Austrian Journal of Statistics
MMMMMM YYYY, Volume VV, 1–xx.
http://www.ajs.or.at/

On representativeness of Internet data sources for real estate market in Poland

Maciej Beręsewicz
Poznan University of Economics

Abstract

Shifting paradigms in Official Statistics lead to a widespread use of administrative records to support or to create an alternative for census and surveys. At the same time demand for diversified detailed information is increasing. Official Statistics in order to meet this demand need to seek for new data sources. Internet data sources or more general – Big Data – could be one of them. Potential usefulness of these new sources of statistical information should not be neglected.

The aim of the paper is to assess representativeness of Internet data sources (IDS) for real estate market in Poland. These sources could be used for describing demand and supply on secondary real estate market in more detailed way that is done with existing methodology. In order to assess representativeness, information from official surveys and other data sources will be used. Due to lack of sufficient literature on this issue, own research will be conducted to enhance information from official statistics. For the purpose of the paper Internet data sources will be defined. Register TERYT containing information on street names was used to correct information taken from Internet data sources. Special program for automated data collection (*web spider*) was developed. All the calculation was done with R statistical software and additional packages (**XML**, **RCurl** and **httr**).

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

1. Introduction

Increasing information needs at a low level of aggregation lead not only to the development of small area estimation but also stimulate the search for new data sources that could support or enhance existing sources (reporting, census or surveys). This process has been continuing since 1970s when statisticians and NSIs started using and adopting administrative records into statistical system (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 means 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 target population, which sometimes could be erroneous (see Gołata (2014), Zhang (2014)). However, in the environment of electronic economy, characterized by the increasing use of the Internet both by households and companies and 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 is problematic to handle with 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 investigated widely in the light of statistical data source or 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 Internet data sources.

The paper has the following structure. The second section defines and presents Internet data sources in the context of survey methodology. The key concept – representativeness – will be defined and discussed in the context of new data sources in the Internet. Relation to the characteristics of Big Data will be underlined in the light of statistical data sources. The third section will be devoted to data sources for the real estate market in Poland and possibilities of the use of IDS for obtaining statistical information will be presented. The penultimate section will be devoted to an empirical evaluation of representativeness based on the example of the Polish real estate web portal – nieruchomosci-online.pl. For this purpose special R (R Core Team 2014) program was written to automatically obtain data from the web portal. The last section will be devoted to the discussion of results and final remarks.

2. Internet data sources

While the access to the Internet in households is increasing (Mohorko, Leeuw, and Hox 2013), the way of communication between people as well as companies is changing (e.g. C2C, B2C, 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* conducted by MIT downloads data from over 60 countries and calculates price indexes and measures of macroeconomics phenomenon (Cavallo 2012, 2013). However, the exact methodology and web scraping technique is protected by PriceStats, a start-up founded at MIT. 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 recent years new research devoted to Big Data and Internet data sources 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 and quality of this new data source it is important to define precisely what kind of data sources are discussed and compare them with the existing data sources. First of all, it should be emphasized that Internet data sources are not defined in a statistical system nor in the literature. For example, according to the United Nation Economic Commission for Europe Internet data sources can be a part of administrative sources, which are defined as *data collected by sources external to statistical offices*. On the other hand, the project *Big Data for Official Statistics* lead by UNECE classified Internet data sources as a 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 Internet of Things (machine-generated data). Internet data sources 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 of data sources that describes the characteristics of the data. 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 (3V). The first V refers to the amount of data counted in tera- and petabytes, which is hard to analyze within the existing infrastructure. The second V denotes how this data is generated and it changes in time (e.g. web-logs, photo uploads). The last V indicates that this data occurs in different formats, such as photos, texts, logs, videos etc. In comparison to classical data sources, like census or surveys, this data requires more effort to process it in order to get meaningful information. Some types of Big Data can occur in administrative sources, i.e. traffic sensors, patients registers, land photos or car registers. However, in most cases this data is 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 this data as a potential source of information about people or business activity.

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

Internet data sources and Big data have recently been under evaluation by NSIs for the production of statistics and replacement or enhancement of existing data sources. Reports raise different aspects connected with the use of such data, e.g. privacy or legal aspects of

using this data. They are outside the scope of this research and therefore will not be discussed in this article. 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 many provides many definitions of representativeness, but none is given explicitly. Kruskal and Mosteller (1979a,b,c) made a comprehensive literature review and collected nine definitions of representativeness. In various papers their authors refer to the following aspects: general opinion about data, lack of selective forces, the scale-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 scale-down population. The measurement of representativeness of Internet research mainly refers to Internet surveys, Internet 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 is checked if it contains units and then representativeness is assessed by linking to existing sources or aggregated for comparison with other sources. Daas and Puts (2014b) proposed using cointegration tests to measure representativeness of trends.

