Austrian Journal of Statistics
April 2015, Volume 44, 45–57.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i2.79

On the Representativeness of Internet Data Sources for the Real Estate Market in Poland

Maciej Beręsewicz
Poznan University of Economics

Abstract

Shifting paradigms in Official Statistics lead to the widespread use of administrative records in an effort to support or create an alternative for censuses and surveys. At the same time the demand for diversified detailed information is increasing. In order to meet this demand Official Statistics needs to seek new data sources. Internet data sources (IDS), or more generally, big data could be one of them. The potential usefulness of these new sources of statistical information should not be neglected.

The aim of the paper is to report on a study intended to assess the representativeness of IDS for the real estate market in Poland. These sources could be used for describing the demand and supply on the secondary real estate market in a more detailed way than is possible with the existing methodology. The degree of representativeness is assessed on the basis of information from official surveys and other data sources. Due to the shortage of relevant literature on the subject, the article provides a definition of IDS and draws on insights from a study conducted by the author to enhance information from Official Statistics. The study involved using information on street names from the National Official Register of the Territorial Division of the Country (TERYT) to harmonize street names obtained from IDS. A special program for automated data collection (*web spider*) was developed. All the calculations were made with R (R Core Team 2014) statistical software and additional R packages (**XML**, **RCurl**, **httr** and **ggplot2**).

Keywords: Big data, Internet data sources, secondary real estate market, web scraping, R.

1. Introduction

Increasing information needs at a low level of aggregation not only encourage the development of small area estimation but also stimulate the search for new data sources that could support or enhance existing sources (reporting, censuses or surveys). This process has been continuing since 1970s when statisticians and National Statistical Institutes (NSIs) started using and adopting administrative records into their statistical systems (Wallgren and Wallgren 2014). However, the statistical theory underlying the use of administrative registers is currently the subject of research and development (Zhang 2011, 2012). Nonetheless, the process has brought about a change in thinking about statistical data sources. In the literature and during statistical conferences this process is often described as a change of paradigm in Official Statistics, which involves the adoption of existing data sources instead of creating new ones.

Although administrative records provide unit-level data, their scope is usually limited to a specific field that was crucial to the register's administrator. Initially, registers were not created for statistical purposes, which means that these sources need to be transformed to become a statistical data source. In addition, it is assumed that registers cover the whole target population, which is not always the case (see Gołata 2014; Zhang 2015). However, in the environment of electronic economy, characterized by the increasing use of the Internet (both by households and companies) and the Internet of Things (e.g. mobile technologies), administrative registers as well as surveys tend to lag behind the changing setting. Therefore, information gaps in certain fields are growing and new data sources should be examined to improve information coverage.

In this context the term *big data* has gained wide recognition as a potential source of statistical information, although it does not have a clear definition. In an information system, it refers to data that cannot easily be handled within the existing infrastructure. From the statistical point of view, it is considered as a potential source for describing ongoing changes in society. The following sources are discussed in the context of Official Statistics: mobile networks (e.g. to track movement, travel routes), social networking sites (e.g. Facebook, Twitter, Linkedin), e-commerce (e.g. eBay, Amazon, price comparison services) or Google search trends. However, they are not being investigated widely as a statistical data source or from the point of view of estimation theory. The purpose of this paper is to bridge this gap by discussing the representativeness issue in the context of new data sources, specifically concentrating on IDS.

The paper has the following structure. The second section defines and presents IDS in the context of survey methodology. The key concept – representativeness – is defined and discussed in the context of new data sources in the Internet. Relation to the characteristics of big data is underlined in the light of statistical data sources. The third section is devoted to data sources for the real estate market in Poland and possibilities of using IDS to obtain statistical information. The penultimate section contains an empirical evaluation of representativeness using data from the Polish real estate web portal – `http://www.nieruchomosci-online.pl`. The data were collected from the web portal automatically by means of a special R (R Core Team 2014) program. The article ends with the discussion of results and final remarks.

2. Internet data sources

While Internet access in households is increasing (Mohorko, Leeuw, and Hox 2013), the way people or companies communicate is changing as well (e.g. Customer to Customer C2C, Business to Customer B2C, Business to Business B2B). This process opens new opportunities for statisticians to track and measure economy and society. For example, it is possible to use web services to assess auctions (e.g. e-Bay), compare prices on the Internet with off-line prices using e-commerce services (e.g. Amazon) or access hard-to-reach populations. In the literature we can find evidence of using data scraped from web pages to measure inflation, predict unemployment or flu risk. For instance, *the Billion Price Project*[1] conducted by the Massachusetts Institute of Technology (MIT) web-scrape data from over 60 countries and calculates price indexes and measures of macroeconomic phenomena (Cavallo 2012, 2013). However, the exact methodology and web scraping technique is protected by PriceStats[2], a start-up based in Cambridge MA. Another well known project that has highlighted the potential usefulness of the Internet is Google Flu Trends, which was widely discussed a few years ago (Ginsberg, Mohebbi, Patel, Brammer, Smolinski, and Brilliant 2008).

