POZNAŃ UNIVERSITY OF ECONOMICS
AND BUSINESS

DOCTORAL DISSERTATION

Internet data sources for real estate market statistics

Author:
Maciej BERĘSEWICZ

Supervisor:
dr hab. Elżbieta GOŁATA, prof. nadzw. UEP
Auxiliary Supervisor:
dr Marcin SZYMKOWIAK

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in the

Department of Statistics
Faculty of Informatics and Electronic Economy

2016

POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Acknowledgements

This dissertation would not have been possible without the help and support of many people. I would first like to express my deepest appreciation to my supervisor, Professor Elżbieta Gołata, and my auxiliary supervisor and friend Dr Marcin Szymkowiak, whose encouragement, guidance, patience and support from the initial to the final stage have helped me to develop a better understanding of the subject. Without their guidance, endless hours of discussions and persistent help this dissertation would not have been possible. I would also like to acknowledge the contribution of Professor Jan Paradysz, who introduced me to the topic and supervised me in my first year of PhD studies.

I would like to thank my colleagues at the Department of Statistics of Poznań University of Economics and Business and the Center for Small Area Estimation at Statistical Office in Poznań. Without their support I would not have reached this stage of my life.

This research has been supported by the National Science Centre, Preludium 7 grant no. 2014/13/N/HS4/02999. I am grateful for their support and the confidence they have placed in me.

Lastly, I would like to thank my family, for all their love, encouragement, and always believing in me. I would like to thank my friends Maciek, Krzysiek, Agata and Jakub for their support. Most importantly, I would like to show my gratitude to my beloved fiancée Alicja, for her unfailing love, patience, and encouragement throughout my work. Without her encouragement and understanding it would have been impossible for me to finish this work.

Contents

List of Tables

List of Figures

List of Diagrams

List of Abbreviations

AIC	**A**kaike **I**nformation **C**riterion
AP	**A**dvertisements **P**opulation
API	**A**pplication **P**rogramming **I**nterface
APPOR	**A**merican **A**ssociation for **P**ublic **O**pinion **R**esearch
ARB	**A**bsolute **R**elative **B**ias
ARIMA	**A**uto**R**egressive **I**ntegrated **M**oving **A**verage
BD	**B**ig **D**ata
BIC	**B**ayesian **I**nformation **C**riterion
BLUE-ETS	BLUE – **E**nterprise and **T**rade **S**tatistics Project (`www.blue-ets.eu`)
BP	**B**rokers **P**opulation
BPP	The **B**illion **P**rice **P**roject
CBS	**C**entraal **B**ureau voor de **S**tatistik (ang. *Statistics Netherlands*)
CCI	**C**ustomer **C**onfidence **I**ndex
CPI	**C**onsumer **P**rice **I**ndex
CSO	**C**entral **S**tatistical **O**ffice in Poland (pol. *Główny Urząd Statystyczny*)
EU-SILC	The **E**uropean **U**nion **S**tatistics on **I**ncome and **L**iving **S**onditions
GPS	**G**lobal **P**ositioning **S**ystem
GREG	**G**eneralized **REG**ression Estimators
GT	**G**oogle **T**rends
GTF	**G**oogle **F**lu **T**rends
HTTP	**H**yper**T**ext **T**ransfer **P**rotocol
ICT	**I**nformation and **C**ommunication **T**echnologies
IDS	**I**nternet **D**ata **S**ource
IDSP	**I**nternet **D**ata **S**ource **P**opulation
IDSs	**I**nternet **D**ata **S**ources
ILI	**I**nfluenza-**L**ike **I**llness
ILO	The **I**nternational **L**abour **O**rganization
IoT	**I**nternet **o**f **T**hings
IP	**I**nternet **P**opulation
IREIS	The **I**ntegrated **R**eal **E**state **I**nformation **S**ystem
ISI	**I**nternational **S**atistical **I**nstitute
IT	**I**nformation **T**echnology
JSON	**J**ava**S**cript **O**bject **N**otation
LAU	**L**ocal **A**dministrative **U**nit
LFS	**L**abour **F**orce **S**urvey
MAR	**M**issing **A**t **R**andom
MCAR	**M**issing **C**ompletly **A**t **R**andom
MIT	The **M**assachusetts **I**nstitute of **T**echnology
MNAR	**M**issing **N**ot **A**t **R**andom
MPG	**M**egapanel **PBI**/**G**emius survey

MS	**M**obile **S**ensors
NACE	Statistical Classification of Economic Activities in the European Community
NBP	**N**ational **B**ank of **P**oland (pol. *Narodowy Bank Polski*)
NDSs	**N**ew **D**ata **S**ources
NSCI	The **N**ational **S**tatistics and **C**ensuses **I**nstitute
NSI	**N**ational **S**tatistical **I**nstitution
NSIs	**N**ational **S**tatistical **I**nstitutions
NSP	The National Population and Housing Census 2011
NTTS	**N**ew **T**echniques and **T**echnologies for **S**tatistics
NUTS	**N**omenclature of **T**erritorial **U**nits for **S**tatistics
OECD	**O**rganization for **E**conomic **C**o-operation and **D**evelopment
PESEL	Universal Electronic System for Registration of the Population (register)
REGON	The **N**ational **O**fficial **B**usiness **R**egister
REST	**RE**presentational **S**tate **T**ransfer
RB	**R**elative **B**ias
RP	**R**egister **P**opulation
SAE	**S**mall **A**rea **E**stimation
SAIPE	**S**mall **A**rea **I**ncome & **P**overty **E**stimates
SD	**S**canner **D**ata
SM	**S**tatistical **M**atching
SOAP	**S**imple **O**bject **A**ccess **P**rotocol
TERYT	National Official Register of the Territorial Division of the Country
TP	**T**arget **P**opulation
UNECE	**U**nited **N**ations **E**conomic **C**ommission for **E**urope
VIN	**V**ehicle **I**dentification **N**umbers
WWW	**W**orld **W**ide **W**eb
XML	E**X**tensible **M**arkup **L**anguage

List of Symbols

$\boldsymbol{\Omega}_{TP}$	Target population (size N_{TP})
$\boldsymbol{\Omega}_{IP}$	Internet population (size N_{IP})
$\boldsymbol{\Omega}_{IDSsP}$	Internet data sources population (size N_{IDSsP})
$\boldsymbol{\Omega}_{IDSs}$	Observed Internet data sources population (size N_{IDSs})
$\boldsymbol{\Omega}_{AD}$	Advertisements population (size N_{AD})
$\boldsymbol{\Omega}_{BP}$	Brokers population (size N_{BP})
$\boldsymbol{\Omega}_{RP}$	Register population (size N_{RP})
$\boldsymbol{\Omega}_{TP,t,d}$	Target population in period t for domain d (size $N_{TP,t,d}$)
s_{TP}	Sample of $\boldsymbol{\Omega}_{TP}$ (size n_{IDSs})
s_{IDSs}	Observed sample of $\boldsymbol{\Omega}_{IDSs}$ (size n_{IDSs})
r_{IDSs}	Statistical units for which target variable is not missing (size m_{IDSs})
t	Denotes time period $t = \{1, ..., T\}$, where T is the total number of periods
d	Denotes domain $d = \{1, ..., D\}$, where D is the total number of domains
k	Denotes data source $k = \{1, ..., K\}$, where K is the total number of data sources
I_i	Indicator variable for sample inclusion for unit i
π_i	Probability of inclusion of unit i into sample
R_i	Indicator variable for response for unit i
ρ_i	Response propensity of unit i
y	The target variable
v	Proxy variable that has similar definition to y in register system
θ	Estimated characteristic of y, v (e.g. means, fractions, quantiles)
$\theta_{t,d}$	Estimated characteristic of y, v in period t for domain d
$\hat{\theta}$	Estimator of θ based on sample survey
$\tilde{\theta}$	Estimator of θ based on register-based survey
$\breve{\theta}$	Estimator of θ based on IDS-based survey
$\boldsymbol{x}$	Matrix of auxiliary variables
$\boldsymbol{X}$	Vector of known population totals of $\boldsymbol{x}$
$\boldsymbol{z}$	Matrix of other auxiliary variables(e.g. sampling scheme, paradata)
$\boldsymbol{w}$	Vector of initial sampling weights
$\boldsymbol{d}$	Vector of calibrated sampling weights
$\boldsymbol{\beta}$	Vector of fixed-effects parameters
$\tilde{\boldsymbol{\beta}}$	BLUE estimator of $\boldsymbol{\beta}$
$\hat{\boldsymbol{\beta}}$	EBLUP estimator of $\boldsymbol{\beta}$
$cor, cor(.)$	Correlation coefficient
sd_y	Standard deviation of y
f	Probability density function (PDF)
f_s	Sample probability density function (PDF)
f_Ω	Population probability density function (PDF)
d_{ijk}	Distance between i-th and j-th unit for k-th variable
$\boldsymbol{D}$	Distance matrix ($\{d_{ij}\}$)

L_i, L_j	Linking indicator variable
D	Dissimilarity index or total variation distance
O	Overlap between two distributions
B	Bhattacharyya coefficient
d_H	Hellinger's distance
df	Degrees of freedom
$\tau_{IDS,REG}$	The offer-to-transaction price index ($Q_{y,IDS,\alpha}/Q_{y,REG,\alpha}$)
$Q_{y,IDS,\alpha}$	A quantile (here percentile) for variable y based on IDS
$Q_{y,REG,\alpha}$	A quantile for variable y based on the register data
α	The probability level with values between $[0, 1]$
η	Direct estimate of $Bias(\breve{\theta})$
$\widetilde{\eta}$	EBLUP estimate of $Bias(\breve{\theta})$
ψ	Known sampling variace of $\hat{\theta}$
ξ	Known sampling variace of $\tilde{\theta}$
ϕ	Known sampling variace of $\breve{\theta}$
ω	Known sampling variace used for the model
$\delta_{k,d}$	First differences (e.g. $\tilde{\theta}_{k,d,t} - \tilde{\theta}_{k,d,t-1}$)
$\boldsymbol{Z}_r$	Matrix of r random-effect variables
$\boldsymbol{u}_r$	Vector of r random-effect parameters
σ_r^2	Variance of r random-effect
$\tilde{\rho}$	Autocorrelation coefficient
$\boldsymbol{\Omega}(\tilde{\rho})$	Toeplitz matrix reflecting autocorrelation structure AR(1)
$\boldsymbol{\Delta}$	Vector of parameters estimated in mixed models
r_{cond}	Conditional residuals
r_{cond}^{stand}	Standardized conditional residuals
$\mathbf{1}$	Vector of 1s
$\boldsymbol{I}$	Identity matrix
$\boldsymbol{J}$	Matrix of 1s ($\mathbf{1}\mathbf{1}'$)
$diag(.)$	Matrix diagonal
$col(.)$	Columnn stacked matrices of vectors
$\times$	Multiplication symbol
$\otimes$	Kronecker product
$\oplus$	Direct product
$\prime$	Transposition symbol
$exp(.)$	Exponential $e^{\cdot}$
R	Number of repliaces

For Alicja, my family and friends

Introduction

The research problem

New data sources, particularly the Internet and big data, are described in the literature on computer science, marketing and social science (Abramowicz, 2013; Lazer et al., 2014; Lynch, 2008; Miller, 2011; Varian, 2014). The Internet also plays an important role in the real estate market, especially in the process of buying and selling residential properties (Strączkowski, 2011) or predicting real estate market indicators (Wu and Deng, 2015; Wu and Brynjolfsson, 2014). However, the suitability of these sources for statistics is not widely discussed in the literature. The non-probabilistic character and complexity of these data is often neglected in consideration of their size. The survey methodology attaches particular importance to the quality of data sources. Hence, before Internet data sources can be used for statistical purposes, especially for official statistics, they need to be assessed in terms of such fundamental issues as representativeness, non-sampling error or bias.

Justification for the research topic

Growing 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 at National Statistical Institutes (NSIs) started using and adopting administrative records as part of their statistical systems (Wallgren and Wallgren, 2014). The use of administrative sources not only significantly reduces survey costs but also usually offers a better coverage of the population and can provide more accurate and timely statistics (Wallgren and Wallgren, 2014). However, the statistical theory underlying the use of administrative registers is currently the subject of research and development (Zhang, 2011, 2012b). Nonetheless, the process has brought 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 data, their scope is usually limited to a specific field that is relevant for their administrators. Initially, registers were not created for statistical purposes, which means that they 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, scanner data), administrative registers as well as surveys tend to lag behind the changing setting (Citro, 2014; Daas et al., 2015). Therefore, information gaps in certain fields are growing and new data sources should be examined to improve information coverage.

New data sources have gained recognition as a potential source of statistical information (Citro, 2014; Daas et al., 2015; Pfeffermann, 2015; Szreder, 2015). Moreover, the Internet not only generates a great deal of what is termed big data, but also provides ordinary-size data in a more accessible way – for example, access to public opinion polls or to local property records (Citro, 2014). However, these sources are not created by statisticians or for statistical purposes. These issues make new data sources similar to register data. It should be noted that statistical information differs from other types of information; in particular, it is characterised by relevance, accuracy (of estimation), timeliness, accessibility and clarity, comparability (in time and space) and coherence. Moreover, representativeness and non-sampling errors are key factors that can introduce bias. Internet data sources (IDSs) and big data are still not recognized and their suitability as statistical sources is often unknown. Therefore, there is a gap in statistical science, particularly in survey methodology.

Moreover, the Internet plays an important role in the housing market, as a source of information for potential buyers, for price and demand forecasting (Lee and Mori, 2015; Strączkowski, 2011; Wu and Deng, 2015; Wu and Brynjolfsson, 2014). For instance, Statistics Netherlands uses Funda.nl as a data source about the real estate market for statistics (Hoekstra et al., 2012). However, there is no research concerning the suitability of advertisement web services, or generally IDSs for real estate market statistics.

The relevance of the research topic and research gaps

So far official statistics has mainly relied on information collected from censuses and sample surveys, which are used to provide certain characteristics of the target population. Censuses are expensive survey tools, which are conducted every 10 years; the resulting information is often out-of-date due to delays between censuses and released results. On the other hand, sample surveys are constructed to reach a representative group of the target population and are focused on selected socio-economic phenomena.

Nowadays, administrative sources, such as registers, are becoming increasingly important as statistical data sources. The main characteristic that makes registers different from traditional sources mentioned above is that they were not created for statistical purposes and statisticians were not involved in the process. The extensive use of registers in Nordic countries (e.g. Finland, Norway, Denmark) and the Netherlands has demonstrated great opportunities of using registers as data sources for statistics. Poland has also got some experiences of using administrative data. The last National Census of Population and Housing in 2011 was conducted using a mixed approach. Data from administrative registers were used and, additionally, a 20% sample survey was conducted. Using registers for statistical purposes is connected with a number of problems – for one thing, registers are not created to provide statistics, nor are concept definitions coherent with

those used in official statistics. Therefore extensive empirical research has been conducted and theoretical foundations have been proposed to account for specific characteristics of registers (Wallgren and Wallgren, 2014; Zhang, 2011, 2012b).

However, in addition to the many advantages (low costs, no further respondent burden, information on small areas), the use of administrative records is also fraught with problems. The main one is the purpose of creating a register, which is not always consistent with the objectives of statistics; also the content scope of a register may reflect just a narrow segment of socio-economic life. For example, there is no register of households, which is the basic unit in most studies conducted by statistical offices. Another problem of using administrative records is related to estimation and inference. In this respect Zhang (2011) observes that the conceptualization and measurement of precision is one of the major challenges facing the use of administrative source. In addition, Zhang (2011) emphasizes that the use of rigorous statistical concepts, known from sample surveys, such as bias, variance, efficiency and consistency, is crucial in order to assess the benefits and risks faced by the use of new sources of information for statistics. These concepts are new and have only started to appear in the literature on estimation theory Holt (2007) and Zhang (2011, 2012b).

Statisticians and statistical offices which are using administrative records extensively have indicated problems related to insufficient information coverage. This has started the search for new sources. Statisticians directed their attention to other available data sources, such as the Internet or other sources that are created outside the statistical system and are not defined by the legislation. On the one hand, the selection of IDSs is the result of the increasing access to the Internet in developed and developing countries; on the other hand, it is due to the development of electronic economy. New data sources (NDSs) are often discussed in the context of big data, which is not a statistical term, although it is widely used in business as a potential source of new profits. It should be noted that, like administrative records, IDSs and big data are created for purposes other than statistical, mainly for business goals. New data sources can be discussed as a way of developing or replacing current modes of data collection, enhancing existing or creating previously not available statistics, while reducing the respondent burden, survey costs and improving timeliness. The use of these data sources for statistical purposes requires their critical assessment to identify benefits and risks, particularly in the context of estimation theory, as discussed by Zhang (2011, 2012b) with respect to administrative sources.

Internet data sources and big data are present in the literature devoted to information systems and technology, e-commerce or sociology (cf. Abramowicz, 2013; Abramowicz et al., 2002; Miller, 2011). However, the quality and representativeness of these data are often unverified. Unfortunately, only a few papers devoted to statistical theory point out methodological problems related to the use of IDSs as sources of statistical information, particularly in the context of the estimation theory. The first attempts to identify methodological problems regarding estimation and sampling appeared in Shmueli et al. (2005) in a study of online auctions on eBay and Bapna et al. (2006) with respect to e-commerce. Recently, most studies of IDSs and big data come from statistical offices, mainly Statistics Netherlands (Buelens et al., 2014; Daas et al., 2012; Daas et al., 2011; Hoekstra

et al., 2012). New data sources are beginning to be perceived as a potential way of enhancing existing or new statistics. Papers refer to various aspects of big data use, pointing to the possibility of extending the existing sources and research, replacing existing methods of data collection and creating new statistics (Lazer et al., 2014; Miller, 2011). Examples include the study of prices of products and services via the Internet, primary and secondary property market research, measuring the representativeness of social websites (Facebook, Twitter) and the possibility of constructing sentiment indexes. It should also be noted that the use of Internet data sources may pose a potential competition for official statistics, i.e. *The Billion Price Project* (Cavallo, 2012, 2013). Baker et al. (2013) provides a summary of the American Association for Public Opinion Research (APPOR) report on non-probability samples and suitable statistical methods. The APPOR review of the current approach to big data for public opinion research is included in Japec et al. (2015). Additionally, it should be noted that the term representativeness was used mainly with respect to sample surveys (Schouten et al., 2009), but recently it has also been used in the context of registers (Ouwehand and Schouten, 2014). Therefore, a detailed methodological framework is required to specify when and how representativeness of new data sources can be measured.

In the context of the estimation theory, new data sources can be described as self-selected / non-probability samples. Classical methods presented in survey methodology may not be suitable to handle such data. Therefore, other methods should be considered. The statistical literature on unit non-response and selection errors can be classified into two groups: (1) calibration and modelling self-selection to obtain weights used for estimation (Brick, 2013; Chen, Kim, et al., 2014; Kim and Riddles, 2012; Lee, 2006; Lehtonen and Veijanen, 2009; Rosenbaum and Rubin, 1983; Särndal and Lundström, 2005; Szreder, 2007; Szymkowiak, 2007), (2) model-based estimation using sample data or a combination of sample and register/census data (Gelman, 2007; Ghosh and Rao, 1994; Lehtonen and Veijanen, 1998; Pfeffermann, 2002, 2013; Rao and Molina, 2015; Rao, 2003). The model-based approach is widely used in small area estimation, where two possible approaches are identified depending on the level of available data – unit and area models (Rao and Molina, 2015). These models are used to provide more accurate estimates where the direct estimator is unreliable. Another approach is model calibration, which is a combination of the two groups mentioned above (Lehtonen and Vejanen, 2012). The model-based approach is also used for handling non-probability samples (Baker et al., 2013; Szreder, 2010). For instance, Wang et al. (2015) propose the Bayesian approach and a stratification model to predict results of elections on the basis of non-probability samples.

Nonetheless, the above mentioned methods have not been discussed in the context of IDSs, where the selection mechanism is yet unknown. In addition, the self-selection process can be correlated with the target variable, which makes it MNAR. This issue is discussed in the context of complex surveys (Kim and Riddles, 2012; Kim and Skinner, 2013; Pfeffermann and Sverchkov, 2007; Rubin, 1976), but rarely addressed with respect to IDSs. Therefore, the present dissertation is devoted to the evaluation of representativeness, non-sampling errors and bias and its sources in IDSs using data from the secondary real estate market.

The research objectives and hypotheses

The main goal of the dissertation is *to evaluate the Internet as a data source for real estate market statistics*. To achieve this aim the following specific objectives have been defined:

G1 Identification of non-sampling error in Internet data sources about the secondary real estate market.

G2 Assessment of the representativeness of Internet data sources for the secondary real estate market.

G3 Assessment of bias in Internet data sources for the secondary real estate market.

To achieve these goals, the following hypotheses have been formulated. The main hypothesis verified in the dissertation is: *Internet data sources enable acceptable estimation of real estate market characteristics*. The specific hypotheses are:

H1 The approach proposed by the author to measure representativeness can be effectively used to assess the representativeness of Internet data sources about the secondary real estate market.

H2 Internet data sources are biased and this bias varies between sources and domains.

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

H4 Internet data sources can be used to estimate the offer price per m^2 in the secondary real estate market with acceptable error measured by absolute relative bias.

The dissertation structure

The structure of the dissertation has been designed to support its goals and provide the basis for the verification of the hypotheses. The first chapter is devoted to a comprehensive literature review and a summary of recent work on new data sources for official statistics. The review covers the historical background of statistical data sources and developments in survey methodology. The review of the literature indicates that there is no single and clear definition of an Internet data source (IDS) or big data. Therefore, the following definition of IDS is proposed: *an Internet data source is a self-selected (non-probabilistic) sample, which is created through the Internet and maintained by entities external to NSIs and administrative regulations*. For purposes of the dissertation this definition of the Internet is limited to the World Wide Web. The chapter continues with a review of current experiences in the use of the Internet as a data source. The chapter ends with a summary of challenges and risks involved in the use of IDSs for official statistics.