3. Data sources on real estate market in Poland

The Polish real estate market is partially covered by official data sources. Main research devoted to this market consists of three surveys supported by administrative registers and non-official data bases - management of housing resources, property sales and residential and commercial property prices. Research is conducted by the National Bank of Poland (NBP) in co-operation with the Central Statistical Office in Poland (CSO); it concerns both primary and secondary market. Since it is NBP that is mainly responsible for the analysis, the report mostly covers aspects connected with the macroeconomic analysis at the country and city level. Research is conducted as a survey of brokers who deliver information about the primary and secondary market. In addition, data from various administrative sources are collected, for instance, the number of brokers and other market participants are taken from the REGON[1] register, which contains companies listed in the statistical system or the Register of Prices and Market Value of Property (pol. *Rejestr Cen i Wartości Nieruchomości*, PVP) that is administered by local government at LAU1 level (poviats). In addition, non-official databases 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 how NBP databases are created or what the quality of PVP register is.

Nonetheless, the Register of Prices and Market Value of Property 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). According to the Acts notaries are obliged to inform local government authorities about transactions involving land and property. Each transaction is described with detailed

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

characteristics (e.g. surface, location, building characteristics) and the transaction price. However, as stated by the law, the PVP register should cover all transactions at the LAU1 level. There are no reports on the quality of data that it contains nor about how it is used for statistics. In addition, access to the register is limited and granted only for the purpose of valuating new properties or to NBP/CSO employees.

The study results in the publication of two reports. The first one is devoted to information on prices (offers and transactions) on the primary and secondary market on a quarterly basis (National Bank Of Poland 2014a). It contains point estimates and hedonic indexes for 13 biggest Polish cities aggregated at the city 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 covers in detail macroeconomic indicators and characteristics of the real estate market for 13 biggest cities in Poland excluding their agglomerations (National Bank Of Poland 2014b). However, this 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. As a result, there is a growing interest in reports and surveys created by non-official institutions in Poland. On the other hand, there is no research devoted to the assessment of quality and uncertainty related to this non-official information, especially considering 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 could be (excluding buyers and tenants) brokers and owners and properties can be put up for sale directly by the owner or by agents. The legal basis for properties offered by agents is provided by two types of agreements - exclusive and open offer. The exclusive offer states that only one broker can offer a given property on the market and they are responsible i.a. for promotion. 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 the research – relations between properties for sale and owner/broker could be of the "many to many" type and identification of units can be problematic. In particular, when different agents are using web-portals devoted to the real estate market may, offers may be duplicated. Nonetheless, in order to sell, brokers or owners need to inform potential buyers about properties for sale and the Internet can be one of the channels.

Examples of the use of IDS for the real estate market could be found in working papers of the Dutch NSI. Statistics Netherlands uses Funda.nl (Hoekstra *et al.* 2012), owned by the association of Dutch brokers (nl. *Nederlandse Vereniging van Makelaars*) which is responsible for the majority of transaction on the Dutch market, to obtain data about the secondary market and to link it with registers. To achieve this, Statistics Netherlands relies on a web-scraping technique, whereby all necessary information is downloaded automatically (Hoekstra *et al.* 2012). The IDS in the case of 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 portals that aggregate other portals. The classification is important in terms of quality. For instance, one broker's official web page contains nearly 4100 offers of flats for sale on the secondary market in Warsaw, Poland, while on portals that offer brokering assistance the same agent presents from 4500 to 5000 offers. 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) otodom.pl, gratka.pl, domiporta.pl and szybko.pl offer respectively 304 000, 380 000, 321 000 and 167 000 flats on the secondary market in Poland. 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 that reflects the quality of research of the real estate market is penetration of the Internet. The Central Statistical Office 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 (97% in 2011) 98.6% of companies in section L had Internet connection, 74.5% have their own webpage (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 and 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, one cannot directly estimate the penetration of the Internet in the group of agencies and brokers that deal with the secondary real estate market. However, it could be assumed that this level is higher in bigger cities. In addition, the survey does not address the use of external portals. For instance, on Polish web portals that offer property brokerage it is possible for brokers to have their own web pages. 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 in the sense that there is no legal control over who is offering the property and where it is being offered. However, taking into account that most buyers are young people, IDS seem to be one of the data sources that should not be neglected as a source of statistical information.