Nonetheless, new data sources have not been discussed widely in the statistical literature. The first reference to the statistical aspect known to the author is mentioned in Shmueli, Jank, and Bapna (2005) and is devoted to on-line auction research. Bapna, Goes, Gopal, and Marsden (2006) discusses the problem of data-driven research in e-commerce studies. However, in

[1] `http://bpp.mit.edu`
[2] `http://www.pricestats.com`

recent years new research addressing big data and IDS in the context of statistical data source has been growing. Below are the main topics and selected literature:

- Predicting unemployment - Fondeur and Karamé (2013), Xu, Li, Cheng, and Zheng (2012);
- Source of information for small area estimation - Pratesi, Pedreschi, Giannotti, Marchetti, Salvati, and Maggino (2013), Pratesi, Giannotti, Giusti, Marchetti, Pedreschi, and Salvati (2014), Porter, Holan, Wikle, and Cressie (2013)
- Opinions / Sentiment analysis - Daas, Roos, van de Ven, and Neroni (2012); Daas and Puts (2014b), Miller (2011)
- Indexes - Vosen and Schmidt (2011)
- Representativeness and quality - Buelens, Daas, Burger, Puts, and van den Brakel (2014), Daas and Puts (2014b)
- General on new data sources - Choi and Varian (2012), Daas, Roos, de Blois, Hoekstra, ten Bosch, and Ma (2011); Daas and Puts (2014a), Hoekstra, ten Bosch, and Harteveld (2012)

However, in order to assess the issue of representativeness of this new data source it is important to define precisely what kind of data sources are involved and compare them with the existing ones. First of all, it should be emphasized that IDS are not defined in a statistical system or in the literature. For example, according to the The United Nations Economic Commission for Europe (UNECE) IDS can be a part of administrative sources, which are defined as *data collected by sources external to statistical offices*. On the other hand, the recent project *Big Data for Official Statistics*[3], led by UNECE, classified IDS as one type of big data sources. The project divides big data into three groups: Social Networks (human-sourced information), Traditional Business systems (process-mediated data) and the Internet of Things (machine-generated data). According to this division, IDS could be classified into the first two classes as Social Networks, Internet searches or E-commerce.

Big data is not a statistical term, but more of a general description used to capture certain characteristics of data sources. There is no specific time when the term was introduced but references can be found in Bayer (2011) and Bayer and Laney (2012). The definition consists of three aspects - high volume, high velocity and high variety (often described as 3V). The first V refers to the amount of data counted in tera- and petabytes, which are hard to analyse within the existing infrastructure. The second V denotes how these data are generated and change in time (e.g. web-logs, photo uploads). The last V indicates that big data occur in different formats, such as photos, texts, logs, videos etc. In comparison to classical data sources, like censuses or surveys, big data processing requires more effort in order to extract meaningful information. Some types of big data can occur in administrative sources, i.e. traffic sensor data, patients registers, land photos or car registers. However, in most cases such data are generated by users of specific types of portals (e.g. social networks) or services (e.g. mobile apps). That is why it is important to consider big data as a potential source of information about people or business activity.

For the purpose of this study, IDS are defined as *data collected and maintained by units external to statistical offices and administrative regulations available (mainly) on the Internet (through web-based databases)*. The definition contains two main aspects – first it explicitly states that data are collected by units other than official institutions and the purpose is not defined by official regulations. This element is crucial since the majority of data sources on the Internet are created by private companies. The definition also excludes official web pages that contain reports or statistics resulting from surveys or the use of registers (e.g. Eurostat Database, STATcube). The second part states that these data sources are available on the Internet through queries (e.g. via web-forms). Such portals could be devoted to price comparison, e-commerce, portals that include unit-data (e.g. offers on real estate market) or reports and aggregated data (e.g. Google Trends).

[3] http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics

IDS and big data have recently been under evaluation by NSIs for the production of statistics and enhancement or replacement of the existing data sources or data collection techniques. NSIs working papers address different aspects connected with the use of such data, for instance privacy or legality. These issues are outside the scope of this study and therefore will not be discussed in this paper. However, before new data sources can be used for statistics, they should meet the criteria that are applied to classical data sources. The following aspects should be discussed in the future: *conceptualisation, representativeness, selectivity, nonsampling errors, measurement of uncertainty, sampling, estimation (e.g. model-based estimation, Bayesian approach) or the place in statistical information system.*

The literature provides many definitions of representativeness, but none is given explicitly. Kruskal and Mosteller (1979a,b,c) provide a comprehensive literature review and list nine definitions of representativeness that refer to the following aspects: a general opinion about data, the lack of selective forces, the scaled-down version of the population, typical/ideal cases, whether it reflects variability of the population, how it refers to specific sampling methods (equality of probability of inclusion), whether it provides good estimation, whether it fits specific purposes. Most of the definitions in the statistical literature refer to respondents (people or companies) (see Schouten, Cobben, and Bethlehem 2009) and the suggestion of using propensity weighting. Bethlehem (2009) defines representativeness with respect to the sample when relative distributions are the same in the sample and in the population. It means that the sample is representative when characteristics of the sample and the population are the same. Following Kruskal and Mosteller (1979a,b,c) this statement can be understood to mean that a representative sample is the same as a scaled-down population. The measurement of representativeness of Internet research mainly refers to online surveys, online panels and pop-ups (Bethlehem 2008; Bethlehem and Biffignandi 2011). Buelens *et al.* (2014) recently proposed a diagram flow to measure selectivity of big data. In the first phase unit-level data are checked if they contain units and then their representativeness is assessed by linking them to existing sources or aggregating them for comparisons with other sources. Daas and Puts (2014b) proposed using co-integration tests to measure representativeness of trends.

3. Data sources on real estate market in Poland

Poland's real estate market is partially covered by official data sources. Data about this market come from three surveys on the management of housing resources, property sales and residential and commercial property prices supported by administrative registers and non-official data bases. The survey is conducted by the National Bank of Poland (NBP) in co-operation with the Central Statistical Office in Poland (CSO); it concerns both the primary and secondary market. Since NBP is mainly responsible for the analysis, the report mostly covers aspects connected with the macroeconomic analysis at the country and city level. It is a survey of brokers who deliver information on the primary and secondary market in the biggest cities in Poland. In addition, data from various administrative sources are collected, for instance, the number of brokers and other market participants are obtained from the National Official Business Register (REGON[4]) register and the transaction data are taken from the Register of Prices and Market Value of Property (pol. *Rejestr Cen i Wartości Nieruchomości*, PVP) that is administered by local government at Local Administrative Unit 1 level (LAU1 level, counties) and contains information on transactions on both markets. In addition, non-official databases (created in collaboration with brokers' associations) are used as well as databases created and supplied with information by NBP employees. However, from the statistical point of view, the methodology of this research is not clear. For instance, there is no information on the quality and response rate of survey data, nor is it clear how NBP databases are created or what the quality of the PVP register is.