The second chapter focuses on Internet data sources in the real estate market. First of all, basic statistical concepts, such as population, statistical unit and target variable are applied to IDSs. Furthermore, existing data sources used for official statistics are compared with IDSs. Possible errors are identified by applying a two-phase life cycle of integrated statistical micro data proposed by Zhang (2012b) to IDSs. As a result, a number of errors are identified, which are present in IDSs; these errors are then exemplified by referring to the real estate market in Poland. The presentation and discussion of statistical and non-statistical data sources in the secondary real estate market in Poland. The second chapter concludes a list of the key points that need to be considered when selecting IDSs, integrating IDSs with the statistical system and linking them with existing data sources.

The third chapter discusses possible approaches to measuring the representativeness of IDSs. First, various concepts of representativeness are presented based on Bethlehem (2009) and Kruskal and Mosteller (1979a,b,c) and are discussed in the context of IDSs. Drawing on this theoretical basis the author proposes his own procedure to measure the representativeness of IDSs. The proposed approach considers several cases where official (survey, census, reporting, statistical registers) and non-official (administrative) sources are available. The approach takes into account data linkage at unit and domain level. The chapter ends with a summary of possible measures of representativeness that can be used to assess IDSs. These measures are classified into two groups based on the availability of data.

The approach proposed in the third chapter is used to conduct an empirical assessment of representativeness, which is described in the fourth chapter. The study begins with the assessment of the overall suitability of IDSs as a data source about the real estate market. This part involves assessing the use of the Internet by companies classified into the category of businesses conducting Real Estate Market Activities. Representativeness is verified by examining the distribution of the number of rooms and floor area in three selected IDSs and a survey conducted by the National Bank of Poland and the Central Statistical Office. The results are compared with data from the Register of Real Estate Values and Prices for Poznań to identify differences in the distribution of the offer and transaction price. The results presented in the fourth chapter reveal substantial differences between IDSs, register and survey data.

The fifth chapter provides the theoretical basis for the measurement of bias in IDSs and the empirical assessment of bias observed in IDSs with respect to the average price per m^2. The author extends the approach proposed by Fosen and Zhang (2011) and Zhang (2012a) in order to apply it to IDSs. It is assumed that the official data are unbiased: they are treated as a *gold standard*. The author then moves on to propose a new small area estimator to estimate bias in IDSs, which takes into account four random effects and uncertainty in both sources. The proposed model makes it possible to assess the impact of the domain (city) and data source effect on bias. The theoretical considerations are followed by an empirical assessment of bias in the offer price per m^2 in 12 cities for 12 quarters. The NBP/CSO survey of real estate brokers, which provides statistics about the secondary real estate market, was used as a reference. The results are then evaluated.

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

Data sources

The selection of data sources, domains and the reference time period was dictated by the availability of data. The Central Statistical Office (CSO) and the National Bank of Poland (NBP) conduct a survey entitled *Residential and commercial property prices* that covers, among other areas, the secondary real estate market. Estimates of offer prices based on the NBP/CSO survey were used as reference official statistics. The evaluation of the representativeness of IDSs was based on data from three online advertising services devoted to the real estate market – Dom.Gratka.pl, OtoDom.pl and Nieruchomosci-online.pl. A detailed description of the selection process is described in Chapter 2 and Chapter 3. Current and historical data from Nieruchomosci-online.pl were obtained using a web-scraping technique, while data from Dom.Gratka.pl and OtoDom.pl were obtained directly from the owners in an aggregated form. There were differences between domains and time series covered by these sources, so to enable comparison, the scope of data was limited to 12 domains (the biggest cities in Poland, Białystok, Gdańsk, Katowice, Kraków, Łódź, Lublin, Olsztyn, Opole, Poznań, Szczecin, Warszawa, Wrocław) and and the period from first quarter 2012 to the fourth quarter 2014 (2012Q1 to 2014Q4). In addition, data from the Register of Real Estate Values and Prices were used in the fourth chapter. However, due to the lack of access to data from different cities the comparison was limited to Poznań.

Chapter 1

New data sources for statistics

1.1 Development of data sources for statistics

Until recently traditional data sources for statistics have been censuses, whose aim is to enumerate all units of the target population at a particular time and geographical location. Such practices can be dated back to Babylon (~ 3800 BC), the ancient Egypt (~ 3000 BC), the Roman Empire (~ 600 BC) and China (2 BC). The growing need for information about the state and society has led to an increase in the use of this method of data collection. The importance of this method can also be observed in the 21st century, with some countries continuing to collect data through full enumeration. Despite their undoubted importance, censuses are expensive to organize and conduct. Data collection and data processing is often highly time-consuming. Censuses are not error-free (e.g. undercoverage, item or unit non-response, which causes problems related to estimation. These problems inspired the search for new solutions that could improve or replace the traditional approach. Sources created by other entities were used, for instance parish records or bills of mortality. Another approach was to obtain information on the population of interest only by analysing a sample. This led to the development of the representative method, which has changed statistics and has become one of the most frequently used methods of obtaining statistical information (Bethlehem, 2009, ch. 1.3).

The first mention of the representative method was made by H. Kiær in 1895 at the meeting of the International Statistical Institute (ISI) (Kiaer, 1897). Zhang (2011) cited Jensen (1924) "Although (Kiær) he was unable to defend the approach on theoretical grounds, the practice continued to evolve thereafter and in 1924 the ISI formed a committee to investigate the application of the representative method". Its report stated: "When ISI discussed the matter twenty-two years ago, it was the question of the recognition of the method in principle that claimed most interest. Now it is otherwise. I think I may venture to say that nowadays there is hardly one statistician, who in principle will contest the legitimacy of the representative method. Nevertheless, I believe that the representative method is capable of being used to a much greater extent than now is the case" (Jensen, 1924; Zhang, 2011). Indeed, the theoretical breakthroughs did not arrive until some 30–40 years after Kiær's initial presentation (Zhang, 2011). The theoretical basis of the representative method had not been developed until 1934 and the pioneering work "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection" of Jerzy Spława Neyman (Neyman, 1934). In the statistical literature, his work is regarded

as the beginning of the new branch of statistics – survey methodology (and survey sampling). The Neyman approach is still widely used in the world, and its comprehensive description can be found, for example, in Bethlehem (2009), Groves et al. (2011), and Lohr (2009). The development of estimation theory using sample surveys began with an approach based on a sampling scheme and an estimator proposed by Hansen and Hurwitz (1943) and Horvitz and Thompson (1952). Groves (2011b) also refers to this period as *the Era of Innovation*, which helped to establish survey methodology as the most important method of obtaining information about society and economy.

The modern use of administrative records dates back to 1970, when Norway and Finland used administrative data to support the census for the first time. In 1981 Denmark used all available registers, creating a new type of census - a virtual census. The new approach is characterized by the use of existing data sources (primarily administrative) to estimate characteristics of the target population. It has changed the paradigm of statistics, marking a transition from relying on censuses and sample surveys that are planned and conducted by statisticians to using existing and available data sources, which have been created for other purposes than statistics. A comprehensive description of the use of registers in the Nordic countries can be found in Statistics Finland (2004), UNECE (2007) and Wallgren and Wallgren (2014). A growing interest in the use of administrative records to replace the traditional census can be observed at present (see also Baffour et al., 2013; Coleman, 2013). In addition, registers are used to construct sampling frames for surveys and as a source of auxiliary variables for the model-based approach. Administrative sources also became the main data source for information about enterprises (see for example the BLUE-ETS project on Enterprise and Trade Statistics[1]).

Despite its many advantages, the use of administrative records is also fraught with problems. The main one is the purpose of creating a register, which is not (always) consistent with the objectives of statistics; also the content scope of a register may reflect just a narrow segment of socio-economic problem of interest. For example, there is no register of households, which is the basic unit in most studies conducted by statistical offices. Another problem is that register holders tend to focus on variables that are crucial for administrative purposes, hence obligatory. As a result, other information is not obligatory, which results in missing data (item or unit missing values). A further problem of using administrative records is related to the possibility of estimation and inference. For example, the register of real estate prices and values contains information on transactions concerning land and buildings in Poland. Transactions might involve one or several properties. Therefore, there is a problem of converting records of transactions to records of units and then to statistical units. Zhang (2011) observes that the conceptualization and measurement of precision is one of the major challenges facing the use of administrative sources. In addition, Zhang (2012b) emphasizes that the use of rigorous statistical concepts, known from sample surveys, such as bias, variance, efficiency and consistency, is crucial in order to assess the benefits and risks involved in the use of administrative sources of information for statistics. These

[1]For more information on BLUE-ETS project on Enterprise and Trade Statistics. See `http://www.blue-ets.istat.it/index.php?id=57`.

concepts are new and have only started to appear in the literature on estimation theory (Holt, 2007; Zhang, 2011, 2012b).

However, problems with estimation, in particular in the case of small domains, have brought about a change in the direction of model-assisted (Särndal et al., 2003) and model-based estimation (Rao, 2003). The first approach involves using information from the sample or known population totals (e.g. GREG estimator), while the second approach uses auxiliary information for domains of interest outside the sample (e.g. registers) or from previous surveys. Fay and Herriot (1979) and Ghosh and Rao (1994) might be regarded as the beginning of a new branch of statistics – small area estimation (SAE). A comprehensive summary and review can be found in Pfeffermann (2002, 2013), Rao et al. (2011), and Rao (2003). What makes this approach innovative is the use of auxiliary information from previous surveys or outside the sample for the domain of interest. As a result, this approach reduces estimator variance and might provide information for out-of-sample domains (often called *borrowing strength over time or space*) (Ghosh and Rao, 1994; Rao et al., 2011). Sources used in SAE are mainly external to a given survey – censuses or administrative sources (registers) – because of assumed full population coverage (hence lack of sampling error). SAE is used by National Statistical Institutions (NSIs) that conduct sample surveys, primarily to obtain more precise estimates in small domains and reduce costs of surveys. For instance, in the USA Local Area Unemployment Statistics are provided based on structural models for over 7300 unique geographic areas (Bureau of Labour Statistics, 2015), Statistics Netherlands provide monthly estimates of unemployment based on Dutch LFS and SAE for 437 municipalities broken down by sex and age Boonstra and Buelens (2011) and Statistics Netherlands (2015) or Census Bureau lead Model-Based Small Area Income & Poverty Estimates (SAIPE) for School Districts, Counties and States that provide income and poverty statistics (Census Bureau, 2015). On the other hand, researchers and NSIs (in particular from Nordic countries) have noticed that administrative sources can be used for the purpose of deriving statistical information instead of sample surveys. This can be seen as the start of the next era or a paradigm change in statistics (Gołata, 2014; Groves, 2011b; Zhang, 2012b).

Finally, statisticians and NSIs which are using administrative records have indicated problems related to insufficient information coverage (Citro, 2014; Wallgren and Wallgren, 2014). This has given rise to the search for new information sources, with statisticians directing their attention to the Internet or other sources (e.g. scanner data, mobile phone data) that are created outside the statistical system and are not defined in the law and the legislation. On the one hand, the use of Internet data sources (IDS) is the result of the increasing access to the Internet in developed and developing countries; on the other hand, thanks to the development of e-economy and activities of people and entities it can largely be observed through the web. New data sources (NDS) are also discussed in the context of *big data* (BD), which is not a statistical term, though it is widely used in business and marketing as a potential source of information and profits. It should be noted that IDS and BD are created for purposes other than statistical (like administrative records), mainly for business goals. These new data sources are discussed mainly

in the context of enhancing, or even replacing and creating new statistics, thus reducing the respondent burden, survey costs and the time between data collection and the release of results. However, the use of these data sources for statistical purposes requires their critical assessment to identify benefits and risks, in particular in the context of estimation theory, as discussed by Zhang (2012b). Therefore, the following chapter will focus on introducing the statistical background and statistical evaluation of new data sources.

1.2 New data sources

Recently, most of new data sources have been classified as BD, which is not a statistical term. The most known definition proposed by Laney (2001) and then discussed by Bayer (2011) and Bayer and Laney (2012) lists three characteristics *3V* – (high) volume, (high) velocity and (high) variety[2]. The first *V* refers to the amount of data counted in giga-, tera- or petabytes, which are difficult to analyse within the existing infrastructure (increase in traffic and data creation is exponential). The second *V* refers to how these data are generated and changed in time. For example according to Intel's infographic[3] during one minute on the Internet nearly 350 000 tweets are posted on Twitter, 4.1 million searches are made in Google, 6.9 million messages are sent on Facebook, and Netflix and YouTube account for 1/2 of all traffic (together over 160 000 hours of video is watched in one minute on the Internet). The last *V* indicates that BD occur in different formats, such as text files (e.g. text, XML, JSON), photos (e.g. posted on Instagram or Facebook), web-site logs, videos (e.g. camera surveillance), recordings or geocoded data. In comparison to classical data sources, like censuses or surveys, BD processing requires more effort not only to "clean" the data but also to provide basic information about persons that generate data. We need to mention that some types of BD can occur in administrative sources, i.e. traffic sensor data, patients' registers or aerial and satellite photos. However, in most cases such data are associated with specific types of business and its customers – on-line services (e.g. Google, Facebook, Twitter), banking and finance (e.g. stock markets, global banking companies), shops (e.g. Wal-mart, Tesco) or telecommunication (e.g. AT&T, Verizon, T-Mobile, Orange).

BD is not a term used widely in official statistics systems. However, Eurostat[4] defines a statistical data source as "*data collected exclusively for statistical purposes, including data from statistical registers, statistical surveys (including censuses), and other official statistical agencies, as well as data from combinations of these sources*" and non-statistical data source as "*data not primarily collected for statistical purposes, including data from administrative sources, commercial sources and automatic monitoring and recording devices*". In the light of these

[2]There are other definitions which include extra two V's – Veracity and Value.

[3]See http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html.

[4]See http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=DSP_GLOSSARY_NOM_DTL_VIEW&StrNom=CODED2&StrLanguageCode=EN&IntKey=21239003&RdoSearch=CONTAIN&TxtSearch=source&CboTheme=&IsTer=&IntCurrentPage=1&ter_valid=21862.

definitions BD is classified as a non-statistical data source. On the other hand, the United Nations Economic Commission for Europe (UNECE) defines administrative sources as *sources containing information which is not primarily collected for statistical purposes*[5] (narrow definition) or *data collected by sources external to statistical offices*[6] (wider definition), whereas Eurostat clearly states what administrative data sources are *A data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations* [7]. The narrow definition of administrative sources includes only the public sector, while a wider definition includes the private sector as well. The wider definition, in contrast to Eurostat's, can be seen to include BD sources as administrative records, albeit the narrow one classifies BD as a secondary (non-statistical source) (see UNECE, 2011, p. 2).

However, in 2014 UNECE launched the "BD and Official Statistics" project. One of its aims was to bridge the definition gap in the UNECE glossary by explicitly defining BD as *data that are difficult to collect, store or process within the conventional systems of statistical organizations. Either, their volume, velocity, structure or variety requires the adoption of new statistical software processing techniques and/or IT infrastructure to enable cost-effective insights to be made* (UNECE, 2014). This definition mainly refers to the characteristics of these data sources and does not specify the purpose of creating these data or identify the holder. UNECE (2014) also proposed a classification of BD into three main groups which treats the data creation process as a factor for distinguishing sources of information:

- *Human-sourced information* (social networks) – this information is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video. Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. Data are loosely structured and often ungoverned: social networks (Facebook, Twitter), blogs, comments, private pictures, Internet searches, text messages, email, user-generated maps;

- *Process-mediated information* (traditional business systems) – these processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected is highly structured and includes transactions, reference tables and relationships, as well as the metadata that sets its context. Traditional business data is the vast majority of what IT managed and processed, in both operational and business intelligence systems. Usually

[5] See: http://www1.unece.org/stat/platform/display/adso/Administrative+Sources.

[6] See: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=75564299.

[7] See: http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=DSP_GLOSSARY_NOM_DTL_VIEW&StrNom=CODED2&StrLanguageCode=EN&IntKey=20159574&RdoSearch=CONTAIN&TxtSearch=source&CboTheme=&IsTer=&IntCurrentPage=1&ter_valid=0.

structured and stored in relational database systems. (Some sources belonging to this class may fall into the category of "Administrative data") : medical records, commercial transactions, banking/stock records, e-commerce, credit cards;

- *Machine-generated information* (Internet of Things) – derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches.: fixed sensors (weather, traffic, surveillance), mobile sensors (tracking, GPS, mobile phones), and data from computer systems – logs and web logs.

The first definition of BD which takes into account the statistical characteristics of these data was proposed by Horrigan (2013) (Associate Commissioner at Bureau of Labour Statistics USA) who regarded these data as *non-sampled data, characterized by the creation of databases from electronic sources whose primary purpose is something other than statistical inference*. Horrigan (2013) proposal included one of the most neglected, owing to the high number of observations, aspects of BD – a non-sampling character of these data.

Groves (2011a,b) classify statistical data sources into two groups: *designed* and *organic*. Designed data refer to concepts and sources designed in official statistics (for instance surveys or census). In contrast, *organic* data denote *data collectively assembled by society which reflect massive amounts of its behaviours and can be regarded as an ecosystem that is self-measuring in an increasingly broad scope*. The main difference between these two groups is that *designed* data are created by statisticians for statistical purposes. The term *design* refers to a designed-based statistics where a specific sample from the population is selected and is used to answer specific questions (that are consistent with the methodology and definitions used in official statistics). However, *organic* data proposed by Groves (2011a) are, in fact, `designed` albeit for business purposes. For example, Twitter allows users to write only 140 characters and offers a special character # (*hashtag*) to tag messages, which speeds up the search; Facebook classifies user information into groups that contain demographic, personal or activity data; Google corrects the spelling errors but also classifies search queries according to categories presented in Google Trends. In addition, services that offer support in sales have special forms to input information and characteristics of products or services. Therefore, *organic* data are becoming *designed* and structured in order to collect as much information on products and users as possible. Keller et al. (2012) followed Groves (2011a,b) provided examples of such data sources:

- *designed data*: administrative data (e.g. tax records), federal surveys, censuses of population and “other data collected to answer specific policy questions” (e.g. LFS);

- *organic data:* location data (cell phone "externals", E-ZPass transponders, surveillance cameras), political preferences (voter registration records, voting in primaries, political party contributions), commercial information (credit card transactions, property sales, online searches, radio-frequency identification), health information (electronic medical records, hospital admittances, devices to monitor vital signs, pharmacy sales) and other organic data (optical, infra-red and spectral imagery, meteorological measurements, seismic and acoustic measurements, biological and chemical ionizing radiation).

Recently, Constance Citro (director of the Committee on National Statistics at the National Academy of Sciences, USA) discussed BD and the Internet in the context of multiple data sources. Citro (2014) proposed a classification of existing data sources into four groups: census and probability surveys, administrative records, commercial transaction records and individual interactions with the Internet, giving the following examples:

- *Surveys and censuses*, or a collection of data obtained from responses of individuals, who are queried on one or more topics as designed by the data collector (statistical agency, other government agency, and academic or commercial survey organization) according to principles of survey research with the goal of producing generalizable information for a defined population.

- *Administrative records* or a collection of data obtained from forms designed by an administrative body according to law, regulation, or policy for operating a program, such as paying benefits to eligible recipients or meeting payroll. Administrative records are usually ongoing and may be operated by government agencies, or non-governmental organization.

- *Commercial transaction records*, or a collection of data obtained from electronic capture of purchases (e.g., groceries, real estate) initiated by a buyer but in a form determined by a seller (e.g., bar-coded product information and prices recorded by check-out scanners or records of product and price information for Web sales, such as Amazon).

- *Interactions of individuals with the World Wide Web* by using commercially provided tools, such as a Web browser or a social media site. This category covers a wide and ever-changing array of potential data sources for which there are no straightforward classifications. One defining characteristic is that individuals providing information, such as a Twitter post, act as autonomous agents: they are not asked to respond to a questionnaire or required to supply administrative information but, instead, are choosing to initiate an interaction.

It can be argued that the classification proposed by Citro (2014) consists of two main groups – public and commercial data. The first group encompasses surveys conducted by statistical agencies but also other government agencies and includes censuses. The classification is not restricted to statistical data but is more

general and includes non-statistical surveys that are "designed by the data collector", much like in the classification proposed by Groves (2011a,b). It also includes administrative and government data that are designed by the law, regulations or other policies. These data are also not designed by statisticians but by government agencies. To sum up, the common denominator is that these data are created to meet specific goals of the government. The second group consists of *commercial transaction records* and *interactions of individuals with Internet/WWW*. The common denominator for these data is the purpose of creation, which is not statistical. Moreover, Citro (2014) does not only focus on the size of new data sources but also points out that "*in regard to the Internet it not only generates a great deal of today's "BD", but also provides ordinary-size data in a more accessible way - for example, access to public opinion polls or to local property records.*". This comment should also be taken into account when considering data sources for statistics. Not only popular BD examples, such as social media or mobile phones, are of interest to statisticians but also medium-sized, such as on-line price comparisons or on-line real estate market advertisement services.

Finally, Nagler and Tucker (2015) discuss in a more detailed way (than Horrigan (2013)) how to treat BD or the Internet using the example of political studies. Because Nagler and Tucker (2015) regard BD as a sample, one needs to consider:

- whether it is representative of the population,
- whether the population from which the sample was drawn is a population of interest and/or whether the population it was drawn from generalizes to other populations,
- because BD is big, more caution needs to be exercised than is typically the case when interpreting measures of statistical precision, in particular variance (and sampling error).

Nagler and Tucker (2015) point out that even though a BD or the Internet data source may be a good predictor for a socio-economic phenomenon, it does not follow that it is representative of the target population.

In conclusion, there are several problems of defining BD or, more generally, new data sources from the perspective of official statistics. It can be argued that the use of the term *BD* is a proper way of describing emerging data sources for statistics. From the perspective of official statistics the size of data or variety of data types are not important; what really matters is the way these data originated and their non-probabilistic character. Therefore, it will be useful to provide a systematic review of the literature on the subject and concepts of new data sources for statistical purposes.