4. Empirical evaluation of representativeness

For the purpose of the study secondary real estate market was limited to flats that were offered in Poznań, Poland. It was motivated by availability of official data and limited scope of this paper. As an example of Internet data sources portal nieruchomosci-online.pl (NOPL) was chosen. In comparison to other portals that was mentioned earlier it offers free of charge access to archived unit data. This allowed to speed up the research due to possibility of compare prices in the time without waiting for new scraped data. For this study special program in R was developed to scrape information from the portal. **XML** (Lang 2013), **RCurl** (Lang 2014) and **httr** (Wickham 2014) packages were used for this purpose. The algorithm is as follows:

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 throught 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

As a result of the procedure text file containing all scraped information was created. It contained information on prices, surface, rooms and other characteristics that could be used for the identification of units. In the process of data cleaning register on street names and addresses TERYT[2] was used to harmonised street names. In addition other methods for extracting information from text (especially long descriptions) was used to check if contain the same information as in description on the page. Offers that had erroneous price per square

[2] http://bip.stat.gov.pl/en/teryt/

meter was excluded from the analysis. After this process data was cleaned and deduplicated using **RecordLinkage**. For comparison purposes data was aggregated quarterly and the table 1 present number of observations for each quarter. Number of observations vary over time and it is connected with availability of data for the beginning of 2012.

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

Quarter	2012Q2	2012Q3	2012Q4	2013Q1	2013Q2	2013Q3	2013Q4	2014Q1	2014Q2
Nobs	2896	3904	6095	6447	6569	9483	13079	11159	4477

The main goal of the paper is to assess representativeness of Internet data sources for real estate market. For this purpose definition proposed by Bethlehem (2009) was adopted and distribution of three characteristics was compared with official reports produced by National Bank of Poland and Central Statistical Office. Due to lack of sufficient number of variables in the reports following was chosen - price per square peter, number of rooms and flat surface. Two last variables was harmonized for comparison reasons. In result, number of rooms and surface had four categories 1 room, 2 rooms, 3 rooms, 4+ rooms and below 40 m2, <40 m2, 60 m2), <60 m2, 80 m2), over 80 m2 respectively. For the sake of comparison both primary and secondary market data was used. Flowing plots present distribution of characteristics on quarterly basis. Each point indicates estimates from official statistics reports and NOPL service. For the sake of clarity on each figure local polynomial regression was added with function *geom_smooth* from **ggplot2** (Wickham 2009) package which indicates trends in measured variables.

Figure 1 present price per square meter according to the NOPL source (red colour) and NBP/CSO source (blue and green). It could be observed that offer price for NOPL in the beginning of the period is closer to the transaction price however, in the end trend changes and increased to the level of offer price. Shape and direction of the trend for NOPL source is similar to offer price in NBP/CSO source. Relation that could be observed in figure 1 indicates that for the new quarters, price per square meter obtained from NOPL sources could reflect state of real estate market.

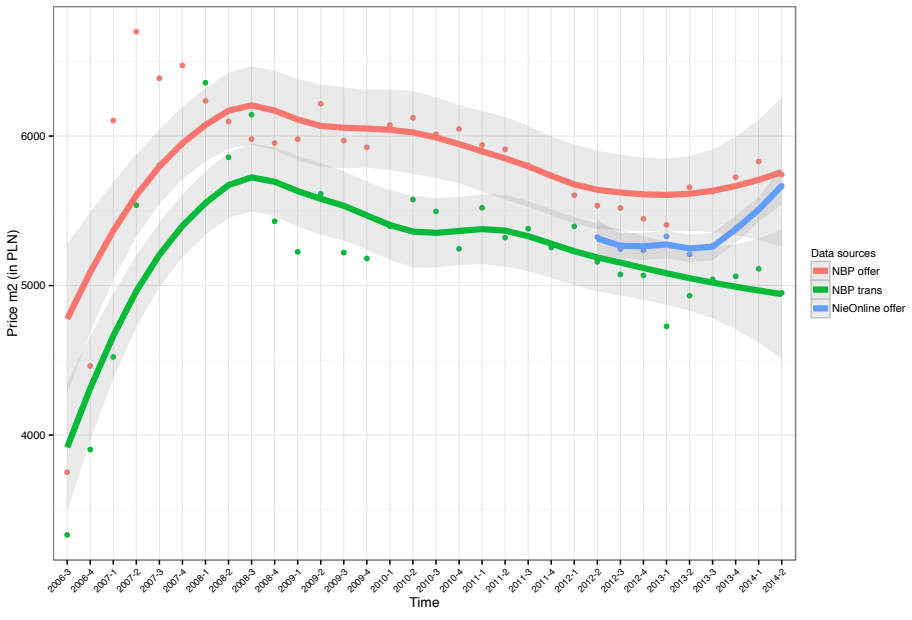
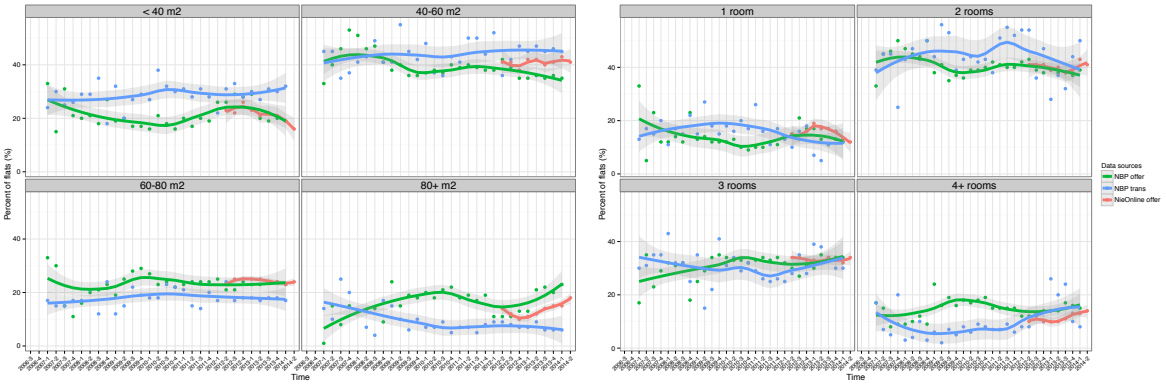