[4] http://bip.stat.gov.pl/en/regon/

Nonetheless, the PVP register is an important data source of statistical data for researching the real estate market. The legal basis is described in the Act of Geodetic and Cartographic Law with amendments (1989) and the Regulation on the Land and Buildings by the Minister of Regional Development and Construction in Poland (2001). Under these two Acts notaries are obliged to inform local authorities about transactions involving land and property. Each transaction is described with detailed characteristics (e.g. floor area, location, building characteristics) and includes the transaction price. As stated by the law, the PVP register should cover all transactions in the biggest Polish cities at the LAU1 level. There are no reports on the quality of data that PVP contains nor about how PVP is used for statistics. In addition, access to the register is limited and granted only for the purpose of evaluating new properties or to NBP/CSO employees (as at the end of 2014).

Results of the research are published in two reports. The first one is devoted to information on prices (offers and transactions) on the primary and the secondary market on a quarterly basis (National Bank Of Poland 2014a). It contains point estimates and hedonic indexes for 17 biggest Polish cities aggregated at the LAU1 level and is based on a survey of brokers, non-official data and the PVP register. The second report is delivered on a yearly basis and provides a detailed description of macroeconomic indicators and characteristics of the real estate market for 17 biggest cities in Poland excluding their agglomerations (National Bank Of Poland 2014b). However, the second report is produced and published with a delay, for instance information for 2013 was available at the end of the 2014, which indicates that information is outdated and does not reflect the current state of the real estate market. In consequence, one can observe a growing interest in reports and surveys created in by other institutions in non-official settings. On the other hand, there is a lack of research devoted to the assessment of quality and uncertainty regarding this non-official information, given that the main data source is the Internet and web portals.

For the sake of clarity it should be noted how the Polish real estate market is organized. Market participants (excluding buyers and tenants) are brokers and owners and properties can be put up for sale directly by the owner or brokers. Properties are offered by agents under two types of agreements – exclusive and open. An exclusive agreement states that only one broker can offer a given property on the market. This type of agreement is not popular owing to the limited number of possible ways of reaching potential buyers. The second type of agreement is more popular and allows brokers to co-operate and exchange information on properties for sale. The organization of the market affects research – relations between properties for sale and owner/broker could be of the "many to many" type, making identification of units difficult. In particular, when agents are using web-portals devoted to the real estate market, offers may appear more than once. Nonetheless, in order to sell, brokers and owners need to inform potential buyers about properties for sale and the Internet is becoming the main channel.

IDS have been used for real estate market research in the past. Examples of such studies can be found in working papers of Statistics Netherlands (CBS). CBS uses Funda.nl (Hoekstra *et al.* 2012), maintained by the association of Dutch brokers (nl. *Nederlandse Vereniging van Makelaars*), which is responsible for the majority of transactions on the Dutch market, to obtain data on the secondary market and to link it with registers. To achieve this, CBS has adopted a web-scraping technique, whereby all necessary information is downloaded automatically (Hoekstra *et al.* 2012). IDS concerning the real estate market can be classified into four groups – brokers' portals, brokers' association portals, portals offering brokering assistance (both for agents and owners) and services that aggregate other web-pages. The proposed classification is important in terms of quality and coverage. For instance, one broker's official website contains nearly 4,100 offers of flats for sale on the secondary market in Warsaw, Poland, while portals offering brokering assistance feature 4,500 to 5,000 offers posted by the same agent. The differences can be seen not only between brokers' activities but also between web portals. For example, four biggest web portals in Poland (measured in terms of the number of visitors) `www.otodom.pl`, `www.dom.gratka.pl`, `www.domiporta.pl`

and www.szybko.pl offer respectively 304,000, 380,000, 321,000 and 167,000 flats on the secondary market in Poland[5]. Certainly, these numbers are biased for different factors – selectivity connected with preferences in the selection of portals, duplicate adverts within and between portals, outdated, erroneous or false sale offers.

Another issue reflecting the quality of research of the real estate market is the extent of Internet coverage. The CSO conducts *Information and Communications Technologies* (ICT, Central Statistical Office 2014) survey, which is part of The Digital Agenda for Europe programme run by the European Commission. According to this survey in 2012 98.6% of companies in section L (described below) had an Internet connection (97% in 2011), 74.5% have their own website (63,3% in 2011) and 37.7% used it to present their products and prices. In Poland companies are classified into different sectors and sections. Section L refers to the real estate market and consists of four groups of companies – *purchase and sale of property on one's own account*, *leasing and management of one's own or leasehold property*, *property brokerage* and *freelance property management*. Given the level of aggregation within this section, it is difficult to directly estimate the coverage of the Internet in the group of agencies and brokers that operate on the secondary real estate market. However, it could be assumed that this level is high in the 13 biggest Polish cities. In addition, the ICT survey does not measure the use of external portals. For instance, Polish web portals devoted to the real estate market enable brokers to have private websites within their domains. Another issue is that brokers can specialize in different aspects of the real estate market - houses, flats, commercial property, sale or renting, which could affect the use of the Internet. Moreover, the Polish property market is not regulated: there is no legal control over what agent is offering a property and where the original offer has been placed. However, taking into account that most buyers are young people, the IDS should not be neglected as a source of the statistical information.

4. Empirical evaluation of representativeness

For the purpose of the study the secondary real estate market was limited to flats (units of interest) that were offered only in Poznań, Poland between 2nd quarter of 2012 to 2nd quarter of 2014. It was motivated by the availability of official data and the limited scope of this paper. The http://www.nieruchomosci-online.pl portal (NOPL) was chosen, which, unlike other portals mentioned in section 3, offers free-of-charge access to historical unit-level data. However, it should be noted that the proposed approach can be extended not only to other cities or websites but also to different fields, where IDS could be used for statistics. For this study special R code was developed to scrape information from the portal.[6] **XML** (Lang 2013), **RCurl** (Lang 2014) and **httr** (Wickham 2015) packages were used for this purpose. Algorithm 1 presents the pseudo-code for the web-scraping.

Data: Web pages, N - number of search result pages (i), n - number of results on search page (j)
Result: Text file with scraped data
Send query through form on webpage and save link to results;
Set cookies for session ;
for $i \leftarrow 1$ **to** N **do**
  Enter i result page;
  Set n ;
  for $j \leftarrow 1$ **to** n **do**
    Scrape data from j result from search result page and write it into text file;
    Enter j page from the search result page;
    Scrape all text data from j page and write it into text file;
  end
end

Algorithm 1: Pseudo-code for the algorithm for web-scraping

[5] Information on 2014-10-07

[6] Available at GitHub https://github.com/BERENZ/Papers-supplements/blob/master/AJS/Codes/NOPL_scraper.R

The algorithm produces a text file containing all scraped information on prices, floor area, number of rooms and other characteristics that could be used to identify units. In the process of data cleaning the Register on Street Names and Addresses (TERYT) was used to harmonize street names. In addition, long text descriptions included in the ads were compared with information in the remaining ads. Offers that had erroneous price per square meter (eg. lower than 1,000 PLN/m^2 or higher than 100,000 PLN/m^2) were excluded from the analysis. In the next step data were cleaned and de-duplicated using probabilistic record linkage (Fellegi and Sunter 1969) implemented in **RecordLinkage** (Borg and Sariyar 2015). Probabilistic record linkage takes into account numeric, character and missing values to link records. A 80% threshold was set to determine the probability of two records referring to the same unit. To measure the degree of representativeness data were aggregated by quarter and the number of observations for each quarter can be found in Table 1. The number of observations varies over time and is connected with the availability of data for the beginning of 2012.

Table 1: Number of Poznań real estate offers from the secondary market obtained from http://www.nieruchomosci-online.pl

Quarter	2012Q2	2012Q3	2012Q4	2013Q1	2013Q2	2013Q3	2013Q4	2014Q1	2014Q2
Nobs	2,896	3,904	6,095	6,447	6,569	9,483	13,079	11,159	4,477

The main goal of the paper is to assess the representativeness of IDS for the real estate market. For this purpose, the definition proposed by Bethlehem (2009) was adopted and the distribution of three characteristics – price per square meter, number of rooms and floor area – were compared with official reports produced by NBP/CSO. The variables describing the number of rooms and floor area were harmonized to reflect the values in the official statistics data. As a result, the number of rooms and floor area had four levels: 1 room, 2 rooms, 3 rooms, 4+ rooms and under 40 m^2, [40 m^2, 60 m^2), [60 m^2, 80 m^2), over 80 m^2 respectively. For the sake of comparison, both the primary and the secondary market data reported by NBP/CSO were used. On the following plots each red, green and blue color indicates NBP/CSO offer estimates, NBP/CSO transaction estimates and NOPL estimates respectively. Due to the variability of the estimates, only trends for the three variables are compared. Trend estimation was conducted using loess regression (with default value for `span = 0.75`) implemented in `stat_smooth` function from **ggplot2** (Wickham 2009). `stat_smooth` is a wrapper for the `loess` function from **stats** package (R Core Team 2014). The loess regression was used for three reasons: the time series for comparison is short (9 quarters) and the application of time series models (e.g. AR, MA or ARMA) is difficult. Second, estimates obtained from the NBP/CSO and NOPL vary over time and are nonlinear, which makes direct comparison of data points impossible. Finally, estimates obtained from NOPL website contain nonsampling errors, which may introduce bias in the point estimates. That is why a comparison of the trend that is estimated from the data may indicate whether the changes in the trend are in the same direction as in NBP/CSO. In addition, the loess regression approach is often used for the aggregation of polls (see Bergman and Holmquist 2014).

Figure 1 presents the price per square meter according to NOPL (blue) and NBP/CSO (green and red). At the beginning of the period of interest we can observe an increase in the price; however, this change is probably due to the quality of the data in the first years of the research. From the 2nd quarter of 2008 the trend slightly decreases until the end of the 2012. The NBP/CSO price keeps increasing from the beginning of 2013 to the end of the period analysed. Between 2nd quarter of 2012 to 3rd quarter of 2013 the NOPL offer price is closer to the transaction price reported by NBP/CSO than to the offer price. However, the trend is comparable to the NBP/CSO price due to stability in time. At the end of 2014 the NOPL price rapidly rises to catch up with the NBP/CSO offer price, with little difference between the two trends of estimates. A comparison of the direction and level of price per m^2 indicates that with respect to this variable NOPL is not representative in the light of the Bethlehem (2009) definition. However, the variable is considered without detailed information on flats

and the limitations of published data do not allow comparisons in subgroups to detect which groups are under- or over-represented. In addition, due to the fact that price per m^2 is considered as an output statistics, the relative distribution of flat characteristics need to be investigated.