1.3 Experiences in the use of new data sources

As Internet access in households is increasing (Mohorko et al., 2013), as is the range of ways in which individuals and entities can communicate over the Web (i.e., social media, web sites, web forums, price comparison and advertisement

services). This process opens new opportunities for statisticians to track and measure the economy and social phenomena. For instance, 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 approach hard-to-reach populations. In the literature we can find many examples of research that involves several types of new data sources. In this section we focus on the literature review and consider two settings. We focus on non-official and official statistics research. As regards non-official statistics, we are interested in literature that addresses various aspects of using BD and IDS for socio-economic investigations, whereas in the area of official statistics we focus on literature relating to issues of survey methodology and inference.

New data sources, in particular BD, are widely discussed in computer science literature. However, this type of literature mainly covers topics related to hardware (e.g. data management, parallel processing, high performance computing) and machine learning (e.g. pattern and voice recognition, natural language processing), which are beyond the scope of this dissertation. Recently, BD has also become popular in socio-economic research (Einav and Levin, 2014; Japec et al., 2015; Lazer et al., 2014). However, statistical properties concerning BD and IDS problems are not widely discussed. Research presented in this section is mainly conducted by non-official statisticians, hence does not fully address the topics related to survey methodology.

The first reference, known to the author, that reflects the statistical perspective on the Internet data is Shmueli et al. (2005) and concerns on-line auction research. Shmueli et al. (2005) outlined the problem related to sampling and underlined that this issue appears to be completely disregarded in the current literature. These issues are raised once again in Jank and Shmueli (2010) on modelling on-line auctions. Bapna et al. (2006) discusses the problem of data-driven research in the context of e-commerce.

The first world-wide popular example that emphasised the possibility of using the Internet for statistics was Google Flu Trends (GFT). Ginsberg et al. (2008) presented the possibility of using Google search data to estimate current flu activity around the world and compared results for several countries. Statistics provided by U.S. Centres for Disease Control[8] and European Influenza Surveillance Network of the European Centre for Disease Prevention and Control[9] were used for reference. The basic idea presented in the paper was to use the number of search terms related to the flu that were entered in the Google search engine. Searches associated with keywords such as "flu" or "influenza" were used as possible indicators of the illness for over 25 countries. In addition, a linear model was used to compute the log-odds of an influenza-like illness (ILI) physician office visit and the log-odds of an ILI-related search query Ginsberg et al. (2008). However, in 2015 the publication of Google Flu Trends was deprecated and is no longer available[10].

[8] See http://www.cdc.gov.

[9] See http://ecdc.europa.eu/en/activities/surveillance/EISN/Pages/index.aspx.

[10] Only historical data are available online. For more information see https://www.google.org/flutrends/about/.

Figure 1.1 presents the estimated flue activity by GFT and flu reports for US and Poland. Flu search activity is based on aggregated Google Search query data, normalized to make data more comparable across regions. The "baseline" level for each region (shown as 0) is its average flu search activity, measured over many seasons. Activity levels for each region represent how much flu search activity differs from that region's "baseline" level. If two regions have similar flu search activity levels, it means those regions differed from their "baseline" activity levels by a comparable margin. The GFT up to 2013 was considered to be a good proxy of flu activity. In the 2013 GFT overestimated reported flu activity twice. However, the relation between Google Flu Trends and Influenza estimates for Poland are more similar and stable than for the US.

FIGURE 1.1: Google Flu Trends in the United States and Poland

Note: based on `https://www.google.org/flutrends/about/`

Results presented in Ginsberg et al. (2008) started a worldwide discussion whether BD, such as GFT, can be used instead of statistics provided by government agencies[11]. Lazer et al. (2014) discussed not only the GFT but also other search data and social media and summarised the on-going debate on query data as follows *instead of focusing on a "BD revolution", perhaps it is time we were focused on an "all data revolution", where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of our world.*

[11] Number of citations of Ginsberg et al. (2008) is 1850 according to Google Scholar (view on 18.01.2016).

However, lately researchers have mainly focused on predicting official indicators using new data sources, in particular Google Trends (GT) indexes that are regarded to be a proxy of socio-economic phenomena. The main field for the use of GT is to study the possibility of predicting the situation on various labour markets. Xu et al. (2012) presented a machine learning approach to predicting the unemployment rate in the US and used weekly GT categories "Local/Jobs" and "Society/Social Services/Welfare & Unemployment" between 2004 and 2012. Xu et al. (2012) provided a review of machine learning methods rather than explaining how GT can be assessed as a proxy for unemployment. Fondeur and Karamé (2013) focuses on nowcasting youth unemployment rates in France for the population aged 15-24, discussed the choice of the right keywords and categories from GT and applied the modified Kalman filter. His results indicate that including GT data improves unemployment predictions for this specific age group. Vicente et al. (2015) discusses a way of predicting monthly registered unemployment in Spain between 2004 and 2013. Vicente et al. (2015) used the ARIMA model with two variables. The first one was the Employment Confidence Indicator, which reflects the balance between positive and negative opinions of industrial firms about the current employment situation and their perspectives three months ahead. The second one was the GT index computed separately for "employment offers" (*oferta de empleo*) and "work offers" (*oferta de trabajo*). Results indicated that using GT based variables significantly improved the proposed ARIMA model.

Web searches, in particular GT indexes, are not only used for nowcasting unemployment but also for other purposes. For example, Goel et al. (2010) summarise current trends in using Web searches for predicting customer behaviour with the focus on commerce (video games, music and movies). Choi and Varian (2012) provide an overview of possible uses of GT for predicting retail, automotive and house sales. Wu and Brynjolfsson (2014) focus on predicting housing prices and sales at state level in the US. Scott and Varian (2014) presented a Bayesian structural time series for forecasting weekly initial claims for unemployment and monthly retail sales using the GT index. Vosen and Schmidt (2011) propose an indicator for private consumption based on GT in relation to two survey-based indicators (the University of Michigan Consumer Sentiment Index[12] and the Conference Board Consumer Confidence Index[13]). The results presented in Vosen and Schmidt (2011) indicate that the construction indicator based only on GT is a better predictor of consumption than the two above mentioned indicators (in the sense of predictive power and relative out-of-sample performance).

Finally, we review the literature devoted to general concepts of new data sources. Couper (2011) and Couper (2013) discuss the future of data collection from the perspective of a public opinion researcher. Couper (2011) underlined potential data sources that, in the future, could support or replace existing data collection. In addition, Couper (2013) addresses drawbacks connected with using new data sources, in particular BD, stressing the importance of representativeness,

[12]It should be noted that these indicators are not based on a probabilistic sample, see `http://www.sca.isr.umich.edu`.

[13]It should be noted that these indicators are not based on a probabilistic sample, see `https://www.conference-board.org/data/consumerconfidence.cfm`.

privacy, access, arguing that these data sources will replace existing opinion research. A report by Japec et al. (2015) assesses the usefulness of BD for public opinion research. The report is a comprehensive review of existing definitions, data sources, techniques, drawbacks and benefits of new data sources. Japec et al. (2015) concludes that surveys and BD are complementary, not competing data sources and suggests that the American Association For Public Opinion Research (APPOR) should take actions to inform about risks and benefits of using these data sources. Edelman (2012) presents possible ways of using the Internet and web-scraping for economic research. Varian (2014) discusses the possibility of research using BD sources and techniques that make it possible to manage these data efficiently. Antenucci et al. (2014) presents how social media (i.e. Twitter) can be used to measure labour market flows giving the example of The University of Michigan Social Media Job Loss Index, which was created for this purpose[14]. Ann Keller et al. (2012) discusses how BD can be used to gain knowledge about cities and provide examples of data and potential uses. Einav and Levin (2014) describe economic research in the BD era, noting the growing use of non–publicly available data in economic research.

The first example is *the Billion Price Project* (BPP)[15] conducted by the Massachusetts Institute of Technology (MIT). The aim of BPP is to estimate daily, monthly and quarterly Consumer Price Index (CPI) in over 60 countries. The BPP CPI is based on a web-scraped panel dataset that covers information on each product, including prices, product and category ID. Cavallo (2012) provides a detailed description of the data collection process (web-scraping) and the methodological issues of the comparison with offline prices. Figure 1.2 presents official and BPP based estimates of CPI. For the United States, BPP CPI, based only on Internet data, is coherent with the official indicators, while in the case of Argentina large discrepancies are visible. These differences were the main reason to conduct the research by Cavallo (2013) in an attempt to confirm the widespread suspicion that the government has been manipulating the CPI since January 2007, when it interfered with the work done by the National Statistics and Censuses Institute (NSCI). Cavallo (2013) argued that the differences were caused by the government intervention, in contrast to other countries in the region (Brazil, Chile, Columbia and Venezuela) where there were no such large discrepancies as in Argentina. Furthermore, Cavallo et al. (2014a) used BPP data to investigate currency unions and exchange rates, the impact of joining a currency union (Cavallo et al., 2014c) and other multinational comparisons (Cavallo et al., 2014b).

The BPP has also been discussed in US newspapers[16], in particular in reference to the official CPI calculated and published by the Census Bureau. Horrigan (2013) discusses problems with the CPI produced by BPP and points out the following problems: (1) BPP does not take into account the same representative bundle on a daily basis; (2) BPP does not include sampling weights derived from websites which it collect the prices from, and finally, (3) it lacks a clear methodology that underlies the calculated CPI. In addition, Horrigan (2013) also mentions scanner data collected by AC Nielsen that is also used for CPI calculation.

[14] See `http://econprediction.eecs.umich.edu`.

[15] For more details, see `http://bpp.mit.edu`.

[16] See the press articles on BPP, `http://bpp.mit.edu/press/`.

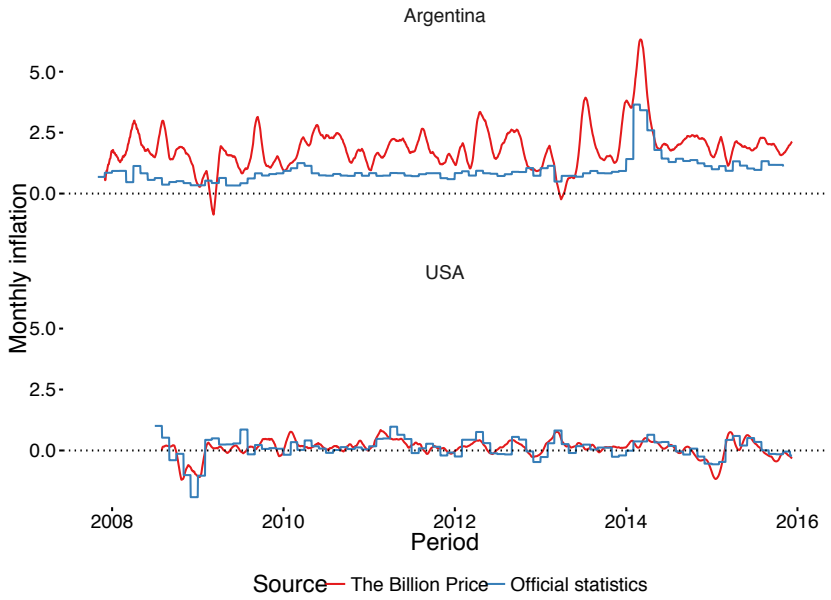


FIGURE 1.2: Official and The Billion Price Project monthly CPI for Argentina and United States

Another type of new data source that is used for official statistics are the above mentioned scanner data (SD), which are *electronic records of transactions that establishments collect as part of the operation of their businesses*. The most familiar and now ubiquitous form of scanner data is the scanning of bar codes at checkout lines of retail stores (Feenstra and Shapiro, 2007). SD provide an opportunity to switch from manual data collection by means of probabilistic sampling to using electronic databases maintained by private entities. Another precursor, apart from Switzerland and Norway, in using SD for price indexes is Statistics Netherlands (CBS). The use of SD for CPI started in June 2002 (with two supermarket chains Haan, 2002). In 2010 CBS started a co-operation with six biggest chains (Grient and Haan, 2010). Feenstra and Shapiro (2007), Haan and Grient (2011), and Ivancic et al. (2011) discuss problems regarding the use of scanner data for CPI.

NSIs use not only scanner data but also the Web and *web-scrapping*. Web-scraping (also called *web harvesting* or *web data extraction*) is a computer software technique of extracting information from websites. Usually, such software programs simulate human exploration of the World Wide Web by either implementing low-level Hypertext Transfer Protocol (HTTP), or embedding a fully-fledged web browser (Wikipedia, 2015). Hoekstra et al. (2012) describes web-scraping experiences at CBS that include data collection about unmanned petrol stations[17], four airline websites[18] and property offer prices[19]. CBS use sale offers that can be found at Funda.nl to link with the register of transactions. However, those results are not officially published. Hoekstra et al. (2012) focuses mainly on the organizational and cost-reduction aspect and how web-scraped data can be incorporated into the existing CPI. Bosch and Windmeijer (2014) describes

[17] See `www.tinq.nl`.

[18] See `www.klm.nl`; `www.transavia.nl`; `www.easyjet.nl`; `www.ryanair.com`.

[19] See `www.funda.nl`.

technical aspects of applying web-scraping techniques at CBS to obtain price information and Griffioen et al. (2014) provides an example of collecting clothing data from the Internet, underlining methodological problems connected with representativeness, missing data, regularity, redundancy and issues in automated data collection.

The Internet as a data source is not only used in the CBS, but also in other NSIs for various purposes. Barcaroli et al. (2015) describes the use of private companies' webpages for the purpose of Italian National Institute of Statistics (IStat) Information and Communication Technologies (ICT) survey. Barcaroli et al. (2015) applied several data mining procedures in order to obtain information on web sales functionality, order tracking or online job applications. Swier (2015) provides examples of web-scraping for supermarket (Tesco, Waitrose and Sainsbury's) grocery prices as one of BD pilot projects in Office for National Statistics (ONS).

In addition, social media are also in the spotlight of NSIs. One of the main examples are Twitter postings that are used to calculate consumer (confidence) sentiment or detect population flows. Daas et al. (2012) describes how Twitter works and whether it could be used as a data source for statistics in Netherlands. Daas and Puts (2014) provide a in-depth analysis of existing social media in Netherlands to determine which of them can be used to estimate the Customer Confidence Index (CCI). Daas and Puts (2014) used i.a. Facebook, Twitter, Google+ and LinkedIn data collected between June 2010 and November 2013 (over 3 billion messages, of which over 50% are redundant). The data was obtained by the CBS from a private company for BD pilot projects. Daas and Puts (2014) discovered that Twitter was strongly correlated and was co-integrated with the existing monthly CCI. Figure 1.3 shows the relation between the official and Twitter-based CCI between 2010 and 2013. Daas and Burger (2015) further studied Twitter data to extract information on user gender, age or origin (called profiling) in order to assess the selectivity. Munoz-Lopez (2015) presents the use of publicly available Twitter postings[20] of about 12 million accounts in Mexico, from which only 700 000 were geo-referenced. Twitter data in Mexico is intended to be used for internal flows, tourism, border communicating, use of roads, urban influence zones and subjective wellness.

However, we can point out a few important drawbacks of using Twitter to describe socio-economic phenomena. First of all, the key issue is representativeness of Twitter users – whether we analyse people that use social media or refer to the general population. Therefore, assessment of demographic characteristics is necessary. Secondly, limiting the analysis to posted messages (of no more than 140 characters) can influence the estimates, especially, given that people on the Internet behave differently than they do off-line (e.g. haters). Finally, these data are unstructured and their transformation may introduce additional bias into estimates. However, Twitter or similar social media will become more and more popular in the future, and their potential usefulness as a data source for statistics should therefore not be neglected.

On the contrary, data that is highly structured and often of high quality i.e. mobile sensors, road sensors and other devices, referred to as the Internet of Things

[20]Twitter makes only 1% of all postings available without additional cost.

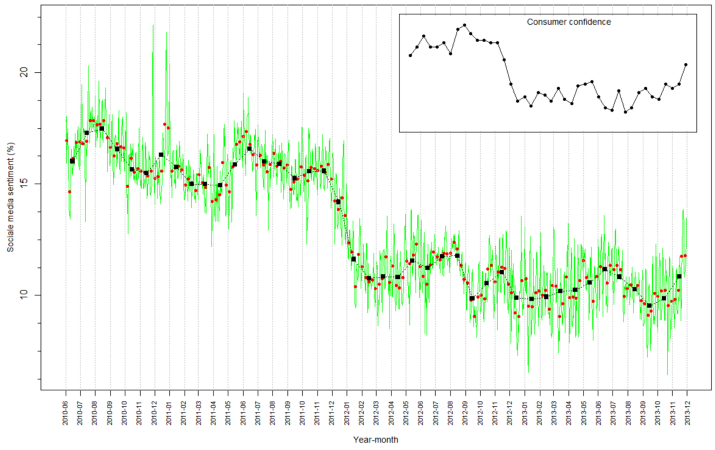


FIGURE 1.3: Twitter-based social media sentiment calculated by Statistics Netherlands

Note: based on Daas et al. (2015).

(IoT), are beginning to be used in Official Statistics. The CBS is the first NSI to include a BD source – road sensors (over 60k road loops in the Netherlands) – in the production of statistics. Figure 1.4 presents results of a study devoted to traffic statistics conducted by the CBS (Daas et al., 2013). Figure 1.4 consists of two figures: the distribution of road sensors on the main roads in the Netherlands (left), and the distribution of the number of vehicles detected in three length categories on 01.12.2011 in the Netherlands (right). The decision to use road sensors was mainly motivated by the high coverage of the main roads in the Netherlands. Only the western and northern parts are less covered; this is due to the small population and road density. In total 4 TB of data for 2010–2014 were obtained (2.7 GB per day), which consisted of 336 billion records (230 mln each day). On completion of the data cleaning process, the size of the final data base was reduced to 1 GB.

Vehicles were categorized into three groups based on their total length. The first group included small vehicles (<= 5.6 meter, 75% of all vehicles) for which the morning peak was observed to be around 6 am (early workers) and rush hours from 8 am to 5 pm. Medium sized (> 5.5 and <= 12.2 meter, 12% of all vehicles) and large (>12.2 meter, 13% of all vehicles) vehicles had different peaks than small vehicles. For instance, during rush hours the number of medium and large vehicles was considerably lower than that of small vehicles. This is the result of deliberately avoiding intensive traffic. Puts et al. (2015) describes in detail the amount of collected data, technical problems and the quality of these data. Puts et al. (2015) applied recursive Bayesian estimation (assuming the underlying Poisson process) to clean the data. Tennekes and Puts (2015) focus on describing technical problems with the positioning of road sensors on the roads as well as the weighting scheme based on road segments lengths that was applied (for more details please refer to Daas et al. (2015)).

Another group of data are mobile sensors (MS), which consist mainly of GPS and positioning of mobile phones. The main aim of using MS is to estimate population flows, migrations and tourism. This issue was highlighted at the New

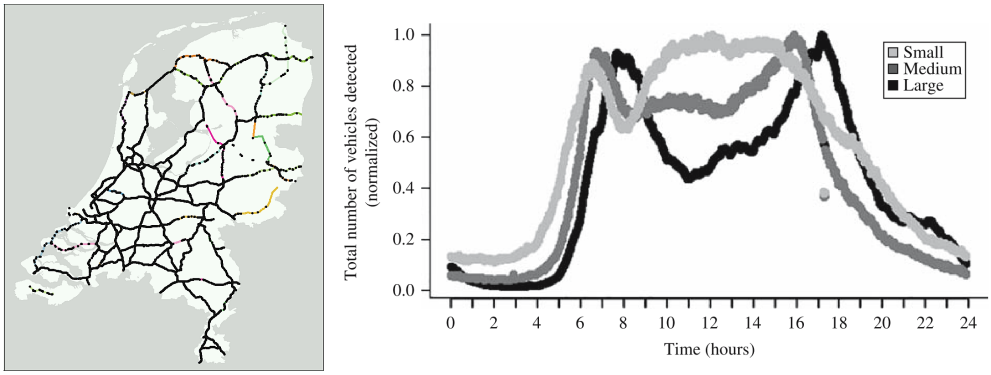


FIGURE 1.4: The distribution of road sensors and the number of vehicles in the Netherlands on 01.12.2011

Note: based on Daas et al. (2015) and Tennekes and Puts (2015).

Techniques and Technologies for Statistics (NTTS) conference in 2015 in Brussels, which featured a special session devoted to *Mobile Phone data as a source for official statistics*. Demunter and Reis (2015) summarise *Feasibility study on the use of mobile positioning data for tourism statistics project* a project coordinated by Eurostat between 2013 and 2014 in Estonia (Eurostat, 2015). The main issues raised by Demunter and Reis (2015) focus on: (1) under- and overcoverage, (2) continuity of data and unstable behaviour and preferences of customers, (3) the problem of reconstructing the existing scope and definitions, (4) the absence of socio-demographic and domain-specific information, (5) lack of existing indicators on regional breakdowns. However, Demunter and Reis (2015) stress that an improvement in the quality (better timeliness, less respondent burden, less recall bias) and increase in coherence was observed. Ahas et al. (2007) discusses the use of mobile positioning data to estimate seasonal tourism spaces in Estonia, which was the preliminary research that led to Eurostat (2015) the project.

The last strand of literature devoted to official statistics are small area estimation studies that take into account auxiliary information (mainly from census or administrative data). Recently, several studies on the use of IDS and BD sources have been published. Marchetti et al. (2015) and Pratesi et al. (2013, 2014) discuss the use of location data obtained from GPS as an auxiliary variable to estimate poverty in Tuscany. Marchetti et al. (2015) notes that in some cases there is a lack of official data at the domain level and BD sources can be used as a proxy to shed light on the socio-economic phenomenon. Pratesi et al. (2014) proposes a model that takes into account measurement errors in auxiliary variables that are produced by the applied instrument (mobile sensor). Porter et al. (2014) applied a spatial Fay-Herriot (SFH) model that took into account GT indecies. The goal was to estimate relative changes in the percentage of Spanish-speaking households in the eastern half of the United States. The main difference compared to previous studies was to treat GT as functional data. In particular, Porter et al. (2014) proposed a SFH model that included functional covariates at area level, which on average reduced the variance of direct estimates. Bacchini et al. (2014) discusses whether

GT can improve the quality of estimates of short-term (monthly, quarterly) socio-economic indicators from the NSI perspective. Bacchini et al. (2014) analyses the unemployment rate, labour demand, trips and vacations at regional and provincial level and compares it with "job" and "job offers" searches in GT. Lahiri (2015) discusses the use of BD auxiliary data obtained from remote sensing (for crop acreage) in several small area models.