Figure 1: Comparison of offer and transaction price per square meter in Poznań, Poland presented in Official Statistics (NBP) and on Nieruchomosci-online web portal

In order to assess representativeness of NOPL data relative distribution of surface and number

of rooms of flats was compared with official statistics. Figure 2 consist of two panels - first (2a) on the left present surface and second (2b) on the right reflects the number of rooms. On both panels red colour denotes NOPL source, blue and green colour stands for offer and transaction from NBP/CSO source. In both cases similar distributions for the smalest (below 40 m^2, 1 room) and middle (<60,90) m^2, 3 rooms) categories could be observed. On the other hand, surface in group of <40,60) m^2 is underrepresented and over 80 m^2 is overrepresented in comparison to the offers in NBP/CSO source. However, category of flats with 2 and 4+ rooms are underestimated for NOPL source.



(a) Surface (b) Numer of rooms

Figure 2: Comparison of offer and transact characteristics of flats on secondary market in official statistics and scraped from NOPL portal

Nonetheless, trends that are created with loess regression indicates the same direct as in NBP/CSO source. For instance, decrease in flat category with surface below 40 m^2 could be first observed in NOPL source and then was reflected in NBP/CSO source because reports from 2013 appeared in the end of the 2014. This indicates that NOPL source could be probably indicator for this category of flats. Similar relationship for big flats with surface over 80 m^2 and 4+ rooms could be observed where increase in trend is first indicated in NOPL source.

However, due to non-sampling character of the data from obtained from the Internet it is challenging to estimate standard errors for the estimated characteristics. In addition, no information on standard errors of estimates in NBP/CSO are presented with again limits scope of comparison of distributions. Therefore visual checking could be useful for detection of trends and its relation with official data sources could be first indicator of representativeness for new data sources.

5. Summary and final remarks

Internet data sources and Big data lately are under evaluation by statisticians in the context of statistical data source. Despite the increasing interest in these new data sources there are several aspect that need to be considered in order to meet criteria of statistical data source. To be able to discuss the representativeness of the IDS the definition of Internet data sources was presented in the paper. In addition, appropriate measure for representativeness was chosen from the wide range of possibilities, which may be useful suggestion for further research.

Research presented in the paper indicates that using IDS for real estate market can be useful for estimation of trends in the considered variables. The main focus in paper was limited only to one data source and city. Therefore comparison to other web portals and cities should be further examined. Furthermore, selection of web portals can affect estimation as well as data cleaning process, which in turn can influence measuring of the representativeness.

The results of the research show that the methodology for surveys should be adopted or

revised to deal with new data sources. One of the remaining problems is lack of reference data from official statistics or its limited scope which leads to problems with measuring representativeness or more importantly uncertainty of estimates. On the other hand, IDS are often impossible to compare with existing research due to lack of coherence and harmonized definitions. Therefore, information obtained from IDS could be treated as a proxy measure of sociological or economical phenomenon (e.g. Google Trends). Moreover, there is lack of statistical literature directly connected with estimation problems in the context of statistical data source. Furthermore, IDS and Big Data should be treated as a non-probability samples, which could as well cause problems with measuring representativeness. Recently Wanga, Rothschildb, Goelb, and Gelmana (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 new possibilities for extending existing statistical sources and it should not be neglected.

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.

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.

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.

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). *The real estate market - Information Quarterly (in polish)*. Finance stability department, Warsaw, Poland. URL http://nbp.pl/home.aspx?f=/publikacje/rynek_nieruchomosci/index2.html.

National Bank Of Poland (2014b). *Report on the situation on the markets of residential and commercial property in Poland in 2013 (in polish)*. Finance stability department, Warsaw, Poland. URL http://nbp.pl/publikacje/rynek_nieruchomosci/raport_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, Gelmana 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 (2014). *httr: Tools for working with URLs and HTTP*. R package version 0.5, 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 (2014). "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 VV *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd