia 14(22)/2, 259-269

2014

DE GRUYTER OPEN

Folia Oeconomica Stetinensia
DOI: 10.1515/foli-2015-0012

WYDZIAŁ NAUK EKONOMICZNYCH I ZARZĄDZANIA
UNIWERSYTET SZCZECIŃSKI

ESTIMATING THE SIZE OF THE SECONDARY REAL ESTATE MARKET BASED ON INTERNET DATA SOURCES

Maciej Beręsewicz, MSc.

Poznań University of Economics
Department of Statistics
Al. Niepodległości 10, 61-875 Poznań, Poland
e-mail: Maciej.Beresewicz@ue.poznan.pl

Received 1 July 2014, Accepted 6 November 2014

Abstract

As a result of the growing digitization of society and the development of electronic economy, current statistical data sources, including administrative registers, do not satisfy the information needs of society. Therefore, there are growing gaps in the statistical coverage of a number of sectors of the economy. One example of such a gap is the secondary real estate market, which is only partially accounted for by official statistical data sources. On the other hand new data sources such as the Internet or Big Data tend to decrease information gap in official statistics. The Web portals that specialise in brokerage on real estate market should be not neglected as a data source for statistics. Therefore, the aim of the paper is to use two Web portals devoted to the housing market to estimate supply measured in the number of flats offered to sale in Poznań, Poland. In addition, classification and quality of Web portals will be discussed.

Keywords: secondary real estate market, internet data sources, statistical data sources, Petersen estimator, capture–recapture, supply on real estate market.

JEL classification: C13, C42, C81, L85.

Introduction

At present censuses and surveys are the main sources of statistical data, which provide information about selected characteristics of study populations. Censuses are conducted at regular intervals (usually every 5–10 years). They are costly to implement and the resulting information tends to be delayed with respect to the time when a census is taken. Representative surveys are designed to avoid these disadvantages and to provide information about selected aspects of socio–economic life of a representative sample of the target population. Both sources are broadly discussed in the statistical theory[1].

Recent years have seen a growing interest in administrative sources (e.g. registers), which are considerably different from traditional sources of statistical information[2]. What distinguishes registers from those classic sources is, among others, their purpose related to the execution of respective laws and regulations. It should be noted that the experience of using registers in Scandinavian countries has demonstrated their significant role as sources of statistical information[3]. The Central Statistical Office in Poland has also gained some experience in the use of administrative registers for statistical purposes (e.g. National Official Business Register REGON, registers of employment offices, The Register of Prices and Values of Real Estate). It should also be mentioned that the last census (NSP 2011) made use of registers[4]. For example, the census provided information about buildings and characteristics of dwellings at the level of *districts* (LAU level 1, Polish 'powiat') and selected *communes* (LAU level 2, Polish 'gmina').

Although in the theory of estimation it is assumed that the whole target population is completely accounted for in registers (or sampling frames)[5] – administrative sources provide only a partial description of the socio–economic environment: they describe only the fragment specified by the regulations and the register administrator. In view of the above, there are gaps in the information coverage of some areas of the economy (e.g. the secondary real estate market). For example, *The Register of Prices and Values of Real* only contains information about transactions made on the real estate market, which are published annually at the level of districts and selected communes. An additional problem is the delay in the publication of survey reports. Considering the above shortcomings, statisticians have been looking for new ways of obtaining information, and focusing their attention on Internet data sources, or more broadly described as *Big Data.* The new data sources (NDS) offer for statisticians timeliness information derived from the data that is often described with detailed characteristics. Moreover, IDS allow measuring new concepts and estimating new statistics that are not in the official statistics and which could be useful for the users of the statistical system. Finally, costs of data collection in

relation to census or survey is low and indicates that new data sources could be considered as a competition to the existing sources in official statistics. For the purpose of this article they will be referred to as new information sources (NIS), while in the context of the real estate market the term 'internet data sources' (IDS) will be used.

1. Internet data sources in the description of the secondary real estate market

Although new information sources (NIS) are mentioned in the literature devoted to information technologies, e-commerce, finance and sociology[6], no comprehensive studies exist about real estate markets[7]. However, they only focus on selected aspects of using NIS, without providing a comprehensive analysis of their usefulness as sources of statistical data, which should be treated as a specific example of an information source. It is generally assumed that statistical data sources should be *representative* and should provide unbiased *estimates* of population characteristics. Consequently, before IDS (e.g. web portals) can be used for estimation purposes just like administrative registers, they must be evaluated to identify opportunities and threats related to their potential use, in particular with respect to the theory of estimation.

The Register of Prices and Market Value of Real Estate (RPMVRE) mentioned above was created on the basis of clause 74 of [the regulation of 29 Mach 2001, no. 38, item 451] and it is administered by district governor's offices. Data stored in the RPMVRE are provided by notaries, who are obliged to report information about transactions made on the primary and secondary real estate market. This means that information stored in the register represents only a fragment of the market (transactions) and does not include the historical record of a given property (e.g. the date of placing an ad, the asking price). This limits the possibility of estimating characteristics of the secondary real estate market, e.g. its size, time-to-sale or the percentage of properties sold (transactions/property ads). Additionally, access to the register is legally restricted, and only real estate appraisers can access the information free of charge[8].

The National Bank of Poland and the Central Statistical Office make use of the RPMVRE in the following surveys devoted to the secondary real estate market: (*i*) the management of housing resources (No. 1.26.01(074)), (*ii*) Property sales (No. 1.26.04(075)), (*iii*) A survey of residential and commercial property prices (No. 1.26.09(078)), additionally relying on questionnaire surveys conducted among real estate agents and information from web portals[9]. However, in January 2014 a number of jobs underwent deregulation, including the job of an estate agent. It is an important development in the context of statistical information, because it affects the possibility of identifying a complete list of all estate agents (and consequently the sampling frame)[10].

The use of IDS calls for the development of methodological principles, which would make it possible to select and evaluate web portals as sources of statistical data. There are two interesting examples of web portals operating on the real estate market: *Funda.nl* and *Nieruchomista.pl*. The first one was created by an association of estate agents in Holland (*Nederlandse Vereniging van Makelaars*, NVM), which is used, inter alia, by the Statistics Netherlands to obtain information about property-for-sale offers in order to link (e.g. via Record Linkage method) them to the register that contain information on transactions[11]. The second example is a Polish portal, offering listings of properties for sale and to rent placed by members of the Poznan Consortium of Estate Agencies REAL NET Ltd. under an exclusivity agreement. However, it should be mentioned that on the market there are web portals (e.g. otoDom.pl, dom.gratka.pl or domiporta.pl), which allows to place property-for-sale for wider group of market participants.

Figure 1 shows the number of estate agents registered at otoDom.pl in Poznań by date of placing and updating a property ad and the number of agencies registered in the REGON register in subclass *68.31.Z – Real estate brokerage* in the period 2009–2013. Data from the OtoDom portal were collected automatically between 2011 and 2013[12]. There is a considerable difference between the number of agents registered in REGON and those registered on OtoDom,

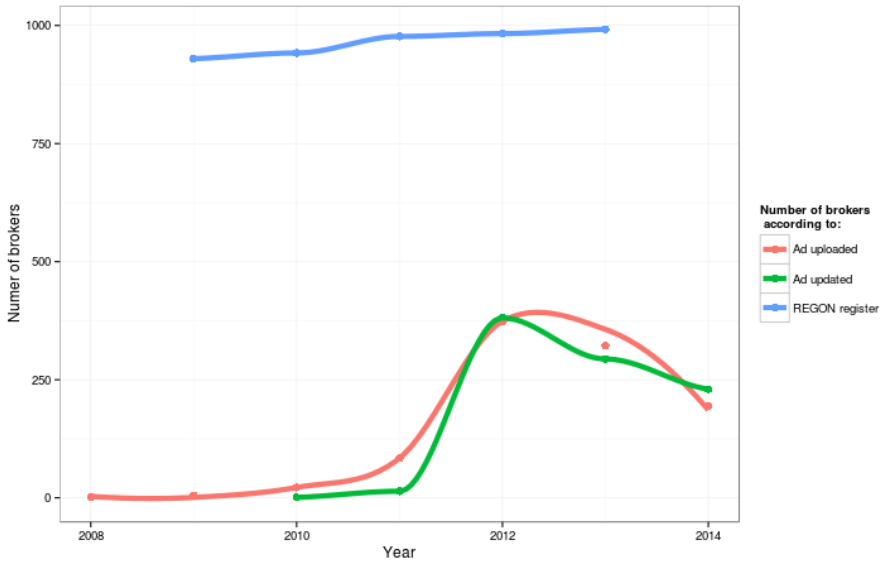

Fig. 1. The number of agents in Poznań according to the REGON register and on OtoDom portal