Marchetti et al. (2015) and Pratesi et al. (2013) summarise possible ways of using new data sources for small area estimation in the following categories: (1) the use of new data sources as covariates in SAE, (2) the use of survey data to remove self-selection bias from estimates using new data sources (3) the use of new data sources to validate small area estimates. Liao (2014) discusses BD in the context of small area estimation in the following way: (1) to apply a conventional sampling-based approach to BD (with respect to administrative data); (2) to combine non-probability sample data with probability sample data: many types of BD, such as data collected by Google/Twitter/Facebook, are not census (population) data. We may treat them as non-probability sample data; (3) to use high-quality small data for measuring and adjusting errors in BD: BD are not only non-representative of the target population, but also carry loads of measurement errors because the construct behind a particular measure in these data can differ from the construct that analysts require.

The last issue is the quality of such data, which is addressed by Struijs and Daas (2014) and Buelens et al. (2014). Struijs and Daas (2014) lists five key ingredients of data quality that are considered in official statistics: (1) relevance, (2) accuracy and reliability, (3) timeliness and punctuality, (4) coherence and comparability and (5) accessibility and clarity, which are not widely discussed in the context of BD. Struijs and Daas (2014) state that the meaning or relevance of data may be difficult to gauge and information about the population behind the records used may be missing. Struijs and Daas (2014) conclude that in particular relevance, reliability, timeliness and coherence are quality requirements that may still be applied when dealing with BD. On the other hand Buelens et al. (2014) discusses the selectivity problem connected with new data sources and proposes a diagram to measure selectivity using existing research, surveys and registers as a benchmark.

The literature presented in this review indicates a lack of systematic approach to the use of the Internet or big data for statistics. Examples of the use of these data source focus on “data mining” (extracting knowledge from the data) rather than providing solid theoretical grounds. Hence, the next sections and the following chapter will attempt to bridge this gap.

1.4 Definition of Internet data sources

In this section a definition of IDS will be proposed together with a classification of data sources. However, to clarify the following discussion some basic definitions concerning the Internet and World Wide Web will be presented. According to the Oxford Dictionary and the glossary of Poland’s Central Statistical Office, the *Internet* is a global and public system of interconnected computer networks

that use the standardized Internet Protocol Suite (TCP/IP). It is a network of networks consisting of millions of local networks and individual computers from all other the world. E-mail, www, FTP and other services are available to use via the Internet. The *World Wide Web* (WWW/the Web) is part of the Internet; it is an information system on the Internet, which allows documents to be connected to other documents by hypertext links, enabling the user to search for information by moving from one document to another. To access the Web, users rely mainly on *web browsers*, which are programs used to navigate the Web by connecting to a web server, allowing the user to locate, access, and display hypertext documents[21] or *mobile applications* that are computer programs designed to run on mobile devices, such as smartphones and tablets[22]. Lately *the Internet of Things* (IoT) has gained recognition as a potential source of information. IoT is defined as the interconnection via the Internet of computing devices embedded in everyday objects, enabling them to send and receive data (independently of the user). Finally, a *web service* is a service offered by an electronic device to another electronic device, communicating with each other via the Web. In practice, a web service typically provides an object-oriented web based interface to a database server, utilized for example by another web server, or by a mobile application, which provides a user interface to the end user[23].

In addition, Bethlehem and Biffignandi (2011) provide the following definitions of surveys that should be taken into account:

Internet survey A general term for various forms of data collection via the Internet. Examples are a web survey and an e-mail survey. Also included are forms of data collection that use the Internet just to transport questionnaires and collected data.

Web survey A form of data collection via the Internet, in which respondents complete questionnaires on the World Wide Web. The questionnaire is accessed by means of a link to a web page.

Self-selection survey A survey for which the sample has been recruited by means of self-selection. Users can decide whether or not to participate in a survey.

In the light of the above, the following definition of IDS is proposed:

An **Internet data source** is a self-selected (non-probabilistic) sample that is created through the Internet and maintained by entities external to NSIs and administrative regulations.

The definition emphasises a number of aspects. First, despite its volume, an IDS should be treated as a *sample*. The reason for that is an IDS does not contain all units from the target population. Secondly, unlike official statistics, which are based on probability selection mechanisms, IDS are the result of the

[21] A particular example is a mobile browser that can be accessed on mobile devices, such as smartphones or tablets.

[22] Mobile applications must first be downloaded and installed on mobile devices.

[23] See `https://en.Wikipedia.org/wiki/Web_service`.

self-selection process: the decision whether to provide information to IDS is left to individuals/entities, which reflects the *non-probabilistic* character of IDS. The definition explicitly states that data are not collected by statistical institutions or public agencies but private/commercial entities. The definition specifies that an IDS only refers to data created by Internet users or by private entities themselves; official databases, such as Eurostat database[24], OECD statistics[25], Statistics Finland StatFin[26] or the Polish Local Data Bank[27] are excluded from the IDS category. Finally, IDS are created via the Internet without the support of on-line questionnaires used in on-line surveys. In fact, *IDS are a new type of Internet surveys*, where data are collected directly from a given on-line service. This type of survey could be named as an *IDS survey* or an *IDS-based survey*. IDS data are the result of interactions of individuals and enterprises with the Internet. IDS might be obtained by web-scraping, specific web services (e.g. representational state transfer REST[28], Simple Object Access Protocol SOAP[29]) or directly from the owner. The access level is set by the data source holder and regulated by terms and conditions of use specified by a given portal. The possibility of automatic data collection should be discussed with the data holder and depend on the level of cooperation with statistical agencies. An IDS rarely consists of statistical units - data tend to be objects (non-statistical units), actions or aggregated data that should be transformed. These data sources might also consist of processed data published in reports or statistics (see Google Trends).

Taking into account the definitions and classifications discussed in the previous section, we include Internet data sources in the classification of data sources in statistics. The proposal is presented in Diagram 1.1.

The first stage identifies statistical data sources, designed by NSIs for statistical purposes and non-statistical data sources that do not meet these criteria. The main examples of statistical data sources are probability based sample surveys (e.g. EU-SILC, LFS), (traditional) censuses that are not based on administrative registers and reporting that provide mainly information on establishments (for business and short-term statistics). In the second stage we distinguish between administrative/government data sources and commercial/private data sources. The main idea that underlies this distinction is the availability of these data sources for statisticians and statistical agencies. Register data can be accessed given specific provisions in regulations (e.g. the use of registers for statistical purposes) while access to private/commercial data sources can only be obtained after establishing collaboration between NSIs and private entities. The main examples of administrative data sources are population, unemployment or social benefits registers. Moreover, this group can also include traffic sensors, surveillance videos, satellite images, transactions on the stock market or other data lawfully collected according to existing regulations (which are regarded as BD). In the last stage we divide private/commercial data sources into two groups – created on and outside the Web by individuals and enterprises. The first group is denoted by the term

[24] See http://ec.europa.eu/eurostat/data/database.
[25] See http://stats.oecd.org.
[26] See http://pxnet2.stat.fi/PXWeb/pxweb/en/StatFin/.
[27] See http://stat.gov.pl/bdlen/app/strona.html?p_name=indeks.
[28] See https://en.Wikipedia.org/wiki/Representational_state_transfer.
[29] See https://en.Wikipedia.org/wiki/SOAP.

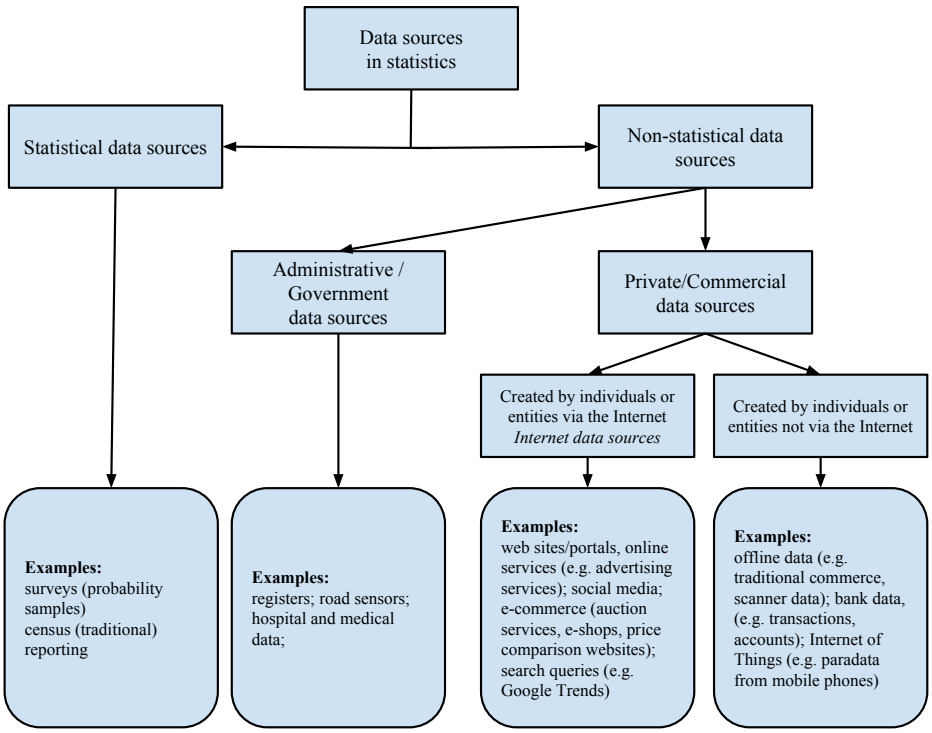


DIAGRAM 1.1: Classification of data sources in statistics including new data sources

Internet data sources (IDS), the second is labelled as *non-Internet data sources* (non-IDS). The common denominator is how these data are created. IDS are created by interactions of individuals or enterprises with the Internet, while non-IDS are the result of passive data collection or without the Internet. For instance, IDS consist of web portals (devoted to specific fields e.g. blogs, news, sports), online services (e.g. real estate ads, car ads), social media (e.g. Facebook, Twitter) or Internet search queries (e.g. Google Trends). In contrast, non-IDS data may come from traditional commerce, data held by telecommunication companies (e.g. SMS, recordings, videos, positioning), e-mails and messages sent via online services (e.g. Gmail, Facebook), transactions made on- and off-line, information about users of online services (demographics, activities), the Internet of Things (e.g. sensors, GPS) or geolocalisation (e.g. geotagging, positioning).

1.5 Internet data sources compared to existing data sources

Censuses, surveys and administrative sources are used by NSIs as well as statisticians/researchers to provide information for a given target population. However, new data sources, in particular IDS, are becoming an interesting alternative. Table 1.1 provides a detailed description of statistical concepts for censuses, surveys, administrative sources and IDS. In addition, the 3V definition is also included.

The first concept refers to *coverage* of the target population. In the case of statistical sources, such as a census or a survey, coverage is assumed to be known. A census aims to fully cover the target population, while a survey is based on a sampling frame, which, ideally, should be free of coverage errors. However, in practice this assumption rarely holds: for instance an under-coverage error is present in census data due to outdated lists of units or addresses. In such cases post-enumeration surveys are often conducted to assess the census coverage. The coverage of surveys depends heavily on the quality of the sampling frame, which influences the probability of a given unit being selected. Because administrative sources and IDS were not created for statistical purposes, they may also suffer from coverage errors. Hence, administrative sources cover *de iure* population, defined by specific regulations. IDS are restricted to a subset of the target population that has access to the Internet (Internet population) on the one hand, and, on the other hand, to units that use a given IDS from the Internet population. Therefore, the coverage error might be high in populations where Internet access and use is low or heterogeneous.

Another aspect of data sources used for statistical purposes is *representativeness*. In the case of censuses and surveys, because of their statistical character, representativeness is known and set a priori. For instance, the EU-SILC sampling scheme ensures reliable estimates at the level of countries or macro regions. When non-response is present in surveys, information about units that did not participate is available. On the other hand, the IDS population is often unknown a priori and its representativeness is difficult to assess without access to raw data. In this respect, this makes IDS similar to registers. For instance, on-line services devoted to the real estate market in Poland provide information about the average offer price or the structure of the market; however, the methodology behind these statistics is often unknown. Hence, the availability of IDS raw data, the application of data cleaning procedures and statistical methods would provide information on the distribution of variables for which representativeness is to be measured. However, owing to the massive, complex character and redundancy of such data, the assessment of their representativeness may be a time consuming process.

Respondent burden is another issue considered for comparison. In a census or a survey a respondent is often approached with a (paper or electronic) questionnaire and is asked to answer several (often many) questions on different topics. This approach often results in high non-response rates, or attrition in panel surveys. As a result, effective sample sizes are reduced significantly. Registers are supplied with obligatory information collected by government institutions, hence no additional burden from NSIs is created. In that sense IDS are similar to registers. IDS are electronically supplied by respondents (IDS users). In addition, often IDS are also supplied with data obtained from passive collection (e.g. collected by mobile devices or software used by IDS users). However, the amount of data provided by IDS respondents depends on their decision. IDS are not obligatory, but often the use of certain on-line services is necessitated by competitors or society. For instance, there are professional web-services for employees (such as LinkedIn), where people present their professional experience in order to find a better job; real estate brokers try to facilitate the sale of flats (especially

TABLE 1.1: A comparison of censuses, surveys, administrative and Internet data sources

Issue	Census	Survey	Administrative	IDS
Coverage	Assumed full coverage of the target population	Based on the sampling frame assumed to have no under- or overcoverage	May have under- and overcoverage	Limited to the Internet population; the coverage varies between sources
Representativeness	Known by design	Often known	Often known, easier to assess	Unknown; may be difficult to assess
Respondent burden	Heavy	Light, but increasing	No additional burden	No additional burden
Self-selection	Negligible	Moderate (depends on unit non-response)	Low; depends on register	High
Bias	Assumed to be unbiased	Assumed to be unbiased	Possibly biased	Unknown and possibly biased
Content	Wide range of information	Narrow range of topics	Variables required for administrative purposes	Restricted to variables for business purposes; redundancy
Concepts	Created for statistical purposes	Created for statistical purposes	Created for administrative purposes	Organic or designed for business or other purposes
Quality control	Designed to minimize errors	Designed to minimize errors	Controlled by the administrative agency, limited	Controlled by the data holder, limited or none
Cost	Expensive	Relatively low cost	Relatively inexpensive	Relatively inexpensive, depends on cooperation
Frequency	Every 10 years; depends on administration	Monthly, quarterly, yearly	Continuous observation	Continuous observation
Timeliness	Major delays	Minimal delays	Depends on the reporting process	No delays; constant registration
Stability	Controlled by statisticians	Controlled by statisticians	Changes due to legislation, regulations or administrative practices	Depends on the data holder (enterprise); statisticians do not control the changes
Coherence	Created to be coherent	Created to be coherent	May not be consistent with statistical data sources; may allow for international comparisons	May not be consistent with statistical data sources; may allow for international comparisons
Volume	Manageable volume	Manageable volume	Manageable volume	Possible problems with management
Velocity	Freeze-framed/bundled	Freeze-framed/bundled	Relatively fast, continuous	Fast, continuous
Variety	Limited	Limited	Limited	Wide

Note: own elaboration based on Brackstone (1987), Kitchin (2015), and Wallgren and Wallgren (2014).

to younger people) by using on-line advertisement services to present their offer. Therefore, the decision whether particular information should be put on line depends on the (in)dependent decisions of individuals or establishments.

Self-selection refers to the situation when individuals or entities themselves decide to participate in a survey or publish information on the Web. Censuses are obligatory, i.e. all units of the population know about the census and are obliged to provide information during the enumeration. Hence, in this case self-selection is negligible. Registers are a similar case. For instance, each individual is obliged to have an national identification number[30] or a new business entity should register its activities. In some registers there area cases of self-selection. For instance, not all unemployed people register or have insurance. Sample surveys and IDS are similar as far as self-selection is concerned. In surveys, especially in web surveys, high non-response is related to respondents' decision whether to participate or not. In IDS individuals or entities decide whether to use a given IDS or not. In both cases selection (participation) results in the measurement of specific units, which might not be representative of the target population. Survey or data source specific self-selection mechanism makes the modelling process difficult and requires both statistical and domain knowledge.

The above mentioned problems often result in *biased* estimates. Censuses and sample surveys, by design are assumed to be unbiased. However, in the presence of non-ignorable non-response (missing not at random, MNAR) in both sources bias may be introduced. Bias and its sources in registers and IDS are often unknown a priori. For instance, in register data may be biased as a result of delays, differences between definitions (measurement error) or populations (Wallgren and Wallgren, 2014, ch. 13). In the case of IDS, which are non-probability samples, bias is observed mainly owing to the self-selection mechanism. For instance, specific units might be observed on-line and self-selection might depend on the target variable. As a result, IDS are similar to surveys in terms of (non-ignorable) non-response, while they resemble registers with respect to the unknown level of bias.

Data sources vary in *content*, or in what kind of statistical information they provide. Censuses are designed to cover a wide range of socio-economic information for each unit of the target population. On the other hand, surveys are created to measure a specific topic, such as living conditions, economic status or ICT use. Registers and IDS are similar in the kind of information that is collected and variables that are required for administrative and business purposes respectively. IDS often contain information which is unstructured and presented in a natural language (e.g. statements, descriptions) and the final content is unknown a priori. In addition, redundancy of information in IDS is substantial and may be outside the scope of statistics. Moreover, paradata that accompany survey data collections or IDS might be within the scope of what statisticians are interested in (Kreuter, 2013; Zhang et al., 2013).

Statistical data sources measure *concepts* that are clearly stated and defined by statisticians and for statistical purposes. On the other hand, registers and IDS use concepts defined by the law or internal regulations of a given entity. Therefore, differences between statistical and non-statistical definitions may be substantial. For instance, a household is a typical example of a statistical concept, which is not present in registers or IDS. IDS concepts differ from statistical ones because they reflect different business purposes that underlie the creation of IDS. However,

[30] In Poland *Powszechny Elektroniczny System Ewidencji Ludności* (PESEL) is a national identification number, which is issued by the Ministry of the Interior and Administration.

there are examples of non-statistical and statistical concepts that are coherent. For example, the definition of residential real estate used in statistics and IDS is the same, which enables direct comparisons.

The level of *quality control* differs substantially between data sources. Censuses and surveys are designed by NSIs and statisticians to minimize errors. For instance, questions included in survey questionnaires should be clear and easy to interpret (to avoid mistakes); respondents are given possible answers that were predefined by statisticians. However, information provided in surveys or censuses is based on respondents' declarations. Administrative sources, owing to their obligatory character and the threat of financial penalties for providing false information, should be of highest quality. However, the level of quality control exercised by government agencies and NSIs may vary significantly. For instance, government agencies may not perform logical checks (e.g. divorced children) that are conducted by NSIs. With respect to IDS, as in the case of administrative sources, quality checks are made by the data holder. In most cases it is automated, for instance, form fields contain certain automatic checks, such as non-negative floor area or price above a given limit. Facebook uses paradata information on the use of their services, uploaded photos, information about friends or mobile phone paradata to verify age, gender or current location. Nonetheless, these methods may differ from those presented in the statistical literature and result in unreliable information on data quality.

Another issue presented in Table 1.1 concerns *costs* of designing and using particular types of data sources. The most expensive ones are censuses and surveys, which can cost thousands or millions of PLN (or EUR, USD). On the other hand, administrative sources and IDS are already available and the cost involved in using them is mainly connected with adapting them for statistical purposes (infrastructure, staff, training, collaboration with data owners, privacy, safety). For instance, the infrastructure available in NSIs is often not suitable to analyse massive data or to continuously observe the Internet (Daas et al., 2015).

Among the factors motivating the use of new data sources for statistics are *frequency* and *timeliness*. Censuses, which are organized once every 10 years provide detailed information but often with substantial delays. For instance, the last set of statistics derived from the National Population and Housing Census 2011 (NSP 2011) was published only in 2015. However, users of such data are often not interested in information provided with a long delay. On the other hand, surveys are conducted every month, quarter or year and provide information limited to a specific issue with a short delay. The National Bank of Poland and the Poland's Central Statistical Office conduct a survey of the real estate market and publish the final report on a given year at the end of the following year. In contrast to statistical data sources, registers and IDS store information from continuous observation, which is required by the relevant legislation in the case of administrative sources, or market considerations in the case of IDS. In addition, IDS operate 24h a day, and timestamps are made automatically, without additional participation of the user. In terms of timeliness, once again administrative sources are limited by the legislation. For instance, on the real estate market in Poland notaries have 30 days to provide information about transactions; authorities, in turn, have a month to input the data, which may result in a maximum delay of 2 months. In the case

of IDS, the delay may also be connected with data processing or changes in the structure of a given data source (e.g. when web-scraping is applied). However, if all information was provided electronically, then statistics could (ideally) be produced "in real time".

Another issue is *stability* of data sources. Censuses and surveys are controlled by NSIs and the implementation can only be interrupted by tragic events (e.g. war) or substantial budget cuts ("twilight of classical census", virtual census). Owing to their statistical character, these data sources are stable and provide comparable statistics over time. A similar situation can be observed for administrative data sources. Registers provide information that is required by the government, regardless of what party is currently in power, which makes them the most stable source of information. In contrast, IDS, which are created for business purposes, may disappear owing to the lack of interest (e.g. MySpace), when the company closes down (e.g. Napster) or is bought and its IDS is merged with another one. However, there are examples of companies, such as Google, Amazon or Twitter, which are not likely to disappear in the next decades. Another aspect is that in the case of registers and IDS statisticians do not control changes that are made in these data sources. Thus, statisticians should analyse the data collection process that adapt it to the new environment and the mechanism of data generation.

Another aspect that should be taken into account for comparison is *coherence*. Official statistics are created to be comparable in time and between countries. For instance, the EU-SILC survey is conducted by each EU member state and is suitable for direct comparisons. Another example is the application of the same methodology (for instance LFS definitions, calculation of CPI) to be comparable between countries all over the world. There are several initiatives that are coordinated by Eurostat or UNECE that focus on ensuring the coherence between national statistics. In the case of administrative sources, coherence may be measured in reference to statistical data sources and for international comparisons. In comparison to censuses or surveys, register data may not be coherent owing to differences of definitions. On the other hand, certain legislations are coherent between countries (e.g. crime statistics, traffic accidents) and can be used for international comparisons. Administrative data sources are being used for statistics because statisticians transform certain registers to be consistent with existing data sources. IDS, owing to the lack of definitions, may not be consistent with other data sources and, like registers, require data transformation. Nonetheless, global companies (e.g. Google, SkyScanner, Booking, Facebook) provide on-line services that are coherent between countries. For instance, the Billion Price Project collects data from global companies, such as H&M, Zara, Ikea or Apple to calculate the CPI for several countries.

The last three concepts presented in Table 1.1 refer to the 3V definition – Volume, Velocity and Variety. Censuses, surveys or administrative records are of manageable size, while new data sources often require special infrastructure to enable the processing of these data. For instance, survey data may contain $\sim$ 100k rows per month (e.g. the Household Budget Survey) while in an IDS a comparable number of rows may be generated each day or each minute. In addition, classical statistical methods are not designed to handle such massive data sources and may not be efficient or even applicable to IDS. In terms of *velocity*

statistical data sources are maintained by NSIs, while IDS are the result of the use of online services. IDS also include the use of the Internet of Things, keeping a record of each activity made by the user. For instance, while using certain advertisement services, information about the user's location based on Wi-Fi or mobile phone is collected, together with user characteristics (e.g. cookies). The last V refers to *variety*. In the case of censuses or surveys, the data structure is often simple and consists of tables with columns for variables and rows for recorded responses. These sources have predefined and coded questions that are not modified by respondents. However, in the case of registers and IDS, there are many possible ways how data can be collected and represented. For instance, both types of sources may contain tables with records concerning units or objects (actions), but also text descriptions, photos or videos (surveillance cameras) or recordings. These data types are not common in statistics, making it problematic for NSIs and official statisticians to apply classical statistical methods.

The analysis of Table 1.1 indicates that IDS have more in common with administrative sources than with surveys or censuses. Similarities are mainly to do with the non-statistical characteristics of these data. Detecting similarities and disparities between sources can help to identify risks and challenges connected with the use of IDS for statistics.

1.6 Risks and challenges in the use of Internet data sources

Table 1.2 contains selected risks and challenges of IDS classified into three groups. The first group refers to opportunities of using IDS as a data source. Traditional statistical sources, such as censuses or surveys, are connected with an increasing response burden, which leads to increasing non-response. The use of IDS might positively influence the response rate by decreasing response burden thanks to relying on data that is already available. Information obtained from IDS could be used instead of asking certain questions in surveys. For instance, Barcaroli et al. (2015) presents how web-scraping could be used to obtain characteristics of company web pages that are studied in the ICT survey.

The electronic format of IDS makes it possible to obtain data using special algorithms, which decreases data collection costs. Instead of manual data collection from web pages, programs for automatic data capture can be developed. These programs might be set to run on a daily, weekly or monthly basis. The data collection process can be designed to decrease the "web page" burden, not generate additional server traffic. For instance, web scraping can collect data overnight or at weekends. However, development of these algorithms requires programming skills, appropriate web services and web technologies used for the creation of web pages. Web scraping programs should be prepared to handle different web pages (e.g. by preparing one algorithm that recognizes common elements) and possible changes in the website structure. In such cases, it may be a better solution to use API to collect data directly from owners of servers instead of scraping web pages. As a result, automatic data collection could decrease the cost of certain surveys.

The use of IDS is a solution to the demand for new data that could enhance or even replace certain existing statistical data sources. The electronic format of IDS enables statisticians to collect data for research in a more convenient way. Moreover, IDS are created as a result of interactions of individuals and enterprises with the Web, which makes them potentially suitable for the measurement of large scale phenomena (e.g. Google Trends). For instance, a comparison of on-line and offline prices might provide insights into how prices behave and whether they differ substantially. The use of IDS for the real estate market might provide additional information on flats or houses, which could then be linked to transaction data. Combining data with IDS might reduce the number of questions that are asked in questionnaires.

Official statistics produces statistics on certain characteristics of the target population at a predefined level. For instance, NSP 2011 generated data at LAU1 and LAU2 level, the Polish EU-SILC provides information only at the NUTS1 level. To cover new fields, topics or domains it is often necessary to change the sample size or even to create a new survey, which obviously increases costs. However, the use of IDS and/or linkage with existing statistical data sources may facilitate the calculation of new statistics. For instance, linkage of real estate or car sale offers with the register of transactions might help to calculate the size of the market (supply), differences between offer and transaction prices or time-to-sale. In addition, information presented online might be used as early indicators for official statistics.

Finally, new data sources could be used as auxiliary variables for model-based estimation, in particular small area estimation. On the one hand, IDS might provide detailed geolocated data which could be aggregated at any level; on the other hand individuals using the Web might provide information that could be helpful in understanding certain phenomena. For instance, the literature review provided examples of Google Trends being used for small area prediction. IDS could be used as proxy variables for official indicators in groups of young people who use the Internet on a daily basis.

The second group consists of selected challenges related to the use of IDS for statistics. The first issue concerns consistency with existing existing statistical data sources. In statistics, the target population, statistical units and concepts are precisely defined (e.g. ILO unemployment classification). In contrast, IDS often consists of objects that are not statistical units or include definitions that are only similar to those used in statistics, in particular official statistics. Hence, it is crucial to develop methods that make estimates based on these data sources coherent.

Another challenge is the measurement of representativeness. Distributions of target variables, data characteristics or the number of statistical units are often unknown a priori. For instance, Twitter users might not provide information that enables identification, multiple advertisements might refer to the same real estate. This gives rise to problems with unit identification. These problems influence the assessment of representativeness and the estimation of basic characteristics, such as distribution of sex, age or employment status. Moreover, IDS often provide only a subset of information that is held by the data owner. Since only part of the information is available, the measurement of representativeness may be limited. Finally, the lack of official data on a given target population makes comparisons

TABLE 1.2: Opportunities, challenges and risks connected with using Internet data sources for statistical purposes

Opportunities
• reduction of respondent burden by using already available data
• cost reduction for selected surveys (automation)
• replacement, enhancement or correction of existing data sources (data resolution)
• obtaining new information that was not available for official statistics before
• use of new data sources as auxiliary variables for model-based estimation
Challenges
• consistency with existing statistical data sources
• measurement of representativeness (e.g. unit identification)
• coverage of domains that are published in official statistics (e.g. sex, age, marital status)
• country-level and international-level law and regulations (data privacy and security)
• the need to co-operate with data owners (e.g. private companies) and obtaining access to the data
• integration into the statistical system
• application of existing methods used in official statistics (e.g. data cleaning and edition, imputation, calibration)
Risks
• data availability and continuity
• data privacy; social resistance against the use of these data by official institutions (e.g. "total surveillance").
• self-selection mechanism
• model-based estimates

Note: own elaboration based on Beręsewicz and Szymkowiak (2015).

difficult or even impossible.

Official statistics provide information on domains such as sex, age or marital status. These statistics are also available at low levels of aggregation (for instance LAU1, LAU2). IDS vary in the available levels at which data could be published. For instance, Google Trends provide search indices at NUTS2 level (for Poland) but there is no information about gender or the age of people for which these statistics are provided. On the other hand, IDS for the real estate market contain characteristics about properties put up for sale, but do not contain information whether a given property was sold. Information stored in IDS are often optional and is not verified. Although there are machine learning techniques designed to extract information about gender, age or other characteristics, obtaining information at the same levels of aggregation as those offered by official statistics might be challenging.

The use of registers for statistical purposes is defined by national regulations and laws. Regulations specify how these data can be used, combined and disseminated by NSIs. Data collected in surveys and censuses are secured and subject to disclosure control. Moreover, not all information is published owing to statistical confidentiality. On the other hand, IDS data are collected by private entities

that have different regulations concerning privacy and data protection. Preserving privacy is a crucial factor, especially when data are geocoded or contain sensitive information.

As stated above, IDS data are created and maintained by entities external to NSIs and do not have to comply with administrative regulations. For this reason, it is necessary to establish co-operation between statisticians/NSIs and private companies in order to obtain access to IDS. These steps often require long negotiations with entities. For instance, Statistics Netherlands conducted four-year negotiations with telecommunications companies, which resulted in obtaining only aggregated data at predefined levels. On the other hand, SN has been receiving on a monthly basis sales data from seven main supermarket chains in the Netherlands since 2010. However, the level of co-operation possibly depends on the company's origin - whether it is part of a globally or locally operating enterprise, and its willingness to share data and the burden resulting from such co-operation.

Another challenge is the integration of IDS into the statistical system. First, integration of IDS requires common identifiers or variables in data sources. For instance, cars have unique Vehicle Identification Numbers (VIN) which provides a direct link with the register of cars; properties can be linked by means of the address and its characteristics. Facebook or Twitter users need to be profiled to extract common variables. Second, the level of integration is crucial: micro integration is not always possible; instead, macro integration might be suitable. Finally, the integration of multiple data sources requires probabilistic methods, a methodology which relies on links and hardware to combine data at a large scale.

Traditional statistical methods used in survey methodology, such as data edition, imputation, calibration or variance calculation are suitable for moderate volume and dimensionality of data. Internet data sources are non-probability samples that probability-based methods might be not suitable for. Data currently used for statistics are designed by statisticians and their structure is known. However, IDS contain data that are rarely used for statistics, for instance photographs, text descriptions, with a completely different data structure (e.g. XML, JSON). Therefore, the application of existing methods in such an environment might require the relaxation of the underlying assumptions of existing methods, the adoption of machine learning models and the implementation of data processing that is suitable for new data structures.

The last group is related to risks involved in the use of IDS for statistics. First of all, since these data are not created for statistics, it is possible that in the future a given data source will become less popular or even disappear. Another problem are ownership changes in private companies that may influence the level of co-operation and, consequently, data availability. However, it should be noted that problems with estimation and the choice of possible methods will be independent of data availability and continuity.

Another issue is connected with social resistance to the use of these data for official statistics. Collection of private data by government agencies, even if NSIs are independent, might raise concerns. NSIs and statisticians might be accused of participating in"total surveillance" by collecting data on individuals from private entities. Privacy preservation and data protection will be crucial.

IDS are generated by means of a self-selection mechanism and largely determined by the behaviour of individuals and companies. The correct specification of the response mechanism is crucial for unbiased estimation. However, if self-selection is connected with the target variable (e.g. income, price of real estate) bias might substantial. In addition, this process might not be constant in time or space. For instance, people in different geographic parts of a given country might have different response mechanisms.

Given that IDS are non-probability samples, the application of classical estimators, such as Horvitz-Thompson is not possible. This problem can possibly be remedied by using model-based estimation. However, models are based on certain assumptions, need strong variables that explain the target variable and correctly specify causality. If the model is wrongly specified or there is a lack of variables that could explain self-selection and the target variable, model-based estimation might provide biased estimates.

1.7 Conclusions

The purpose of this chapter was to provide basic information about statistical data sources as well as new data sources, in particular Internet data sources. A definition of IDS was proposed and illustrated with several examples of such data. The literature review presented various aspects of IDS in official and non-official statistics. The first chapter ends with a summary of potential risks and challenges connected with the use of IDS for statistics. Information covered in the chapter was limited to a general overview of IDS in the statistical context. The next chapter will provide detailed characteristics of IDS in the context of statistics, particularly real estate market statistics.

Chapter 2

Internet data sources on the real estate market

2.1 Definitional vagueness of populations in the real estate market

This chapter is devoted to the study of populations in the real estate market. First, we distinguish the IDS population and its relation to the target population. For the purpose of the study the target population is defined as **residential properties offered for sale in the secondary market** and a statistical unit is defined as **residential real estate**. However, for clarity a shorter name for the population **properties for sale** will be used. In addition, the terms 'target and general population' will be used interchangeably.

2.1.1 The IDS population

The general relation between populations and the sample in the context of IDSs is presented in Figure 2.1. We have distinguished four subpopulations of the general population and the sample. Let $\mathbf{\Omega}_{TP}$ denote *target population* of N_{TP} size. The target population consists of *statistical units* denoted by i and numbered $\{1, 2, ..., N_{TP}\}$. In addition, we assume that $\forall_{i \in \mathbf{\Omega}_{TP}} \pi_i > 0$. $\mathbf{\Omega}_{TP}$ represents all residential properties that are offered for sale in the secondary market.

The second population is the *Internet population* denoted by $\mathbf{\Omega}_{IP}$ of N_{IP} size. It should be noted that the relation between these two populations is $\mathbf{\Omega}_{TP} \subset \mathbf{\Omega}_{IP}$ not $\mathbf{\Omega}_{TP} \subseteq \mathbf{\Omega}_{IP}$. This means that $\exists_{i \in \mathbf{\Omega}_{TP}} \pi_i = 0$, certain units from $\mathbf{\Omega}_{TP}$ are excluded. $\mathbf{\Omega}_{IP}$ consist of properties offered on advertisement services, government web pages, brokers' websites, brokers' Intranet or postings on social media. $\mathbf{\Omega}_{IP}$ take into account all possible sources that are connected to the Internet. The creation mechanism of $\mathbf{\Omega}_{IP}$ is unknown and uncontrolled. In addition, we assume that the Internet population is characterised by overcoverage (e.g. false advertisements, non-existent properties).

The IDS population is denoted by $\mathbf{\Omega}_{IDSsP}$ and its size is equal to N_{IDSP}. $\mathbf{\Omega}_{IDSsP}$ consists of statistical units that fit the IDS definition, are offered on all Web pages (e.g. advertising services, local web pages). As in the case of the Internet population, we assume that $\mathbf{\Omega}_{IDSsP}$ has over- and undercoverage. For instance, a property offered on an advertisement service might refer to a different population or to a non-existent building/flat. $\mathbf{\Omega}_{IDSsP}$ is the result of *self-selection*

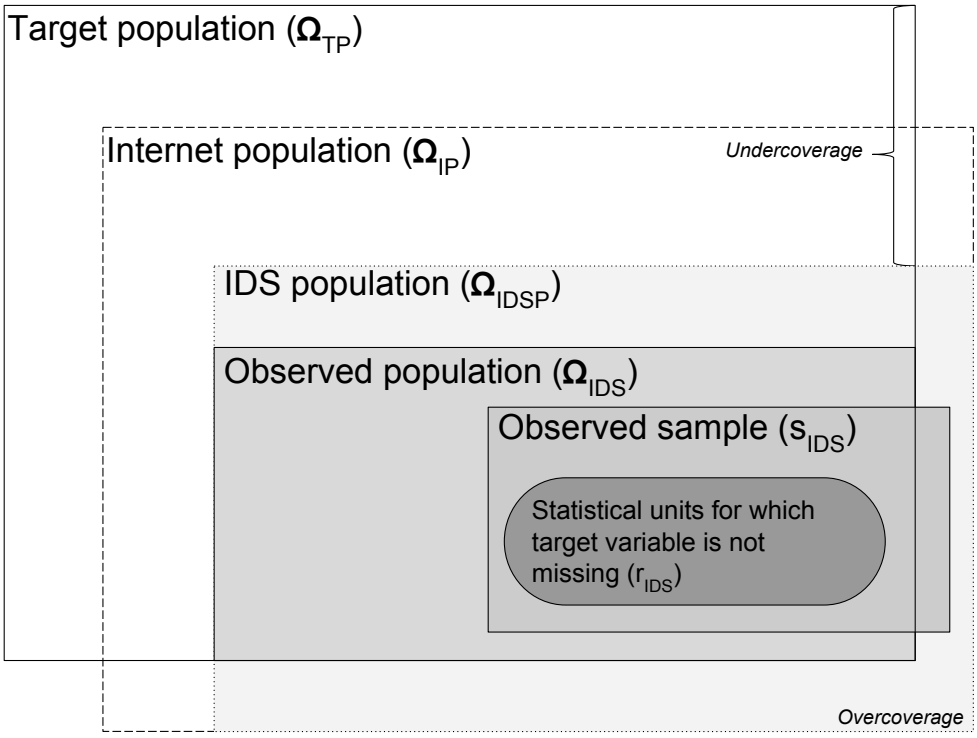


DIAGRAM 2.1: Relation between target and IDS population

(non-probabilistic) mechanism denoted as ρ. This mechanism can differ from the mechanism that generates $\mathbf{\Omega}_{IP}$. In addition, it should be taken into account that $\mathbf{\Omega}_{IP} = \mathbf{\Omega}_{IDSsP}$ (the same units of $\mathbf{\Omega}_{TP}$ are excluded), which rarely happens in practice and $\mathbf{\Omega}_{IP} \subset \mathbf{\Omega}_{IDSsP}$ is more probable. This means that the number of units excluded from $\mathbf{\Omega}_{TP}$ is higher for $\mathbf{\Omega}_{IDSsP}$ than for $\mathbf{\Omega}_{IP}$. Specific members of $\mathbf{\Omega}_{TP}$ may use $\mathbf{\Omega}_{IDSsP}$. For instance, brokers may present properties on Intranet while a certain subset is displayed on web portals.

In an ideal case *IDS population* consists of all IDSs available on the Internet, that is $\mathbf{\Omega}_{IDSsP} = \cup_{k=1}^{K} \mathbf{\Omega}_{IDSP,k}$ where K denotes the number of an IDS. However, in practice, $\mathbf{\Omega}_{IDSsP} \neq \cup_{k=1}^{K} \mathbf{\Omega}_{IDSP,k}$ is more common. In addition, $\exists_{k \neq l} \mathbf{\Omega}_{IDSP,k} \cap \mathbf{\Omega}_{IDSP,l} \neq \emptyset$ because unit i may use several IDSs and $\exists_{k \neq l} \rho_{i,k} \neq \rho_{i,l}$, where l is l-th IDS and k is k-th IDS.

In general, the IDS population and $\mathbf{\Omega}_{IDSP,k}$ could be associated with an *imperfect frame* or an *imperfect list* of all units from $\mathbf{\Omega}_{TP}$. We use the adjective *imperfect* to emphasise that specific units from $\mathbf{\Omega}_{TP}$ are observed in IDSs. This setting can also be associated with the concept of the *sampling frame*. However, the term *sampling frame* has not been used intentionally, because no sampling is conducted. $\mathbf{\Omega}_{IDSP,k}$ are the result of the self-selection process and in fact $\pi_{i,k}$ should be associated with propensity $\rho_{i,k}$ that a given unit i is used in k-th IDS. Hence, the term *self-selected frame* would be most appropriate. However, the term *frame* or *list* will be used for simplicity.

The treatment of IDSs as frames/lists indicates similarities with the problem of multiple, overlapping sampling frames. This problem is not new in survey

methodology, particularly in official statistics. However, research is mainly devoted to cases when the sample was already drawn ($\pi_{i,1}, \pi_{i,2}$ are known), the overlap between frames and samples is known and estimates are calculated based on two (or more) samples (Lohr and Rao, 2006; Lohr and Rao, 2000). In the IDS setting $\mathbf{\Omega}_{IDSP,k} \cap \mathbf{\Omega}_{IDSP,l}$ is often unknown, hence $\rho_{i,k}, \rho_{i,l}$ are also unknown. Moreover, $\mathbf{\Omega}_{IDSsP}$ is not ideal and often suffers from overcoverage ($\mathbf{\Omega}_{IDSsP} \setminus \mathbf{\Omega}_{TP} \neq \emptyset$). The causes of overcoverage will be presented later on in the chapter.

Finally, we distinguish the *observed population*, denoted by $\mathbf{\Omega}_{IDSs}$, and of size N_{IDS}. $\mathbf{\Omega}_{IDSs}$ consist of selected IDSs which owners have provided access to. For instance, it is rarely the case in practice that data from all existing IDSs are obtained; instead, one selects the biggest IDS for a given purpose (Facebook or Twitter, or the biggest (in terms of the number of advertisements) real estate market services, such as OtoDom.pl or Dom.Gratka.pl). The selection of IDSs for statistics will also be discussed later on in the chapter.

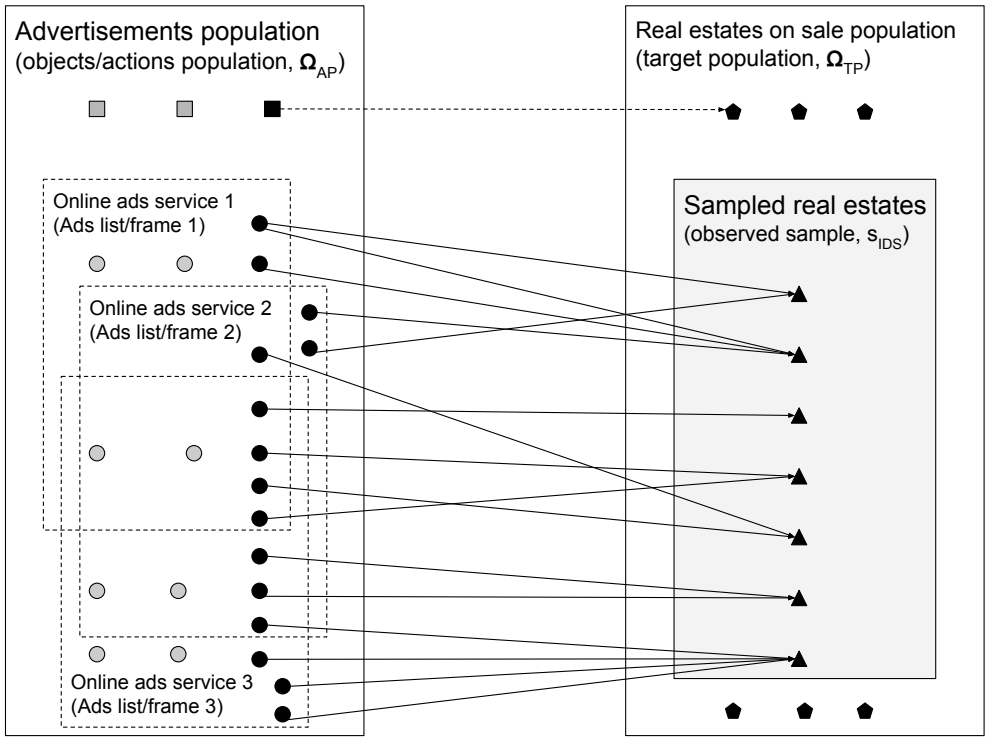
Nonetheless, it is very unlikely that all units from $\mathbf{\Omega}_{IDSs}$ can be accessed and only a *sample* of units is observed, which is denoted by s_{IDSs} of size n_{IDS}. For instance, Twitter allows access to only 1% of data without additional costs; other on-line services may limit the number of units visible on web-site or the data collection process via web-scraping may be interrupted. Despite the data collection process, this sample is in fact self-selected in the sense that units $\mathbf{\Omega}_{TP}$ decide whether to use a given IDS. Hence, to describe the observed sample we will use the term *self-selected* sample. The motivation for this is as follows:

- self-selection describes the non-probabilistic process which generates these data,
- each unit has a propensity measure ρ that indicates its willingness to use a given IDS (unit i decides whether to use a given IDS),
- using the term *self-selection sample* in the IDS context makes it similar to self-selection surveys and indicates possible estimation methods (we have discussed an IDS as a special form of an Internet survey in Chapter 1).

Finally, as in the case of surveys, IDSs contain statistical units that provided information on the target variable. In Diagram 2.1 r_{IDSs} denotes *statistical units that are members of s_{IDSs} and provided information on the target variable*. For instance, all residential properties that provided information on the offer price or the floor area will be members of r_{IDSs}. In sample surveys the concept of *respondent* is used in this situation. However, this concept refers to *a person who replies to something, especially one supplying information to a questionnaire or responding to an advertisement*. Because, the scope of the target population is not limited to persons, a more general term (*statistical unit*) will be used. The size of r_{IDSs} is $m_{IDS} \leq s_{IDSs}$; however, it is more likely to be $m_{IDS} \leq s_{IDSs}$. Hence, the final inference is based on m_{IDS} units that are members of $\mathbf{\Omega}_{TP}$. For the rest of $\mathbf{\Omega}_{TP}$ we do not have information on the target variable.

2.1.2 The population of advertisements

The previous section describes a simple case when statistical units are observed on-line. However, in practice the population observed in IDSs consists of non-statistical units, such as objects (e.g. advertisements, accounts) or actions (e.g. transactions, purchases, conversations). In the context of the real estate market, IDSs are often advertising services containing ads that refer to statistical units. In this setting *the population of advertisements* is denoted as $\mathbf{\Omega}_{AP}$. Diagram 2.2 depicts a hypothetical relation between $\mathbf{\Omega}_{AP}$ and $\mathbf{\Omega}_{TP}$. Dashed lines denote on-line ads services (multiple overlapping frames) that include out-of-scope units (overcoverage). It means that not all objects from $\mathbf{\Omega}_{AP}$ are observed. Black circles denote ads referring to statistical units (residential properties), while grey circles represent ads that refer to a different population (e.g. rentals). Gray circles refer to cases where overcoverage is observed.



Note: Squares (□) and circles (○) refer to advertisements (objects) that are members of $\mathbf{\Omega}_{AP}$. Triangles (△) and pentagons (⬠) refer to properties for sale (statistical units) that are members of the target population $\mathbf{\Omega}_{TP}$. Black shapes in $\mathbf{\Omega}_{AP}$ denote objects that are connected with units from $\mathbf{\Omega}_{TP}$. Grey shapes in $\mathbf{\Omega}_{AP}$ denote objects that are not connected with $\mathbf{\Omega}_{TP}$.

DIAGRAM 2.2: The relation between objects/actions and the target population

Connections between $\mathbf{\Omega}_{AP}$ and $\mathbf{\Omega}_{TP}$ can be many-to-many. Several advertisements may refer to one statistical unit and one advertisement may include

multiple statistical units. This is the result of brokers' or owners' on-line activities. Several brokers may advertise one property under an open agreement with the owner. However, detection of connections between these two populations may not be straightforward. An example is given in Diagram 2.1 based on OtoDom.pl service. Diagram 2.1 illustrates the case when 7 advertisements refer to one residential property; however, the ads differ in their details about the location, price, the number of floors or photographs.

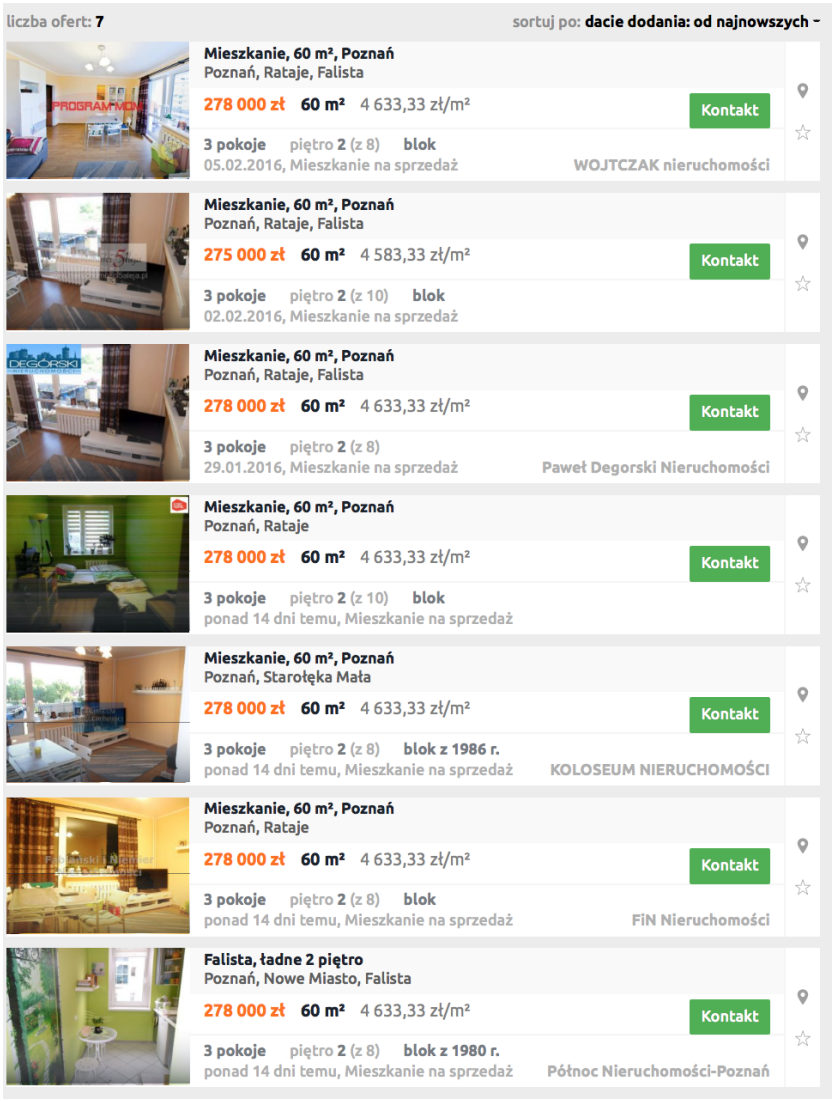


FIGURE 2.1: An example of several advertisements referring to the same property.

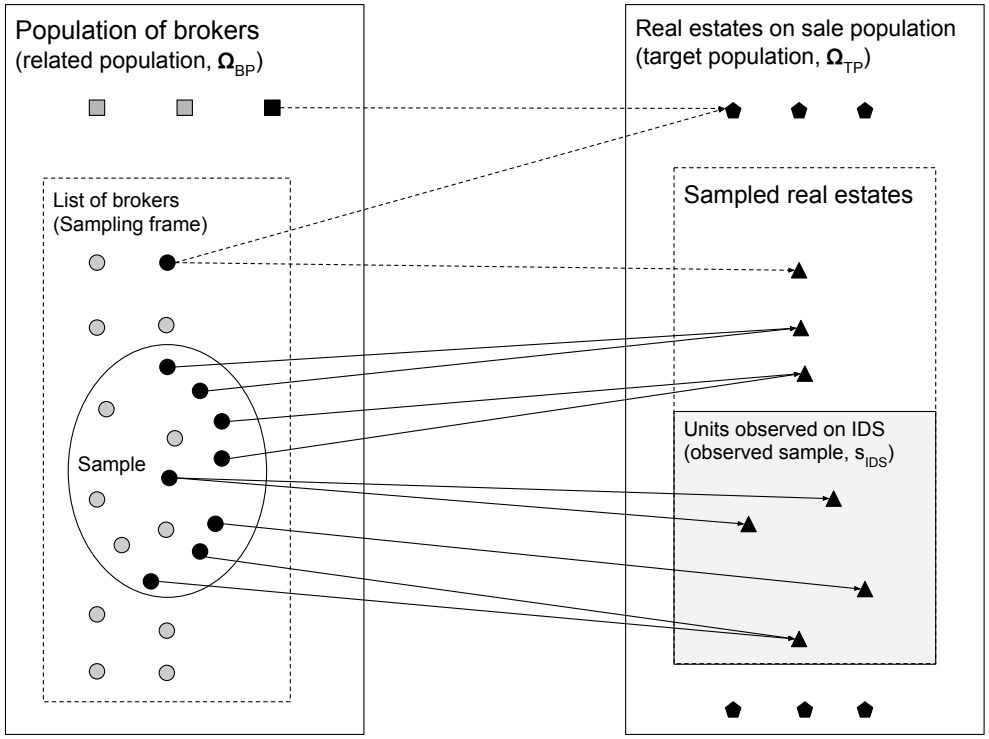
Note: Based on OtoDom advertisement service. Accessed on 08.02.2016. Link to query: `http://goo.gl/gk3KPW`.

Finally, when the network of links between objects and statistical units is established, an observed sample s_{IDS_s} is derived. This sample is denoted by the grey rectangle in Diagram 2.2. Statistical units that were identified in the linkage process are represented by black triangles. The sample s_{IDS_s} may suffer

from undercoverage due to the self-selection mechanism that is responsible for the creation of $\mathbf{\Omega}_{IDSs}$. The causes of undercoverage will be discussed later in this chapter.

2.1.3 The population of brokers

The population of advertisements is not the only population observed in the secondary real estate market. For instance, properties are offered by owners, brokers or government institutions (which generate supply on the market). However, it is unlikely that a list of individuals (private owners) that offer properties for sale exists. In fact, it would be easier to obtain information about statistical units from real estate brokers. Because brokers are required to register business activities, a list of brokers is available. Diagram 2.3 presents a hypothetical linkage scenario for the population of brokers, denoted as $\mathbf{\Omega}_{BP}$ and the general population $\mathbf{\Omega}_{TP}$. Naturally, the list of brokers may be incomplete (e.g. due to delays in registration) and may contain out-of-scope individuals (not all brokers operate in the secondary residential market).



Note: Squares (□) and circles (○) refer to brokers that are members of $\mathbf{\Omega}_{BP}$. Triangles (△) and pentagons (⬠) refer to properties for sale (statistical units) that are members of the target population $\mathbf{\Omega}_{TP}$. Black shapes in $\mathbf{\Omega}_{BP}$ denote objects that are connected with units from $\mathbf{\Omega}_{TP}$. Grey shapes in $\mathbf{\Omega}_{BP}$ denote objects that are not connected with $\mathbf{\Omega}_{TP}$.

DIAGRAM 2.3: The relation between the population of brokers and the IDS population

In practice, not all units from the sampling frame can be contacted; rather a sample is drawn (denoted as a solid circle in Diagram 2.3). Connections between brokers and statistical units can be many-to-many. One broker can offer several different properties and several brokers can offer the same property. When a sample of brokers is drawn, a sample of properties may include both those observed in IDSs (offered on the Internet) and those observed outside IDSs. For instance, brokers may provide information about properties that are not offered via advertisement services or web pages. This type of survey may help to verify the selection mechanism responsible for the creation of Ω_{IDSs}.

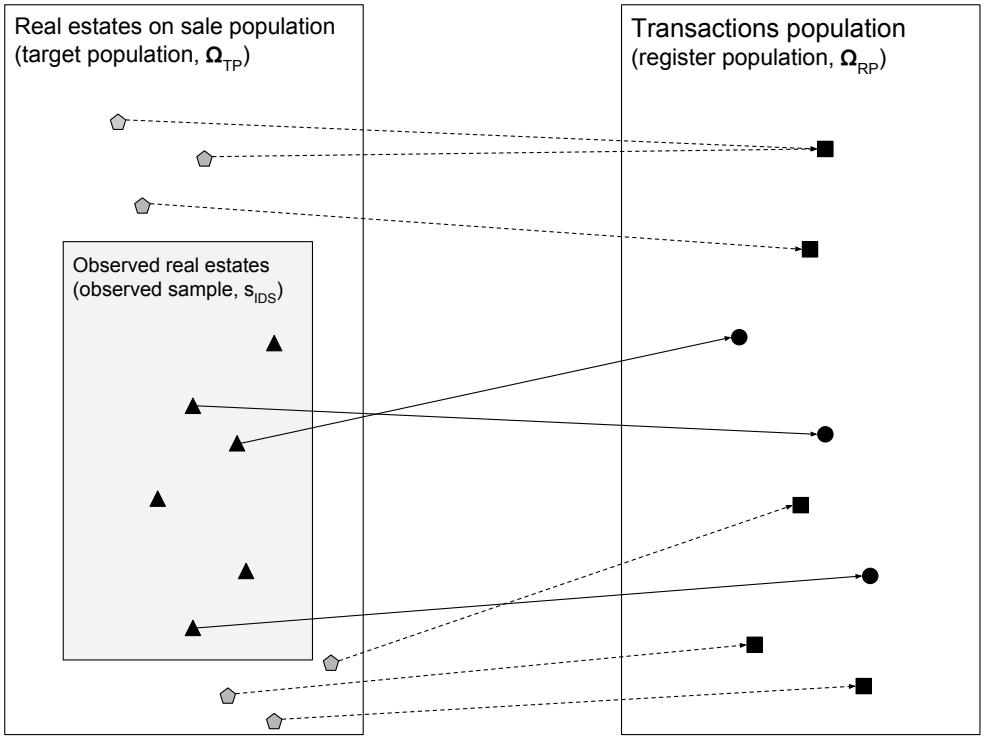
In Poland, NBP/CSO conduct a survey of brokers, which provides information about the secondary real estate market. However, it does not contain information on which units are observed online and which offline. The main motivation for using IDSs is to avoid surveying entities or individuals. Instead, existing data sources should be used to verify differences between IDS and non-IDS population.

2.1.4 The register population

One existing source of information about the real estate market is the register of transactions. It represents another population that is related to the general population – the *register population*. In Poland, each sold property is registered in the *Register of Real Estate Prices and Values*, which is maintained by local authorities at LAU1 level. More information about the register will be provided later in the chapter. The register population consists of actions, in particular transactions. One transaction may involve the sale of multiple different properties. A hypothetical connection between the general population and the register population (Ω_{RP}) is presented in Diagram 2.4. In Diagram 2.4 the IDS population is shown on the left and the register population is shown on the right, with lines indicating transitions between these populations. The observed sample s_{IDSs} is obtained from IDSs and includes sale offers for units that will not be sold (not observed in Ω_{RP}). Transactions denoted by black circles refer to sold properties that were members of the IDS population, while black rectangles refer to those that were not observed online.

There may be different linkage scenarios for s_{IDSs} and Ω_{RP}. For one thing, the period between the placement of offers and transactions is often unknown. More importantly, though, in both data sources the same information should exist in order for them to be linked. However, the use of data that is already available is desirable. Linkage with the register population may provide insights into differences between properties offered online (links) and offline (non-links).

All in all, identification of different populations in the real estate market will be crucial for detecting potential data sources. Linkage scenarios for these populations can provide information about selection mechanisms. The description of sources and their potential usefulness for the verification of the quality of IDS will be presented later in this chapter.



Note: Squares (□) and circles (○) refer to members of $\mathbf{\Omega}_{TP}$. Triangles (△) and pentagons (⬠) refer to properties for sale (statistical units) that are members of the target population $\mathbf{\Omega}_{TP}$. Black shapes in $\mathbf{\Omega}_{BP}$ denote objects that are connected with units from $\mathbf{\Omega}_{TP}$.

DIAGRAM 2.4: The relation between the register and the IDS population

2.2 Data sources about the real estate market in Poland

This section contains a detailed description of the existing statistical, non-statistical and Internet data sources about real estate. However, for the sake of clarity it should be noted how the Polish real estate market is organized. Excluding buyers and tenants, market participants include 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 surveys – relations between properties for sale and owner/broker can 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.

Taking the above into account and due to the scope of the thesis (secondary real estate market) only sources that are devoted to this target population will be discussed. In the section devoted to statistical data sources, surveys conducted by NBP/CSO and REGON statistical register will be discussed. The last round of census, the 2011 Polish Census of Population and Housing, which was partially based on registers, will not be included. CSO conducts two major surveys of the real estate market: Municipal Infrastructure (Pol. *infrastruktura komunalna*) and Housing Stock (Pol. *zasoby mieszkaniowe*), which are based on reports delivered by gminas. However, due to their scope, the CSO surveys will also be excluded from the detailed description.

The part devoted to non-statistical data sources will cover the *Register of Real Estate Prices and Values*, which contains transactions in the primary and secondary market. For simplicity, the name *Register of Transactions* will be used hereafter. Other registers that cover the real estate market are Mortgage Registers (Pol. *Księgi Wieczyste*) and Land and Buildings register (Pol. *Ewidencja Gruntów i Budynków*). These registers contain information on characteristics of buildings and land that are partially used in the Register of Transactions. Other registers, which indirectly refer to legal persons that operate on the market, such as tax registers (e.g. VAT or NIP, Pol. *Numer Identyfikacji Podatkowej*) or business registers (CEIDG, Pol. *Centralna Ewidencja i Informacja o Działalności Gospodarczej*) will only be described briefly. In addition, the Register of Agencies and Brokers lists brokers who operate in the real estate market. However, this register is created and maintained by private associations, such as the Polish Real Estate Federation[1] or The National Chamber of Real Estate Management, and registration is voluntary. Hence, the coverage of this register may not be sufficient to be used for statistics.

The last group consists of selected Internet data sources, in particular those used for the study described in this thesis – OtoDom.pl, Dom.Gratka.pl and Nieruchomosci-Online.pl. The selection was based on the popularity of these services and non-official surveys of brokers [2]. Nieruchomosci-Online.pl was also selected due to the availability of historic advertisements without additional costs.

2.2.1 Selected statistical data sources

Surveys conducted by NBP/CSO

In Poland there is only one (official) survey devoted to the secondary real estate market – residential and commercial property prices (code 1.26.09(079)[3]). The survey is conducted by the National Bank of Poland (NBP) in co-operation

[1]For more information on the Polish Real Estate Federation, see `https://rejestr.pfrn.pl`.

[2]The survey was conducted by the PBS company and commissioned by OtoDom. The main aim was to determine what web services brokers use and how they use them. The methodology as well as the sample size was not reported. The link to the article (in Polish) `http://blog.otodom.pl/2015/05/po-pierwsze-skutecznosc-po-drugie.html`.

[3]The Survey Program of Official Statistics, see `http://bip.stat.gov.pl/dzialalnosc-statystyki-publicznej/program-badan-statystycznych/`.

with the Central Statistical Office (CSO). Since NBP is mainly responsible for the analysis, the report mostly covers aspects connected with macroeconomic analysis at the country and city level.

The main goal of the survey is to provide information about the price dynamics of residential properties (buildings, premises, land) and rents. The scope of the survey includes offers and transaction prices including characteristics of properties (for hedonistic indexes) that refer to premises, houses and land that were offered and sold in the primary and secondary market. The survey is limited only to major cities in Poland and their agglomerations (17 cities in total). In addition, the survey covers rent prices and characteristics of commercial properties. Information in the survey is provided by entities making housing and commercial investments (developers, housing associations), entities that operate in the residential and commercial market (brokers and associations of brokers, housing co-operatives, property managers and owners, consulting enterprises), entities that are responsible for the Register of Prices (local authorities, offices of district governors), tax authorities that issue ownership transfer agreements to real estate, cooperative ownership rights to premises etc. and tax offices.

The survey consists of five questionnaires:

- The questionnaire about the residential real estate market in the primary market (form code NBP NM/RP) – electronic form; filled out by developers, housing co-operatives and tax offices,
- The questionnaire about the residential real estate market in the secondary market (form code NBP NM/RW) – electronic form; filled out by brokers and brokers associations, local authorities and tax offices,
- The questionnaire about the commercial real estate market (form code NBP NK/B) – scope: offices; electronic form; filled out by brokers and brokers associations, consulting companies, property managers and owners,
- The questionnaire about the commercial real estate market (form code NBP NK/H) – scope: retail centres; electronic form; filled out by brokers and brokers associations, consulting companies, property managers and owners,
- The questionnaire about the commercial real estate market (form code NBP NK/M) – scope: warehouses; electronic form; filled out by brokers and brokers associations, consulting companies, property managers and owners.

In addition, the survey draws on data from the Register of Transactions. Each survey is conducted separately by local branches of NBP. 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 survey is not clear. For instance, there is no information on the quality and response rate. However for the purpose of the study, information about the sample size was provided and will be discussed in the fourth chapter.

The result of the survey consists of: (1) the level and dynamics of the average price of premises, houses and land according to the market type (primary

and secondary), the type of price (offer and transaction) and city and (2) the level and dynamics of average rents, vacancy rates, capitalization rates for each commercial property (office, retail and warehouse). NBP/CSO publish two reports that summarise the resulting information on a quarterly (NBP, 2014b, 2015b) and yearly basis (NBP, 2014a, 2015a). The yearly report contains point estimates and hedonistic indexes for 17 biggest Polish cities aggregated at the LAU1 level and is based on a survey of brokers, non-official data and the Register of Transactions[4]. The quarterly report is limited to average prices and indexes for the 17 cities. However, the annual report is produced and published with a delay, for instance information for 2014 was available at the end of 2015, which means that the information is delayed and does not reflect the current state of the real estate market.