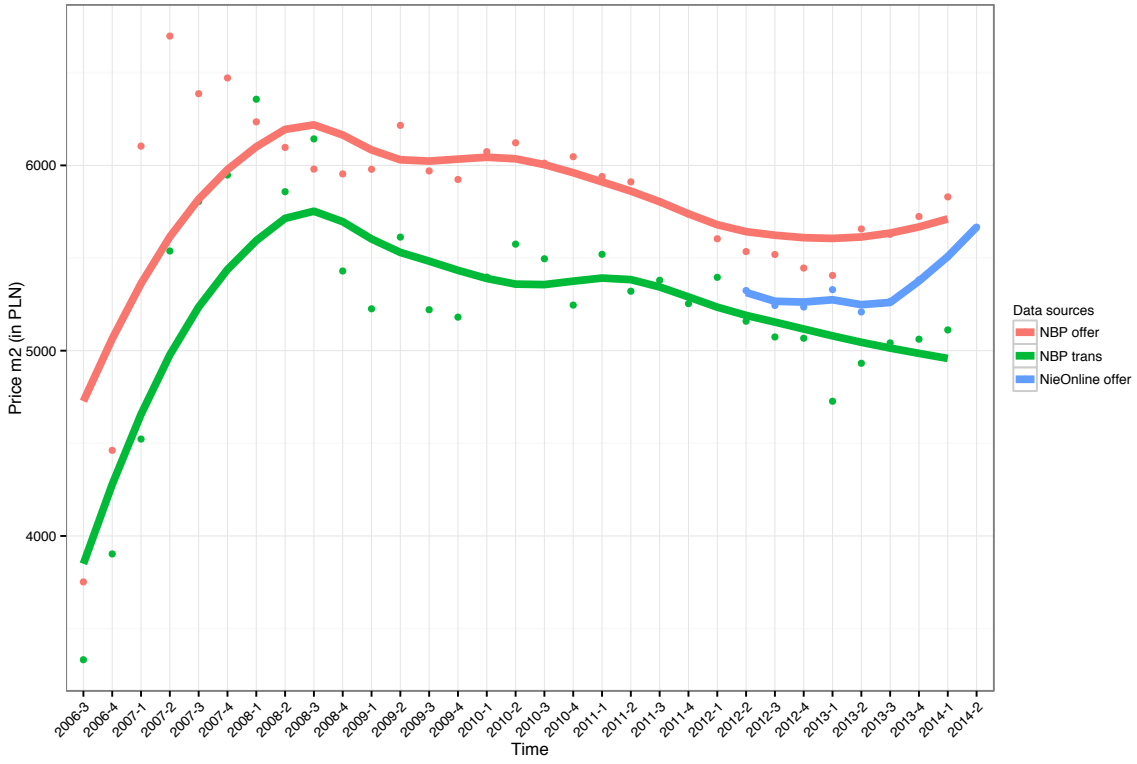

Figure 1: Comparison of offer and transaction price per square meter in Poznań, Poland reported by NBP/CSO and obtained from NOPL

In order to assess the representativeness of NOPL data, a relative distribution of floor area and the number of rooms of flats was compared with NBP/CSO reports. Figure 2 presents a comparison of harmonized floor area of flats in four categories reported by NBP/CSO. The smallest flats (under 40 m^2) and medium-sized (60-80 m^2) are represented at the same level as in official statistics data, and the trends are consistent for these groups. The blue line representing NOPL is close to the red line representing the NBP/CSO offers and is different from the green line denoting transactions. Nonetheless, the biggest (over 80 m^2) and smaller (40-60 m^2) flats are under-represented compared to official statistics data. The difference between the fraction of 40-60 m^2 flats is constant in time and the trend is consistent with that reported by NBP/CSO. In the case of the biggest flats the fraction of flats offered for sale is not at the same level as in official statistics data - the trend has a smaller slope while the direction is the same.

Figure 3 presents a comparison of number of rooms aggregated to the four categories defined in the NBP/CSO reports – 1 room, 2 rooms, 3 rooms and 4 and more rooms. In the case of flats with 1 room the NOPL trend is slightly shifted in time and reaches the same percentage of flats but with one quarter delay. However, the trend is consistent in the sense of shape and direction with the one plotted on the basis of official statistics data. The trend for flats with 3 rooms reaches the same level as in NBP/CSO reports and the differences between the trends are minor. The main differences are visible in the group of biggest flats that are under-represented in comparison with official statistics data. Nonetheless, the trend is comparable to the NBP/CSO flat offers denoted by the red line. In contrast, the trend for 2-room flats for most of the period is consistent, although in the 2nd quarter of 2012 a change in the slope can be observed. which is more comparable with transaction data, or perhaps, it indicates a change in the trend which will be visible in the upcoming report.