Source: own presentation based on data from OtoDom portal.

which may be due to a few causes. For one thing, the subclass includes businesses not only offering property brokerage services but also consulting and appraisal services. Secondly, the subclass does not contain information about the market segment that a given agency specializes in (houses, flats, etc.) Finally, agents registered in Poznań may broker deals involving property located only in Poznań, in the Poznań district or both. It is worth noting that the data indicate a decreasing number of agents.

In view of the differences mentioned above one needs to identify some distinguishing characteristics of web portals. From the point of view of IDS–based studies, web portals can be classified into the following types, which differ not only in terms of data quality, and the business model but also in the number of property ads[13]: (1) *brokering portals* (e.g. HomeBroker.pl, Metrohouse.pl), (2) *portals of real estate associations* (e.g. REAL NET – nieruchomista.pl, PFRN[14] – fagora.pl), (3) *portals offering brokering assistance* (e.g. Allegro – OtoDom.pl, Agora – domiporta.pl), (4) *portals aggregating other portals* (e.g. dom.money.pl).

The choice of portals may significantly affect the information coverage of the secondary estate market, since estate agents, as well as private individuals, may prefer one (e.g. their own) or more portals or may not place ads online. Unfortunately, Polish literature on the subject does not contain studies devoted to estate agents or web portals they use to contact their clients (e.g. what percentage of ads are placed online). Moreover, estate agents may have their subpages on portals offering assistance in the sale of property. For instance, *MetroHouse* carries[15] 498, dom.gratka.pl 453, and OtoDom 327[16] ads of property for sale on the secondary market in Poznań. The existence of preferred portals (e.g. owing to the effectiveness of promotion) is associated with what is termed in statistics as selectivity. Since selectivity can result in biased estimates, the choice of portals is a crucial issue. One example of a real estate market which is not published by official statistics is its size. For this reason, an attempt was made to estimate the size (supply) of the secondary property market (flats) in Poznań at 1 September 2013 based on data from two portals – OtoDom and Gratka.

2. Estimating the secondary property market in Poznań

The problem of estimating the population size is mainly known from environmental studies, in particular studies aimed at estimating the number of species[17]. However, it soon became evident the methodology applied in those studies could be used in social research, for example in the study of *hard-to-reach populations*), such as immigrants, homeless or illegal workers[18]. Statisticians have noticed that similar problems occur when trying to identify the

scope of coverage for censuses and administrative registers. In this way the approach known from environmental studies has been adapted to serve the needs of public statistics, for example to determine the quality of sources (*under-, overcoverage*) or to estimate using two or more sources (*dual system estimator*). A detailed list can be found in[19].

The literature describes two approaches to estimating population size – one based on Petersen estimator and the second one based on log-linear models. This article focuses on the first approach, which is based on the following model – it is assumed that there are two lists (registers, sampling frames, surveys) of the same population (*capture–recapture).*

Table 1. Probabilities of units occurring in lists A and B

		List B		
		included in the list	not included in the list	
List A	included in the list	p_{i11}	p_{i12}	p_{i1+}
	not included in the list	p_{i21}	p_{i22}	p_{i2+}
		p_{i+1}	p_{i+2}	1

Source: own presentation based on Wolter (1986).

Let A denote the first list and B the second list. Let p denote the probability of the occurrence of events described in Table 1, where p_{ijk} denotes the probability of events j (being included or not included in list A) and k (being included or not included in list B) for i-th unit. After drawing a sample, we obtain Table 2, where n denote the number of events j and k.

Table 2. The sample size by unit count in lists A and B

		List B		
		included in the list	not included in the list	
List A	included in the list	n_{11}	n_{12}	n_{1+}
	not included in the list	n_{21}	n_{22}	n_{2+}
		n_{+1}	n_{+2}	$n_{++} = N$

Source: own presentation based on Wolter (1986).

The model is based on the following assumptions[20]: (1) *the population is closed* – of constant size N; there has been no change in population size between the first and the second sample; (2) *Probabilities determined in the table (z p) are parameters of a multinomial distribution*; (3) *Autonomous independence* – lists A and B have been created independently as independent samples from the same population; (4) *no errors in unit linkage* – i.e. units from lists A and B are linked unequivocally; (5) *no undesirable events* – errors in both lists have been removed before estimation; they include duplicates, non-existent units or units from outside the

population; (6) *complete information to identify a unit*; (7) *The probability of occurring in both lists is the same for all units* – this means that $p_{i1+} = p_{1+}$ and $p_{i+1} = p_{+1}$ for every $i \in 1, ..., N$. Under these assumptions, the likelihood function for the model is given by:

$$L_t\left(N, p_{1+}, p_{+1}\right) = \binom{N}{n_{11} n_{12} n_{21}} p_{1+}^{n_1} p_{+1}^{n+1} \times \left(1 - p_{1+}\right)^{N - n_{1+}} \left(1 - p_{+1}\right)^{N - n_{+1}} \quad [1]$$

Petersen's estimator, which maximises the above likelihood function, can be written as:

$$\hat{N} = \frac{n_{1+} n_{+1}}{n_{11}} \quad [2]$$

where n_{1+} denotes the size of list A, n_{+1} denotes the size of list B, a n_{11} the number of common units.

In the case of Internet data sources and the specific character of the secondary property market, assumptions (1), (4), (5) and (7) may not be satisfied. Assumption (1) implies that the size of the market is constant, which means that there are no transactions or no adverts are removed. According to assumption (4), the same units from the two lists can be unequivocally matched, which may not always be possible in the case of adverts containing incomplete information (e.g. location). In the case of IDS on the secondary property market assumption (5) is not satisfied since the same offers may be placed multiple times by the same or different agencies (e.g under an open-ended agreement). The last assumption (7) is not satisfied either, since, in practice, the choice of a particular portal by an agency may be motivated by e.g. a higher chance of reaching buyers. In the literature one can find approaches in which some of these assumptions are omitted[21]. In this article it is assumed that the above assumptions concerning the use of Petersen estimator are satisfied. This simplified approach is also used to estimate the scope of coverage by a census or to study population that are hard to reach[22].

Information for the study was obtained using a software application (a web crawler) specially written for this purpose, which collected selected data from two portals – OtoDom and Gratka. This included information about flats for sale (excluding houses) where adverts were updated within 14 days of 1st September 2013. All the computations were performed using the *R* statistical software[23]. The stage of preparing data involved the following steps: (1) key characteristics of sampled units were identified in text columns based on regular expressions; (2) missing data were imputed using information contained in text descriptions; (3) adverts which did not meet the study requirements (e.g. property for lease, houses) were removed

(4) uniform names of agents were adopted to ensure consistency between the two portals; (5) names of streets verified against the database of STREETS in the TERYT register using the Levenshtein distance to measure the similarity between texts; (6) the last step involved adopting uniform levels and names of variables; (7) Units were identified by means of *probabilistic record linkage*[24]. The threshold was set at 0.8, which corresponded to 80% probability that a pair of adverts refers to the same unit. As a result of the preparation process described above, 2,532 adverts from the Gratka portal (originally 2,780) and 2187 from OtoDom (originally 2,425).

The next stage involved estimating the size of the secondary real estate market in Poznań using Petersen estimator, and its variance by means of parametric bootstrap, which was based on assumption (2) about the multinomial distribution of probabilities in Table 1. The procedure of estimating variance consisted of the following steps[25]: (i) agents were sampled independently from the two portals under Poisson sampling[26]; (ii) Empirical probabilities for a given sample were calculated; (iii) n_{+1}, n_{+1} and n_{11} were generated from the multinomial distribution given the (ii) probabilities; (iv) The population size $\hat{N}$ was estimated.

Table 4. Descriptive statistics of bootstrap estimates of the size of the secondary property market

Min	1st Quartile	Median	Mean	3rd Quartile	Max
2,057	2,779	2,951	2,914	3,086	3,383

Source: own tabulation.

Figure 2 shows the distribution of estimates of the size of the secondary property market obtained by means of parametric bootstrap. It can be seen that the distribution of the population size is negatively skewed and most estimated are concentrated around the value of 3,000 units.

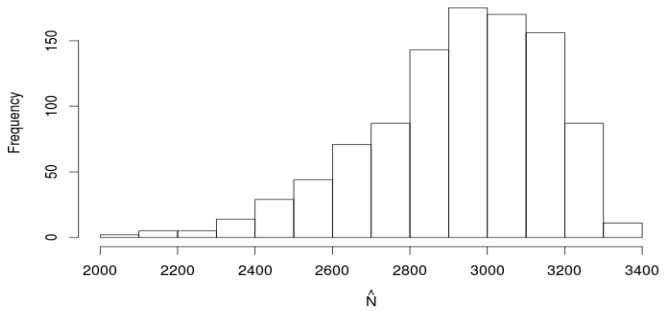

Fig. 2. The distribution of bootstrap estimates of the size of the flat market in Poznań

Source: own presentation.

Table 3 contains basic descriptive statistics. It can be seen that both portals underestimate the population size by 500–1000 flats.

Conclusions

The use of Internet data sources to produce statistical information has not been fully analysed yet and requires a comprehensive evaluation in the context of estimation theory. The secondary property market in Poland is only minimally accounted for by Official Statistics, which makes it difficult to estimate its characteristics, such as market size (supply). On the other hand, web portals provide comprehensive information about processes taking place on this market. The study presented in this article was an attempt to estimate the size of the secondary property market based on data from two portals – Gratka and OtoDom. The results show that the 95% CI of number of flats offered to sale on 1st September 2014 is (2,456; 3,385) with mean 2919. The results obtained by means of Petersen estimator indicate that discussed web portals do not cover the whole market and underestimate it by 500–1000 flats. Not taking into account overlap between two data sources or using only one of them leads to bias in the estimation of market size and therefore of statistics that is based on this characteristic (e.g. the fraction of sold flats).

Acknowledgements

This research was a part of the project *The use of Internet data sources in the context of the estimation theory exemplified by the secondary real estate market in Poland.* The project was financed by the National Science Center grant awarded on the basis of decision number 2014/13/N/HS4/02999.

Notes

[1] For example Groves et al. (2010).
[2] Cf. Paradysz (2007); Wallgren, Wallgren (2014).
[3] Cf. Statistics Finland (2004); UNECE (2007).
[4] Cf. Roszka (2012).
[5] This assumption is not always satisfied (cf. Gołata, Dehnel (2013); Zhang (2015)).
[6] For example Lazer et al. (2014); Miller (2011).

[7] Beręsewicz (2014); Beręsewicz, Klimanek (2013); Daas et al. (2011).

[8] Status as at 21.06.2014. The regulation of the Minister of Infrastructure of 19 February 2004.

[9] Unfortunately, the methodological description does not specify to what extent IDS are used in the study. NBP Study coordinators from NBP in Poznań have mentioned www.gratka.pl and www.otoDom.pl as the primary sources of information, which are then supplemented with data from questionnaire surveys conducted among real estate agents.

[10] There are new trade associations of estate agents (e.g. The Polish Federation of Real Estate Market), which are creating their own systems of certification, and, consequently, databases of agents.

[11] Hoekstra et al. (2010).

[12] Data from Gratka.pl was collected at the end of 2012 and in 2013, which made it impossible to compare the information.

[13] For example, the number of ads offering residential property for sale on the secondary Real estate market on 21.06.2014 in Poland according to OtoDom, Dom. Gratka, Domiporta and Szybko.pl amounted to: 342,109; 414,425; 314,413 and 153,476 respectively, and for Poznan 10,572; 13,410; 6,790 and 6,351 respectively.