REGON

The National Official Business Register (REGON) is a register created and maintained by CSO. REGON was established by article 41 paragraph 1 item 1 of the Law dated June 29th, 1995 on official statistics (Journal of Laws No 88, position 439, with amendments). Detailed rules for maintaining and updating the register are set out in the Regulation of the Council of Ministers of July 27th, 1999 on the mode and methodology of maintaining and updating the business register, including application, questionnaire and certificate templates and detailed conditions and mode of cooperation of official statistical services with other bodies maintaining official registers and information systems in the public administration (CSO, 2015a).

The REGON register is a continuously updated set of information on subjects of the national economy maintained as an IT system in the form of a central database and local databases. The register is used (CSO, 2015a):

- to ensure identification consistency of businesses entered into other official registers and information systems of the public administration,
- to ensure the uniformity of descriptions used in the nomenclature and classification concepts in all official registers and information systems of the public administration,
- to provide general characteristics of businesses operating in the national economy broken down by territorial unit, proprietor, type of activity, legal form, etc.,
- to facilitate the preparation of an address list of active businesses,
- to provide the basis for the creation of databases and data banks on businesses,
- as the main source of information for the sampling frame used for statistical surveys.

[4]The survey covers 16 province capitals (Bydgoszcz, Gdańsk, Katowice, Kielce, Kraków, Lublin, Łódź, Olsztyn, Opole, Poznań, Rzeszów, Szczecin, Warszawa, Wrocław, Zielona Góra) and Gdynia. The location of these cities is shown in Appendix A.3.

From the viepoint of real estate market research it is important that registration in REGON is mandatory for: (1) legal persons, (2) organisational units without the status of a legal person, (3) natural persons conducting economic activities (including private farms), and (4) local units of entities mentioned above. Activity performed, including the principal kind of activity, legal form and form of ownership, name and address of the head office is provided during the registration process. Finally, entities are classified into economic activity sections defined by NACE rev. 2, in this case - real estate activities (L). The structure of this section is presented below.

68 Real estate activities

- 68.1 Buying and selling of own real estate
- 68.2 Renting and operating of own or leased real estate
- 68.3 Real estate activities on a fee or contract basis
  - 68.31 Real estate agencies
  - 68.32 Management of real estate on a fee or contract basis

However, in 2013 The Ministry of Justice began the process of deregulation involving several occupations, including the job of a real estate market agent and broker. Changes in the law came into effect on the 1^{st} of January 2014: as a result, all previously existing formal requirements have been eliminated, including licenses and entry obligation in the list of real estate agents or property managers[5]. 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). Hence, REGON has become the main source of information about one group of participants in the real estate market.

Nonetheless, it should be noted that residential properties in the secondary market may be offered and sold not only by agents but also by owners. Owners are not obligated to register or inform the authorities about their intention to sell; similarly, notarial acts do not contain any information about how a given property was sold. However, a vast majority of properties put up for sale are offered by agents. The reason for that their experience in sales and handling administrative issues connected with transactions.

2.2.2 Selected non-statistical data sources

The Register of Transactions (the full name is The Register of Real Estate Prices and Values) was created on the basis of Dz.U. [Journal of Laws] No. 38, item 454, as amended (2001). It is a public register held by a district governor (Pol. starosta), which includes data about real estate prices provided in notarial deeds and values of real estate quoted by real estate appraisers in appraisal reports, whose extracts are provided to the register of land and buildings. The Register of Transactions is used for several survey programs conducted by CSO, for instance the Real Estate Turnover Survey (1.26.04(076), published on a yearly

[5] See https://www.ms.gov.pl/pl/deregulacja-dostepu-do-zawodow/i-transza/lista-zawodow/.

basis), Residential and Commercial Property Prices Survey (1.26.09(079), published on a quarterly basis), The diversity of the level and dynamics of regional development (1.70.01(248), published on a yearly basis), The statistical information system in rural areas (1.70.03(250), published on a yearly basis) or Urban Audit (1.70.04(251), published on a yearly basis). Since 2015 CSO has been legally obliged to publish quarterly price indexes at NUTS2 level (Dz.U. [Journal of Laws] No. 985, 2015). The information on prices is also obtained from the Register of Transactions.

Notaries are obliged to inform district governors about each ownership change (section I and II in the Mortgate Register) within 14 days of the date when documents take effect. Then, district governors have 30 days to enter transaction details into the Register of Transactions. Finally, CSO is obliged to publish price indexes within 4 months of the end of a given quarter. The indexes are based on the 380 districts (Pol. powiat) that report information.

According to the Dz.U. [Journal of Laws] No. 38, item 454, as amended (2001) the following properties are registered:

- undeveloped agricultural parcel for one purpose (Pol. niezabudowana jednoużytkowa rolna),
- undeveloped agricultural parcel for multiple purposes (Pol. niezabudowana rolna wieloużytkowa),
- undeveloped parcel for purposes other than farmstead development (Pol. niezabudowana przeznaczona pod zabudowę inną niż zagrodowa),
- developed parcel with buildings or residential buildings (Pol. zabudowana budynkiem lub budynkami mieszkalnymi),
- developed parcel with buildings for other purposes (Pol. zabudowana budynkami pełniącymi inne funkcje),
- a building on land in perpetual lease (Pol. budynkowa),
- part of a building constituting a separate property (Pol. lokalowa),
- developed forest parcel (Pol. zabudowana leśna),
- another type of real estate (Pol. inna nieruchomość).

The register population is limited to fully-owned properties. Therefore, cooperative ownership right to an apartment is not covered by the register. In addition, it should be noted that the Register of Transactions contains items (in particular transactions) that can involve many properties. For instance in one transaction it is possible to buy several properties (dwellings and garage) and land parcels (undeveloped). Depending on the condition of the real estate market, properties offered (and sold) in the primary market can then be put up for sale again in the secondary market or one property may be resold multiple times. This should be taken into account when performing data cleaning procedures.

Not all properties are within the scope of the study, which is limited to residential and non-residential properties, which are termed 'lokal' in Polish [6]. Official statistics defines 'lokal'/ as *a room or a suite of rooms separated by durable walls within a building dedicated to the permanent stay of people, which, together with auxiliary rooms serve the purpose of fulfilling their housing needs or which are used according to their dedication for purposes other than residential purposes (garages)*[7]. The definition is taken from the Apartment Ownership Act (Dz.U. [Journal of Laws] No. 85, item 388, as amended, 1994). The Register of Transactions contains the following information about transactions involving apartments:

- ID of the transaction,
- Total value of the transaction and its components,
- ID of the apartment,
- ID in the Mortgage register,
- Total value of the apartment,
- Sequential number of the apartment in the building,
- Type of apartment – residential, non-residential (e.g. garage),
- Number of rooms in the apartment (the CSO definition of room: *A room is a space in a dwelling, separated from other spaces by permanent walls from the floor to the ceiling, with an area of at least 4 square metres, with direct day lighting, i.e. with a window or a French window in the external wall of the building; kitchens are also considered rooms as long as they fulfill the above criteria*[8]),
- Number and types of spaces belonging to the apartment,
- Number of the floor the apartment is located on,
- Usable floor area (CSO definition of usable area: *The total usable area in an apartment or in a residential building containing a single dwelling, i.e. rooms, a kitchen, larders, anterooms, alcoves, halls, corridors, bathrooms, toilets, enclosed veranda, porch, dressing room and other spaces for residential and utility needs of dwellers regardless of their designation and use (among others: ateliers, recreation facilities etc)*[9]),
- Usable floor area of spaces belonging to the apartment,

[6] Because the English terminology does not have a single equivalent term, the term 'apartment' will be used for the sake of simplicity, and will be understood as including garages.

[7] See http://stat.gov.pl/en/metainformations/glossary/terms-used-in-official-statistics/1984,term.html.

[8] See http://stat.gov.pl/en/metainformations/glossary/terms-used-in-official-statistics/133,term.html.

[9] See http://stat.gov.pl/en/metainformations/glossary/terms-used-in-official-statistics/324,term.html.

- Address of the apartment including: (1) city and TERYT ID, (2) city district, (3) street name and TERYT ID, (4) building number and (5) apartment number,
- Information about spaces belonging to the apartment concerns: (1) type and usable floor area, (2) the building where the space is located.

The quality of the Register of Transactions is unknown. For instance, in its report about Real Estate Turnover CSO publishes information only about missing data in the number of rooms, which was 1.8% in 2014 and 2.6% in 2013. Quality measures of other variables included in the register are not published. Nonetheless, because the it includes all transactions involving fully-owned propeties, it can be linked with IDS data.

2.3 Selected Internet data sources

IDSs typically consist of many data sources covering the same topic. In the case of the real estate market there are several competing advertisement services. However, there is no official survey or other kind of study about the use of online services by real estate market participants in Poland. In addition, the selection of IDS has not been discussed in the statistical context. Hence, we propose the following issues that should be discussed prior to the selection of IDSs.

Expert knowledge experts in a given field, in particular real estate, can indicate which web services should be considered for statistics. The decision could consulted with brokers or buyers. The most popular services used by brokers between April 2014 and April 2015 are shown in Figure 2.2. The results are based on a survey of brokers conducted by the *Pracownię Badań Społecznych* (PBS) on and commissioned by OtoDom. The main aim was to determine what web services brokers use and how they use them. Brokers awnsered a multpile choice question: "Which of the following advertisement services do you use as a broker?[10]". The total sample size was 268 brokers; however, no further details about the methodology were provided[11]. According to the survey OtoDom and Dom.Gratka.pl are the most popular websites, Nieruchomosci-Online.pl was used by over 50% of brokers in the survey. The report also stated that almost 80% of brokers used from 4 to 9 web services.

The number of visitors (market share) due to the lack of official statistics on the use of advertisement services, only non-official statistics can be used in the evaluation. However, in most cases a detailed methodology of non-official surveys is often unavailable. Hence, non-official statistics cannot be treated as unbiased estimates and used as trends reflecting the popularity of a given service.

[10] In Polish the question was stated as follows: *Z usług których serwisów ogłoszeniowych korzysta Pan(i) jako pracownik biura nieruchomości?*

[11] Link to the article (in Polish) `http://blog.otodom.pl/2015/05/po-pierwsze-skutecznosc-po-drugie.html`.

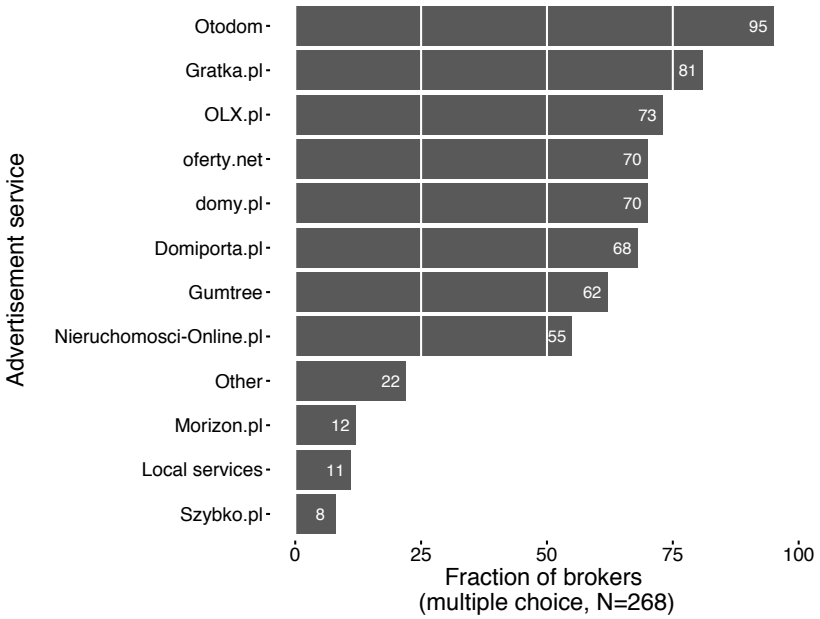


FIGURE 2.2: Services used by real estate brokers between 04-2014 and 04-2015

The main non-official survey in Poland devoted to the use of the Internet is *Megapanel PBI/Gemius* (MPG survey)[12]. The MPG survey measures the popularity of web services based on a web-panel sample. It consists of three components: (1) *site-centric data* – raw data on traffic and data use of all Internet users who visit web sites participating in the MPG survey; (2) Pop-up survey – a random sample of users who visit web sites where the pop-up survey is conducted; (3) Online panel – pop-up surveys encourage users to take part in the online panel survey. Each user who agrees to participate installs on their computer a special program (*netPanel*) that collects data on the Internet use. According to MPG, several thousands users take part in the panel survey and its structure is consistent with the Internet population; (4) *CAPI survey* (NetTrack[13]) which is conducted by MillwardBrown. The survey provides information about the Internet population and its characteristics. It is a monthly CAPI survey on a sample drawn from PESEL register. The reference population is defined as persons between 15 and 75 years. According to the NetTrack web-site the sample consists of 48 000 respondents each year. However, data from the MPG survey is made available only upon prior registration[14]. The following measures defined for the MPG survey could be used to select IDS for statistics[15]:

- Visitors (Real Users, RU) – The number of Internet Users (visitors) in a given target group who visited (generated at least one page view) the selected node(s) in a specified time period. This indicator relates to the actual number of persons – not computers, cookies or IP addresses.

[12] See `http://www.audience.gemius.pl`.

[13] For more information, see `http://www.millwardbrown.com/subsites/poland/services/syndicated/net-track`.

[14] For more details, see `https://audience.gemius.com/en/research-results/poland/`.

[15] For other measures, see: `https://audience.gemius.com/en/methodology/metrics/`.

- Page views – The number of times a web page was requested by a visitor (of a selected target group during a specific time period).
- Page view share – The ratio of page views generated by visitors to all pages in a specified category of web pages. The ratio is calculated for a specific period and specific target groups. This indicator is expressed as a percentage.
- Internet-Reach – The ratio of the number of visitors that visited a specific web page to the number of Internet users in a given month. The ratio is also calculated for specific target groups. This indicator is expressed as a percentage.
- Audience share – The ratio of the number of visitors that visited a website to the number of visitors in the category which the page belongs to. The ratio is calculated for a specific period and specific target groups. This indicator is expressed as a percentage.

MPG survey results are published on a monthly basis for several website categories including *construction and real estate*[16]. MPG results for the construction and real estate category in October 2015 are presented in Table 2.1. The results are aggregated according to the ownership status.

Grupa Allegro[17], the owner of advertisement services OtoDom.pl and Olx.pl, had the highest number of visitors and reach-Internet share in October 2015. It should be noted that OtoDom.pl specializes in real estate and OLX.pl is a classified ads service. The second highest is Grupa ZPR Media[18], the owner of several online advice platforms, for instance Murator.pl or Urzadzamy.pl. These services are out of the scope of this thesis. The third highest ranking company is Grupa Polska Press[19], which owns several newspapers as well as on-line services such as Dom.Gratka.pl, regioDOM.pl or e-budownictwo.pl. Ranked fourth is Grupa Gazeta.pl, the owner of DomiPorta.pl and gazetadom.pl. The top 9 places are occupied by the biggest e-commerce and information companies in Poland. The tenth place is occupied by Trovit.pl, a search engine for classified ads (for instance jobs, cars, real estate, products and rentals). This portal aggregates information from other web services mentioned earlier and operates in 43 countries[20].

Number of advertisements another measure that could be considered in the selection of data sources is the number of advertisements. A vast majority of web portals provide information about the number of ads available to their users, which varies between web portals. However, this number is certainly biased because multiple advertisements may refer to the same property, or may include

[16]Other categories are: Business, Finance, Law; Construction & Real Estate; E-commerce; Education; Erotica; Corporate; Hosting; Information & journalism; Communication; Culture and entertainment; Maps and locators; Automotive; New technologies; job; public; communities; Sport; Lifestyle; Tourism.

[17]See `https://allegro.pl/`.

[18]See `http://www.grupazpr.pl/`.

[19]See `http://polskapress.pl/pl`.

[20]Accessed 2015-11-20.

TABLE 2.1: Results for the category Construction and Real Estate based on the Megapanel PBI/Gemius survey in October 2015

NR	IDS owner	Number of Visitors	Reach-Internet
1	Grupa Allegro	3 078 413	12.5%
2	Grupa ZPR Media	2 744 231	11.1%
3	Grupa Polska Press	1 907 160	7.7%
4	Grupa Gazeta.p	1 829 562	7.4%
5	Grupa Wirtualna Polska	1 707 002	6.9%
6	Grupa Budujemydom.pl - AVT Korporacja	965 642	3.9%
7	Grupa Onet - RASP - Dom	670 872	2.7%
8	Grupa Interia.pl	650 407	2.6%
9	Grupa Morizon - Melog.com	594 102	2.4%
10	trovit.pl - mieszkania.trovit.pl	468 148	1.9%
11	Grupa Świat Kwiatów	452 643	1.8%
12	poradnikogrodniczy.pl	413 450	1.7%
13	gumtree.pl-gumtree.pl	406 113	1.7%
14	nieruchomosci-online.pl	379 807	1.5%
15	klikmapa.pl	338 199	1.4%
16	home-you.com	331 470	1.4%
17	meble.pl-meble.pl	288 583	1.2%
18	dom.pl	255 010	1.0%
19	Grupa Szybko.pl	245 205	1.0%
20	skyscrapercity.com	186 671	0.8%

Note: based on data from Megapanel PBI/Gemius, `http://www.audience.gemius.pl`. The sample size was N = 7 286. The target population: all Internet users over 7 years old. Source for the table `https://www.gemius.pl/wszystkie-artykuly-aktualnosci/wyniki-badania-megapanel-pbigemius-za-wrzesien-2015-2558.html`.

false or outdated ads. The numbers of advertisements of residential properties on selected web services are presented below[21]. The figures are comparable, but only Nieruchomoci-online.pl granted access to archived data through web-service. The smallest number of residential property ads is found on OLX.pl. However, this portal is not limited to real estate listings but publishes ads in other categories.

- OtoDom.pl - 347 285
- OLX.pl - 50 256
- Dom.Gratka.pl - 341 860
- Nieruchomoci-online.pl - 250 140 (with archived 3 038 720)
- Domiporta.pl - 323 205
- Szybko.pl - 348 055

[21] Accessed 2015-11-20.

- Morizon.pl - 344 321

Popularity/search index Google[22] provides the Google Trends service that can be used as a proxy of popularity of a given online service. The methodology behind this index is unknown but it is based on queries made by Google Search users. However, Google Trends may not be the best solution because users tend to enter search terms such as *flats for sale*, *flats to buy* rather than write names specific online services.

Taking into account the above considerations, the following services were selected: OtoDom.pl and Dom.Gratka pl (owing to their popularity) and Nieruchomosci-Online.pl (because of the direct access to archived advertisements). Screenshots of the selected services are presented in Figure 2.3.

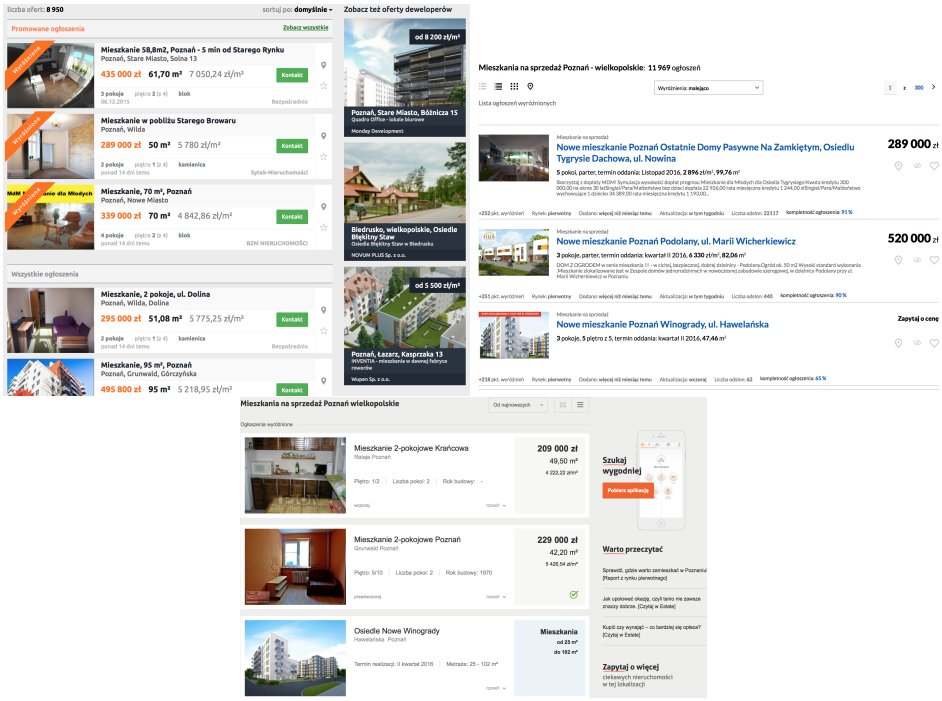


FIGURE 2.3: Real estate market web services

Note: OtoDom.pl (top left), Dom.Gratka.pl (top right) and Nieruchomosci-Online.pl (bottom).

[22]Since 2015-08-11 Google has been part of a multinational conglomerate Alphabet Inc. which was founded by Google's co-founders (Larry Page and Sergey Brin). However, in the dissertation the name Google, instead of Alphabet, will be used. The reason is that Google offers Google Trends or Google Search Engine.

2.4 Integration of Internet data sources into the system of real estate market statistics

This section is devoted to the integration of the above mentioned data sources to provide a more complete information about the real estate market. In particular, a statistical system consisting of registers, surveys and IDSs will be proposed. Several proposals as to how these sources could be linked will also be made. Finally, the pros and cons of using IDSs as the main source for statistics will be discussed.

The main goal of integrating IDSs into the statistical system is to provide more detailed estimates about the real estate market in Poland. Such a solution is mainly motivated by the lack of a systematic approach to providing statistics for this market and to improve timeliness. Another motivation is to provide more detailed characteristics about the primary and secondary real estate market at city level. The proposed integration is intended to bridge this gap.