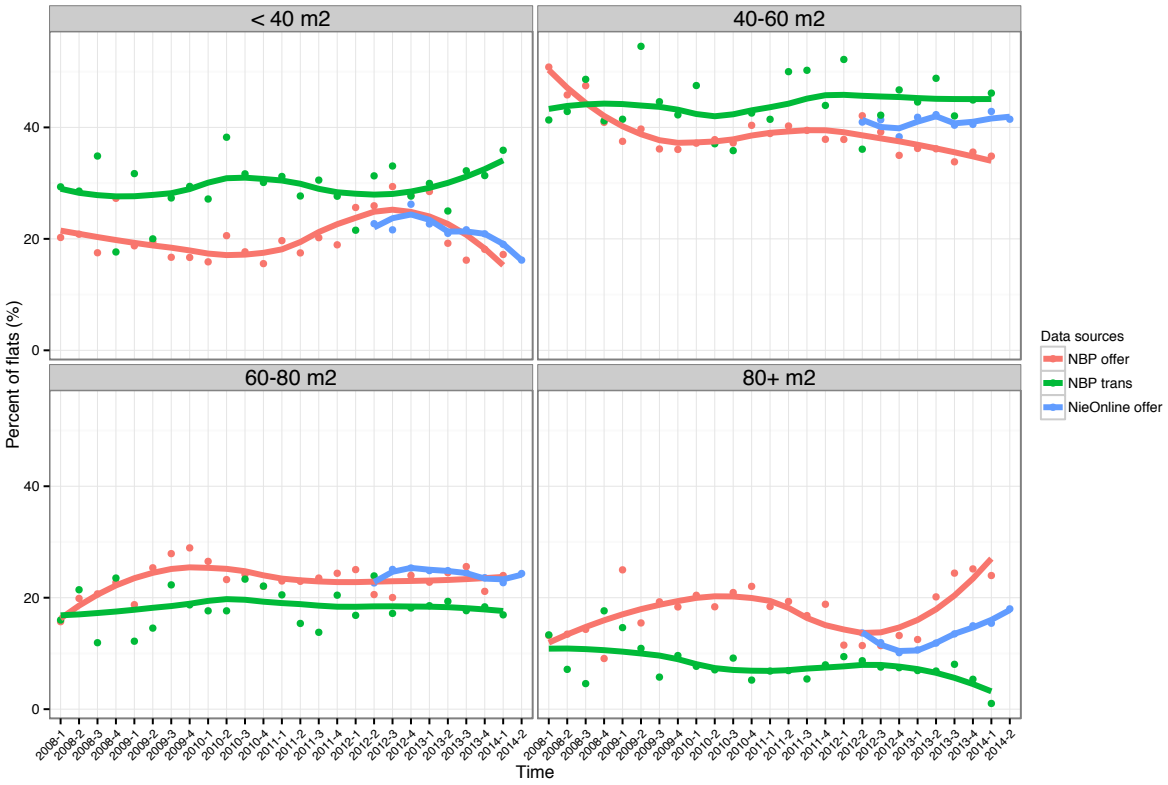

Figure 2: A comparison of characteristics of flats offered and sold on the secondary market as reported by NBP/CSO and obtained from NOPL by floor area

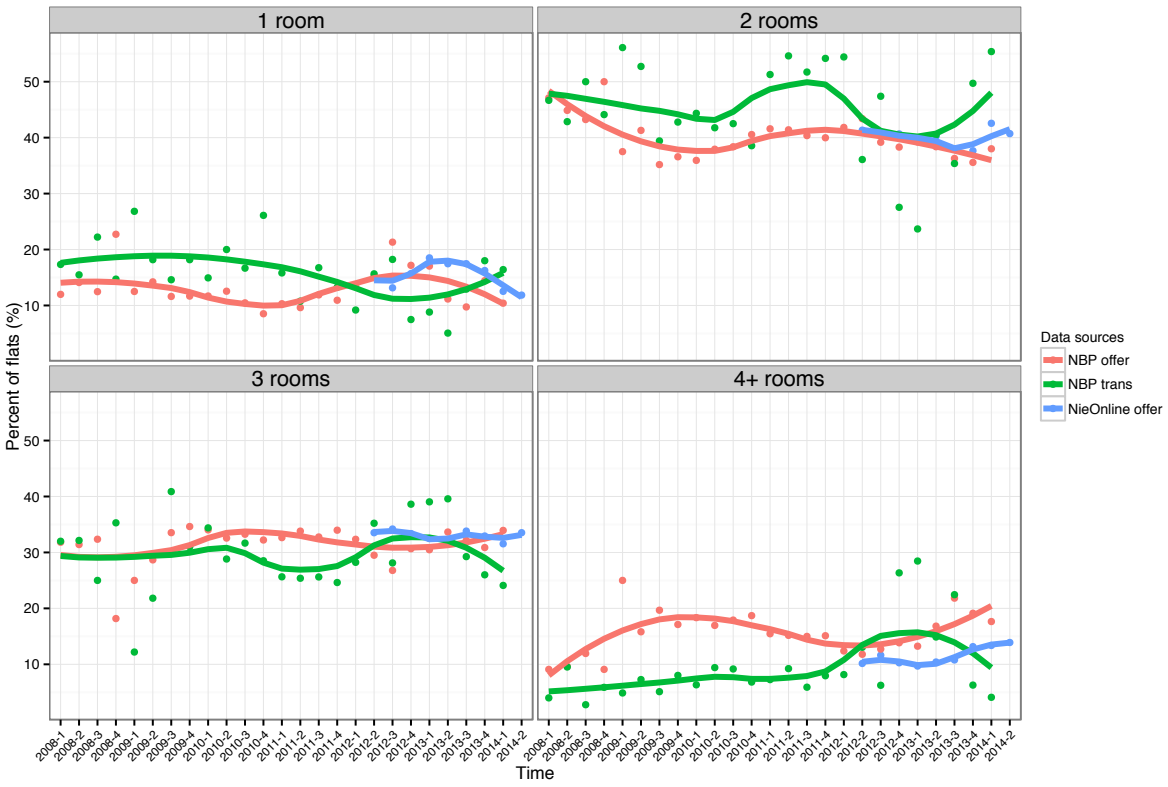

Figure 3: Comparison of characteristics of flats offered and sold on the secondary market as reported by NBP/CSO and obtained from NOPL by the number of rooms

The comparison of trends indicates that data obtained from NOPL are representative for the smallest (below 40 m^2, 1 room) and medium-sized (60-80 m^2, 3 rooms) flats. On the other hand, the group of the biggest (80+ m^2, 4+ rooms) are under-represented in comparison with official statistics data presented in NBP/CSO reports. However, the decrease in the category of under 40 m^2 flats appears first in the NOPL source and is then reflected in the NBP/CSO source, because reports from 2013 appeared at the end of the 2014. This indicates that the NOPL source could be regarded as an indicator for this category of flats. A similar relationship for big flats with floor area over 80 m^2 and 4+ rooms can be observed where an increase in the trend is first indicated in the NOPL source. As a result, the differences in the categories of flats may influence the discrepancies in the price per m^2 presented in Figure 1.