[14] Polish Federation of the Real Estate Market.

[15] Status at 21.06.2014.

[16] In the case of dom.gratka.pl and szybko.pl these are approximate figures, obtained via a questionnaire. The actual number of ads may differ from that one that was declared, e.g. as a result of duplication.

[17] Cf. IWGDMF (1995).

[18] Cf. Lavalleé, Rivest (2012).

[19] Wolter (1986); Zhang (2015).

[20] Wolter (1986).

[21] Zhang (2015).

[22] Cf. Lavelleé, Rivest (2012); Wolter (1986).

[23] R Core Team (2014).

[24] A detailed description of the method can be found in Fellegi, Sunter (1969).

[25] The number of bootstrap replications was set at 999.

[26] The number of agents on OtoDom and Gratka amounted to 103 and 118 respectively.

References

Beręsewicz, M. (2014). *Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena*. *Taksonomia: klasyfikacja i analiza danych – teoria i zastosowania. Taksonomia 22: klasyfikacja i analiza danych – teoria i zastosowania*. Wrocław: Uniwersytet Ekonomiczny we Wrocławiu.

Beręsewicz, M. & Klimanek, T. (2013). *Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w badaniu cen mieszkań*. *Taksonomia 21: klasyfikacja i analiza danych – teoria i zastosowania*. Wrocław: Uniwersytet Ekonomiczny we Wrocławiu.

Daas, P., Roos, M., de Blois, C., Hoekstra, R., ten Bosch, O., & Ma, Y. (2011). *New data sources for statistics: experiences at Statistics Netherlands*. The Hague/Herleen: Statistics Netherlands.

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

Gołata, E. & Dehnel, G. (2013). *Rozbieżności szacunków NSP 2011 i BAEL. Taksonomia 20: klasyfikacja i analiza danych – teoria i zastosowania* (pp. 120–130). Wrocław: Uniwersytet Ekonomiczny we Wrocławiu.

Groves, R., Fowler, M.F.J. Jr., Couper, M., Lepkowski, J.M., Singer, E. & Tourrangeau, R. (2010). *Survey methodology*. New York: Wiley.

Hoekstra, R., ten Bosch, O. & Harteveld, F. (2010). *Automated Data Collection from Web Sources for Official Statistics: First Experiences*. Heerlen. The Netherlands: Statistics Netherlands.

IWGDMF (1995), International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation II: Applications. *American Journal of Epidemiology*, 142, 1059–1068.

Lavallee, P. & Rivest, L.-P. (2012). Capture-Recapture Sampling and Indirect Sampling. *Journal of Offcial Statistics*, 28, 1.1–27.

Lazer., D., Kennedy, R., King, G. & Vespignani, A. (2014). The parable of Google Flu: traps in Big Data analysis. *Science*, 14 March 2014.

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

Paradysz, J. (2007). *Rejestry administracyjne jako źródło zasilania w statystyce regionalnej*. In: *Statystyka regionalna w jednoczącej się Europie*, ed. J. Paradysz. Poznań: Uniwersytet Ekonomiczny w Poznaniu.

R Core Team (2014). *R: A language and environment for statistical computing* [computer software]. R Foundation for Statistical Computing. Vienna. Austria, www.R-project.org.

Roszka, W. (2012). System statystyki publicznej oparty na zintegrowanych źródłach danych. *Przegląd Statystyczny*, 59, 2.

Rozporządzenie z dnia 29 marca 2001 r. Ministra Rozwoju Regionalnego i Budownictwa w sprawie ewidencji gruntów i budynków (DzU 2001.38.454).

Statistics Finland (2004). Use of registers and administrative data sources for statistical purposes – best practices in Statistics Finland. *Handbook* 45. Helsinki.

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

Wallgren, A. & Wallgren, B. (2014). *Register-Based Statistics: Statistical Methods for Administrative Data*. Chichester: John Wiley & Sons.

Wolter, K.M. (1986). Models for Census Data Some Coverage Error. *Journal of the American Statistical Association*, 81 (394), 338–346.

Zhang, L-C. (2015). On modelling register coverage errors. *Journal of Official Statistics* (forthcoming).