In addition, due to the non-sampling character of data obtained from the Internet, it is challenging to estimate standard errors for the estimated characteristics. In addition, NBP/CSO reports do not contain any information on standard errors of estimates, which again limits the scope of comparison of distributions. Therefore, visual analysis can be useful to detect trends and their relations with official data sources could be the first indicator of representativeness for new data sources.

5. Summary and discussion

IDS and big data have recently become the subject of evaluation by statisticians as potential statistical data sources. Despite the increasing interest in these new data sources there are several aspects that need to be considered in order to meet the criteria of statistical data sources. In order to discuss the representativeness of the IDS the definition of IDS was presented in the paper. The comparison of trends estimated by loess regression was proposed as a indication of the representativeness of the NOPL in comparison with official statistics data. Results presented in the paper indicate that for two categories (under 40 m^2 and 60-80 m^2) of flats the NOPL source could be considered representative, while the biggest flats are under-represented. However, it should be noted that the results reflect the secondary real estate market only in one city and such analysis should be extended to include other cities as well as other web-portals. Furthermore, estimation can be affected by the selection of web portals and by the data cleaning process, which in turn can influence the measuring of representativeness.

The results of the study suggest that the existing definitions and methodology, which are valid for existing statistical data sources, should be adopted or revised to deal with new data sources. Another problem is the lack of reference data from official statistics or its limited scope, which makes it difficult to measure representativeness or, more importantly, uncertainty of estimates. On the other hand, IDS cannot often be compared with existing research due to the lack of consistency and harmonized definitions. Therefore, information obtained from IDS could be treated as a proxy measure of sociological or economical phenomena (e.g. Google Trends). Moreover, there is a lack of statistical literature directly connected with estimation problems related to new data sources. Furthermore, IDS and big data should be treated as non-probability samples, whose representativeness is hard to measure. Recently Wanga, Rothschildb, Goelb, and Gelman (2014) proposed Bayesian model-based estimation and post-stratification that could be one of the possible approaches to the problem. Nonetheless, new data sources open up possibilities of extending the set of statistical sources, which should not be neglected.

Acknowledgement

The article and the research has been financed by National Science Centre Poland, Preludium 7 grant no. 2014/13/N/HS4/02999. In addition, I would like to thank two anonymous reviewers for helpful comments.

References

Bapna R, Goes P, Gopal R, Marsden JR (2006). "Moving from Data-Constrained to Data-Enabled Research: Experiences and Challenges in Collecting, Validating and Analyzing Large-Scale e-Commerce Data." *Statistical Science*, **21**(2), 116–130. ISSN 0883-4237. doi:10.1214/088342306000000231. 0609136v1, URL http://projecteuclid.org/Dienst/getRecord?id=euclid.ss/1154979815/.

Bayer M (2011). "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data." URL http://www.gartner.com/newsroom/id/1731916.

Bayer M, Laney D (2012). "The Importance of 'Big Data': A Definition." URL https://www.gartner.com/doc/2057415/importance-big-data-definition.

Bergman J, Holmquist B (2014). "Poll of Polls : A Compositional Loess Model." *Scandinavian Journal of Statistics*, **41**(2), 301–310. doi:10.1111/sjos.12023.

Bethlehem J (2008). "Representativity of Web Surveys–An Illusion?" *Access panels and online research, panacea or pitfall*, pp. 19–44.

Bethlehem J (2009). *Applied Survey Methods: A Statistical Perspective*. John Wiley & Sons.

Bethlehem J, Biffignandi S (2011). *Handbook of Web Surveys*. John Wiley & Sons.

Borg A, Sariyar M (2015). *RecordLinkage: Record Linkage in R*. R package version 0.4-7, URL http://CRAN.R-project.org/package=RecordLinkage.

Buelens B, Daas P, Burger J, Puts M, van den Brakel J (2014). "Selectivity of Big Data." URL http://www.pietdaas.nl/beta/pubs/pubs/Selectivity_Buelens.pdf.

Cavallo A (2012). "Scraped Data and Sticky Prices." *MIT Sloan Research Paper*. URL http://www.mit.edu/%7Eafc/papers/Cavallo-Scraped.pdf.

Cavallo A (2013). "Online and Official Price Indexes: Measuring Argentina's Inflation." *Journal of Monetary Economics*, **60**(2), 152–165.

Central Statistical Office (2014). *Information Society in Poland Statistical Results From the Years 2009-2013 (in Polish)*. Statistical Office in Szczecin, Warsaw, Poland. URL http://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5497/1/7/4/spolecz_inform_w_polsce_2009-2013.pdf.

Choi H, Varian H (2012). "Predicting the Present with Google Trends." *Economic Record*, **88**(s1), 2–9.

Daas P, Puts M (2014a). "Big Data As a Source of Statistical Information." *The Survey Statistician*, **69**, 22–31. URL http://pietdaas.nl/beta/pubs/pubs/Big_data_survey_stat.pdf.

Daas P, Puts M (2014b). "Social Media Sentiment and Consumer Confidence." URL http://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.pdf.

Daas P, Roos M, de Blois C, Hoekstra R, ten Bosch O, Ma Y (2011). "New Data Sources for Statistics: Experiences at Statistics Netherlands." In *Paper for the 2011 European New Technique and Technologies for Statistics conference, February*, pp. 22–24.

Daas P, Roos M, van de Ven M, Neroni J (2012). "Twitter As a Potential Data Source for Statistics." URL http://pietdaas.nl/beta/pubs/pubs/DiscPaper_Twitter.pdf.

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

Fondeur Y, Karamé F (2013). "Can Google data help predict French youth unemployment?" *Economic Modelling*, **30**, 117–125. ISSN 02649993. doi:10.1016/j.econmod.2012.07.017. URL http://linkinghub.elsevier.com/retrieve/pii/S0264999312002490.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2008). "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature*, **457**(7232), 1012–1014.

Gołata E (2014). "New Paradigm in Statistics and Population Census Quality." European conference on quality in official statistics, URL http://www.q2014.at/fileadmin/user_upload/GOLATA_NEW.pdf.

Hoekstra R, ten Bosch O, Harteveld F (2012). "Automated Data Collection From Web Sources for Official Statistics: First Experiences." *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, **28**(3), 99–111.

Kruskal W, Mosteller F (1979a). "Representative Sampling I: Non-scientific Literature." *International Statistical Review*, **47**, 13–24. URL http://www.jstor.org/stable/1402564.

Kruskal W, Mosteller F (1979b). "Representative Sampling II: Scientific Literature Excluding Statistics." *International Statistical Review*, **47**, 111–123. URL http://www.jstor.org/stable/1402564.

Kruskal W, Mosteller F (1979c). "Representative Sampling III: The Current Statistical Literature." *International Statistical Review*, **47**, 245–265. URL http://www.jstor.org/stable/1402647.

Lang DT (2013). *XML: Tools for Parsing and Generating XML Within R and S-Plus*. R package version 3.98-1.1, URL http://CRAN.R-project.org/package=XML.

Lang DT (2014). *RCurl: General Network (HTTP/FTP/...) Client Interface for R*. R package version 1.95-4.3, URL http://CRAN.R-project.org/package=RCurl.

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

Mohorko A, Leeuw Ed, Hox J (2013). "Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time." *Journal of Official Statistics*, **29**(4), 609–622.

National Bank Of Poland (2014a). *Real Estate Market – Quarterly Report*. National Bank of Poland, Finance stability department, Warsaw, Poland. URL http://www.nbp.pl/homen.aspx?f=/en/publikacje/inne/real_estate_market_q.html.

National Bank Of Poland (2014b). *Report On the Situation in the Polish Residential and Commercial Real Estate Market in 2013*. National Bank of Poland, Finance stability department, Warsaw, Poland. URL http://www.nbp.pl/en/publikacje/inne/annual_report_2013.pdf.

Porter AT, Holan SH, Wikle CK, Cressie N (2013). "Spatial Fay-Herriot Models for Small Area Estimation with Functional Covariates." *arXiv preprint arXiv:1303.6668*.

Pratesi M, Giannotti F, Giusti C, Marchetti S, Pedreschi D, Salvati N (2014). "Area Level Sae Models with Measurement Errors in Covariates: An Application to Sample Surveys and Big Data Sources." Small Area Estimation, URL http://sae2014.ue.poznan.pl/SAE2014_book.pdf.

Pratesi M, Pedreschi D, Giannotti F, Marchetti S, Salvati N, Maggino F (2013). "Small Area Model-Based Estimators Using Big Data Sources." NTTS, URL http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_208.pdf.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schouten B, Cobben F, Bethlehem J (2009). "Indicators for the Representativeness of Survey Response." *Survey Methodology*, **35**(1), 101–113.

Shmueli G, Jank W, Bapna R (2005). "Sampling eCommerce Data From the Web: Methodological and Practical Issues." In *ASA Proc. Joint Statistical Meetings*, volume 941, p. 948. URL https://archive.nyu.edu/bitstream/2451/14953/2/USEDBOOK11.pdf.

Vosen S, Schmidt T (2011). "Forecasting Private Consumption: Survey-Based Indicators Vs. Google Trends." *Journal of Forecasting*, **30**(6), 565–578.

Wallgren A, Wallgren B (2014). *Register-based Statistics*. Wiley Series in Survey Methodology, second edition. John Wiley & Sons, Inc. ISBN 9781119942139.

Wanga W, Rothschildb D, Goelb S, Gelman A (2014). "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting. Forthcoming.*

Wickham H (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Springer New York. ISBN 978-0-387-98140-6. URL http://had.co.nz/ggplot2/book.

Wickham H (2015). *Httr: Tools for Working with URLs and HTTP*. R package version 0.6.1, URL http://CRAN.R-project.org/package=httr.

Xu W, Li Z, Cheng C, Zheng T (2012). "Data Mining for Unemployment Rate Prediction Using Search Engine Query Data." *Service Oriented Computing and Applications*, **7**(1), 33–42. ISSN 1863-2386. doi:10.1007/s11761-012-0122-2. URL http://link.springer.com/10.1007/s11761-012-0122-2.

Zhang LC (2011). "A Unit-Error Theory for Register-Based Household Statistics." *Journal of Official Statistics*, **27**(3), 415–432.

Zhang LC (2012). "Topics of statistical theory for register-based statistics and data integration." *Statistica Neerlandica*, **66**(1), 41–63. ISSN 00390402. doi:10.1111/j.1467-9574.2011.00508.x.

Zhang LC (2015). "On Modelling Register Coverage Errors." *Journal of Official Statistics. Forthcoming.*

Affiliation:

Maciej Beręsewicz
Department of Statistics
Poznan University of Economics
61-875 Poznan, Poland
E-mail: maciej.beresewicz@ue.poznan.pl

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 44 *Submitted:* 2014-10-07
April 2015 *Accepted:* 2015-02-18