A proposal for the general integration of registers (e.g. The Register of Transactions, Mortgage registers or TERYT), survey data conducted by NBP/CSO and IDS data is presented in Diagram 2.5. The main assumption behind the proposed approach concerns the availability of multiple sources for each type of data. The first stage of the process involves harmonization and the detection of errors within each data source (register micro-data, survey micro-data and IDS micro-data). For instance, data from the Register of Transactions are made available in a standardized (Pol. *Standard Wymiany Danych Ewidencyjnych*) format, which has been developed to represent spatial objects and provide the description of land and buildings in a flat text file. This format was designed for the Polish system and enables data integration with other real estate registers. In Poland registers for the real estate market are already being integrated. The Head Office of Land Surveying and Cartography coordinates the Integrated Real Estate Information System (IREIS) project[23]. The system should be available by the end of 2018. However, according to the information given in the IREIS' reports, the Register of Transactions will not be integrated with this system. Hence, it is still necessary to connect information about values of properties with geodetic information.

The main source for survey data is the survey conducted by NBP/CSO, which covers both the primary and secondary market. This also includes data collected by the CSO, for instance from reporting made by local authorities. This will result in the creation of a set of *integrated survey micro-data*, consisting of records with statistical units.

The third group are IDS micro-data obtained from selected web services. In Poland such data can come from the biggest online services, such as OtoDom.pl, Dom.Gratka.pl, Domy.pl, Domiporta.pl or Nieruchomosci-Online.pl. The structure of each online service is different and should be standardized during the integration process . Moreover, this step also includes matching objects to statistical units. Possible methods of record linkage will be presented later on in the dissertation. Information obtained from IDSs can include the offer price as well as other characteristics provided by persons that have uploaded their advertisements.

[23] See `http://www.gugik.gov.pl/__data/assets/pdf_file/0018/23166/ZSIN-Faza-II_prezentacja-publiczna_v5.pdf`.

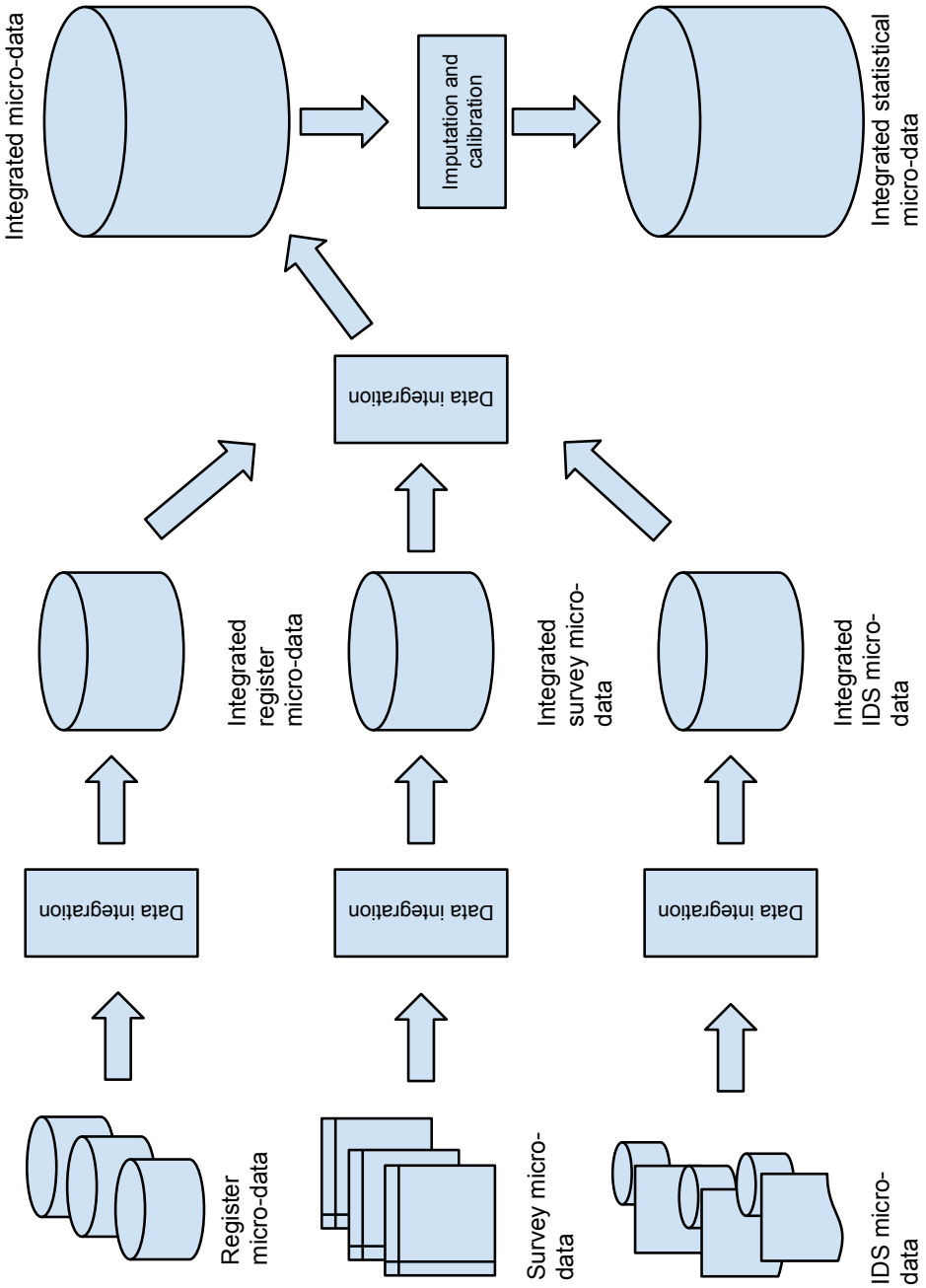


DIAGRAM 2.5: Internet data sources in the statistical system – a general perspective

Finally, all three types of data sources are integrated into one data base denoted as *integrated micro-data*. Diagram 2.6 presents hypothetical linkage scenarios for register, IDS and survey data. In Diagram 2.6 we assume that the main source is the register containing all apartments, buildings and land parcels which IDS and survey data are linked to. However, it should be remembered that in some cases linkage may not be possible (for instance, lack of common linking variables). In Diagram 2.6 four cases are considered:

1. It is possible to link all IDS and survey data to the register and linkage between IDS and survey data is possible (top left on Diagram 2.6);
2. IDS and survey data can be linked to the register but there is no overlap between IDS and survey units (top right on Diagram 2.6);
3. All IDS and survey units can be linked to the register but IDS and survey data partially overlap (bottom left on Diagram 2.6);
4. All survey units can be linked to the register but there are some IDS data that cannot be linked, for instance information about a certain number of IDS units is not sufficient for linkage (bottom left on Diagram 2.6).

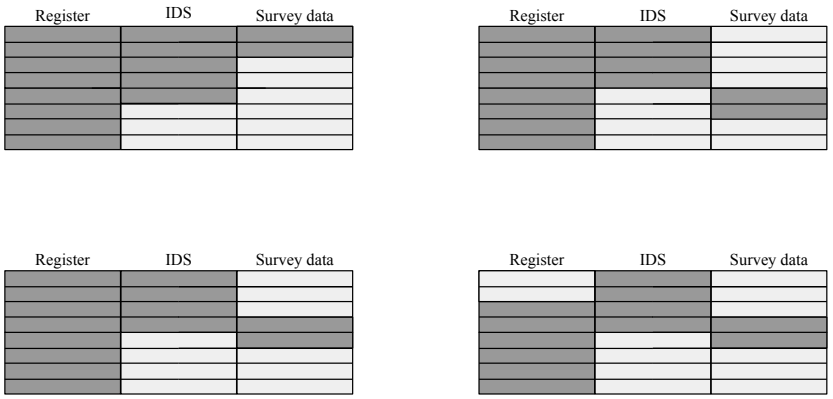


DIAGRAM 2.6: Hypothetical linkage scenarios for registers, Internet data sources and survey data

These data sources could be integrated using deterministic or probabilistic record linkage and statistical matching/data fusion. However, it is very unlikely that IDS and survey data will have the same identifiers as those used in the register, which renders deterministic record linkage inapplicable. To apply probabilistic record linkage or data fusion common variables are required to link records with a certain level of probability/distance. In this case the following variables can be considered: (1) address, (2) floor number, (3) floor area and (4) number of rooms. The correct address is the most important variable that makes identification possible. However, Polish IDSs often contain only the street name with the building number (without the apartment number), so additional characteristics of the flat should be used.

Finally, integrated micro-data presented in Diagram 2.6 should be processed to detect errors and inconsistencies between sources. Hypothetical linkage scenarios for sources do not take into account the fact that variables in all data sources may contain missing data. Hence, imputation and calibration can be applied to correct for this type of error. For instance, information form the Register of Transactions can be used to fill the gaps in variables observed in IDSs, while survey data may provide information that is not available in these sources. After imputation and calibration an *integrated statistical micro-data* is created, which can be used to calculate multivariate statistics based on integrated sources.

The main reason for using IDSs for real estate market statistics is to create a system where IDSs are the main source, which is supplemented by register and survey data. There are several arguments for this solution. First of all, in the Polish statistical system there is no sampling frame or other list of all properties offered for sale in the market. This is due to the fact that the real estate market, and particularly the secondary market, is a hard-to-reach population. Table 2.2 contains a comparative characterisation of a hard-to-reach population and the secondary real estate market. Secondly, brokers or owners look for potential buyers (in particular young people) and the Internet is the most popular place for posting ads. Another reason is that it may be possible to account for unobserved on-line units by linking IDS data with register or survey data. In particular, the Register of Transactions should prove a useful source of information as it contains all transactions made both in the primary and secondary market.

TABLE 2.2: A comparative characterisation of a hard-to-reach population and the secondary real estate market

Hard-to-reach population	Secondary real estate market
1. The population of interest is relatively small	1. A small fraction of properties is offered for sale
2. Members of the population of interest are hard to identify	2. We do not know in advance whether a given property is for sale
3. Lack of sampling frames for these populations	3. There is no sampling frame for properties offered in the secondary market
4. Persons concerned do not wish to disclose that they are members of this population	4. Not all properties are presented to a wider audience, nor are brokers willing to present all the information
5. The behaviour of the population of interest is not known	5. Motivations to put properties for sale vary and are unknown

Note: based on Marpsat and Razafindratsima (2010).

The proposed system, in which IDS is the main data source, is presented in Diagram 2.7. The system starts with a statistical concept (for instance the secondary real estate market), which is used to select an IDS. In the first stage, *data source selection* is made. For instance, in Poland there are various online services devoted

to real estate (advertisement services, websites of brokers or brokers associations, Public Information Bulletin), which differ in terms of the number of offers presented, popularity and population. The selection of an IDS was discussed earlier in the chapter.

Once a number of web services have been selected (denoted $1, ..., K$) data are collected using one of the following approaches:

- data acquisition directly from the owner; the data can be delivered on a weekly or monthly basis in an Excel, text or DB file,
- data acquisition directly from the owner via an API; Statisticians download the data directly from the owners' database using a secure connection (user, password and permissions are assigned),
- data acquisition indirectly from the owner using web-scraping techniques; This approach requires the development of special algorithms that semi-automatically collect data from the web service.

All of these approaches have advantages and disadvantages and require co-operation between NSIs/statisticians and IDS owners. The first one requires the data owner to prepare and send the data to NSIs/statisticians. This process can increase the respondent burden and reluctance to share data. The first option also requires the preparation of a structure of the final data set and the information needed for statistics. In the second solution, the data owner prepares login, password and permissions to acquire data directly from the database. This approach is the most suitable because the owner already has the necessary infrastructure and interfaces, for instance the Application Programming Interface (API), which is a set of routines, protocols, and tools for building software applications. This approach limits the data owner's involvement in providing access. Data collection is made by the statistician and, consequently, the respondent burden is low. Moreover, the structure of these databases is likely to be more stable over time than that of web pages. Hence, this approach would enable a long-term data collection process without the need to change data collection algorithms. The last option requires the statistician to create special web-scraping algorithms, which should be adjusted separately for each online services. The preparation of a web-scraper for one online service will be less demanding than doing the same for several web services with different structures. Moreover, the design of web pages may change over time, which means the algorithm needs to be adapted to new environments. Another drawback is that this method of data acquisition generates additional traffic on the website, which may be undesirable for the data owner. Hence, the best strategy for data collection is to co-operate with the data owner and gain direct access to available databases.

After the data collection process is completed, data are processed to ensure overall coherence of different sources (denoted on the graph as *Data cleaning step I*). For instance, variable names, street names, brokers' names, number of rooms or floor number are harmonized. This process also involves extracting additional information from long descriptions that most advertisements contain. In the second step (*Data cleaning step II*) micro data from the selected IDS are linked and

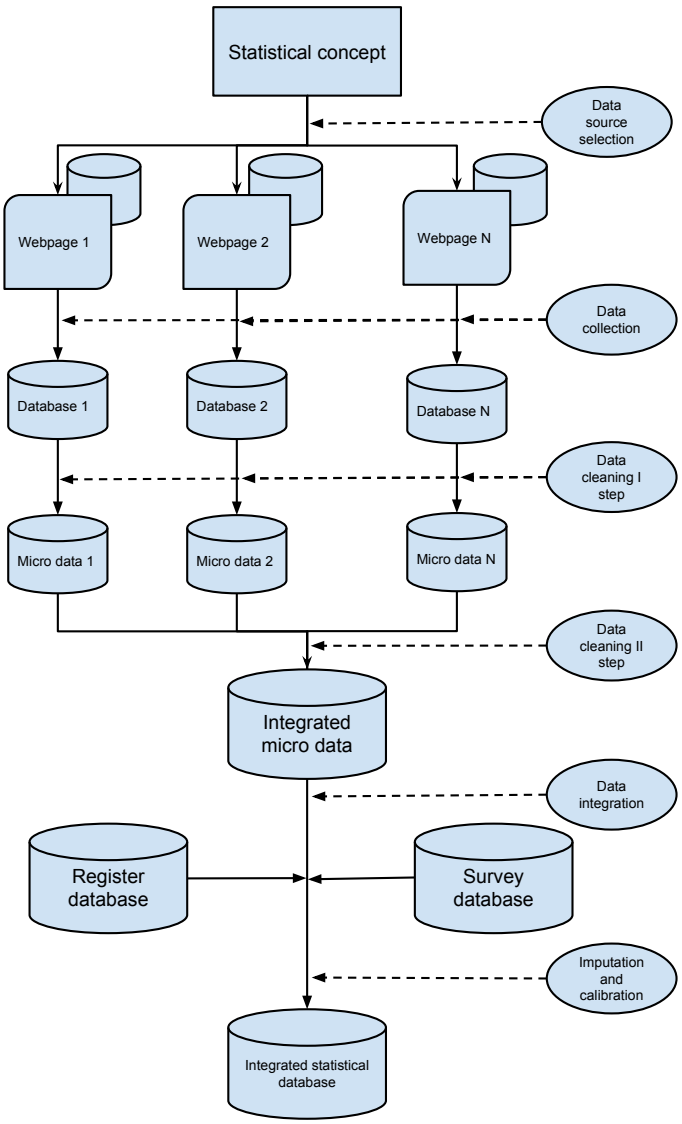


DIAGRAM 2.7: Internet data sources as the main source for real estate market statistics

statistical units are identified in the advertisements. Furthermore, the output *integrated micro data* is de-duplicated and consistency between IDS variables is verified. This step completes data processing and the cleaning of selected IDSs.

Next, the integrated micro IDS data are linked with register (denoted *register database*) and survey (denoted *survey database*) data. In the Polish context, the natural choice can be the Register of Transactions, and the survey conducted by NBP/CSO. The Register of Transactions should first be processed because it contains objects (transactions) that might involve multiple properties. Hence, before the integration process takes place, statistical units should be identified. In the case of the NBP/CSO survey, the information collected already contains statistical units provided by brokers. In the last stage the integration makes it possible to (1) account for coverage and selection errors and (2) verify measurement errors in IDSs. This can be achieved by imputation and calibration conducted after

linkage; finally, the *integrated statistical data base* is created.

Ideally, the output dataset should yield new statistics about the real estate market, such as the number of properties offered in the market, differences between the offer and transaction price or time-to-sell. Moreover, with micro-level data one could produce estimates at the level of domains that have not been published before. Nonetheless, it should be noted that not all information is published online and linkage with the Register of Transactions and survey data may help to reduce undercoverage. For instance, survey weights assigned to brokers participating in the NBP/CSO survey can be shared among properties using the Generalized Weight Share Method proposed by Deville and Lavallée (2006), Lavallée (2009), and Lavallée and Caron (2001).

DIAGRAM 2.8: The most possible outcome of the integrated micro-data when the Internet data source is the main data source

Finally, Diagram 2.8 shows the most likely linkage scenario for IDS, register and survey data. Dark grey represents the linked while light grey the non-linked statistical units. The failure to link units in each case can be explained as follows: (1) an IDS may contain properties that are offered for sale but are not necessarily sold; (2) the Register of Transactions only contains information about fully-owned properties, which excludes properties owned by housing co-operatives; (3) Survey data may partially refer to objects in the register and in an IDS, while brokers may also report information about properties that are not offered online.

2.5 Quality of Internet data sources

Quality assessment is the most important part of the evaluation of statistical and non-statistical data sources. New data sources, in particular IDSs, are not recognized by official statistics. Hence, possible errors and their sources should be carefully studied to indicate potential sources of bias. To bridge this gap, we have decided to adapt the classification of errors connected with the *two-phase life cycle of integrated statistical micro data* proposed by Zhang (2012b). In this classification, errors are identified at two stages: the first is connected with one data source (before integration), the second phase involves many integrated sources. Table 2.3 shows errors broken down by representation and measurement. The first group is connected with objects and units, the second group refers to variables. Representation covers the following errors: frame error, selection error, missing data and redundancy, coverage error, identification error and unit error; the second group

consists of validity error, measurement error, processing error, relevance error, mapping error and comparability error. Table 2.3 includes potential sources of error in the case of IDSs as well as examples.

2.5.1 Representation errors

One data source – representation Following Zhang (2012b), the first error is *frame error*, which is associated with differences between the target set and an accessible set of the target population. The accessible set may be associated with a sampling frame. However, it should be taken into account that such a list is not always available. For instance, no sampling frame of households is available in Poland; instead, a list of units for a different but related population is used (a list of addresses, other cases were studied in Lavallée (2009)). In the secondary real estate market in Poland there is no list that covers all residential properties on sale. Instead, there are several advertisement services or brokers web pages that partially cover the target population. These web services may refer to specific types of residential properties, such as flats, houses or lofts. As a result, only a certain subpopulation is covered. Moreover, in the Internet one company often owns several web services that target the same population (the same target audience). On the other hand, brokers may place offers on several web services. As a result, multiple, overlapping frames can be observed online. However, certain units may may still be excluded. *Frame error* may be substantial if web services represent only a small fraction/subpopulation of the target set, or unobserved units significantly differ from those observed online (self-selection). This may cause serious bias, which is why *frame error* should be carefully studied to detect such differences.

In the case of IDSs, *selection error* is associated with the response (self-selection) mechanism. Respondents r_{IDSs} independently decides whether or not to use a given IDS. However, if self-selection is connected with the target variable (for instance the price), then informative self-selection or non-ignorable missing data is observed. Hence, if specific units are excluded from an ID, then IDS-based characteristics may be severely biased. In addition, selection error can me associated with the following situations: (1) IDS owners do not provide access to the whole data; (2) the owner may only provide limited access to data for web-scrapers or through Application Protocol Interface (API); (3) finally, the data collection process may be limited to a specific period. If certain types of objects/units are excluded owing to limitations made by data owners, additional bias is introduced.

TABLE 2.3: Quality of Internet data sources

Nr	Component	Source of error	Example
		REPRESENTATION	
1	Frame error	Differences between the target set of all units and the available sampling frame.	An IDS should be treated as an imperfect frame/list. Brokers and owners use one or several advertisement services to offer properties. However, not all properties are presented online. In addition, these advertisement services overlap owing to business considerations (several services are owned by the same owner) or customer behaviour (use of different services to reach a wider audience). These services do not fully cover the target population, which gives rise to error.
2	Selection error	The result of using a specific sampling method. The Internet is connected with the self-selection mechanism and the data collection process.	Advertisement services are created by their users, so no sampling process is involved. Decisions of brokers or property owners are the basis of IDSs. In addition, IDS owners may provide limited access to API (e.g. Twitter or Facebook limit access to public tweets or information), web-scraping may also introduce selection error depending on the stability of a given scraper, privacy or protection of data source (e.g. services may block scrapers).

Continued on the next page

Table 2.3 – *continued from previous page*

Nr	Component	Source of error	Examples
3	Missing data and redundancy	Units may not provide information for a given variable; this error can be introduced during data collection or through technical problems; unverified information referring to a different population, erroneous or false information; information may differ between web services (quality checks).	IDS often contain objects that refer to the same or non-existing units; Information in a given field can be missing while being available following meta-data or natural language processing; During data cleaning it is possible to verify what information is redundant.
4	Coverage error	Not all units from the target population are observed in the Internet; not all units from the Internet population are observed in a given IDS.	IDSs that differ in the number of objects/units; an IDS may contain false or outdated objects; certain units may be presented only on special websites (e.g. producers' or brokers web-site). In addition, coverage error may be the result of data cleaning and linkage. The popularity of a given IDS may depend on its business concept and what kind of subpopulation it intends to reach.
5	Identification error	The relation between individual objects (actions, non-statistical units) and composite units.	Advertisements may refer to residential properties offered with a garage, additional space or may refer to several properties.
6	Unit error	Error occurs when objects (or composite units) are transformed into statistical units.	Multiple occurrences in IDSs may cause the following types of relations between objects: one-to-one (one ad to one statistical unit), one-to-many (many properties listed in one ad) or many-to-one (different ads refer to one property).
		MEASUREMENT	

Continued on the next page

Table 2.3 – *continued from previous page*

Nr	Component	Source of error	Examples
1	Validity error	Differences between concepts used in an IDS and the intended concept.	A overlap or discrepancy between concepts used in official statistics, for instance floor area, number of rooms or building type are consistent with definitions used in official statistics, while the measurement of sentiment (via natural language processing) may not be consistent with the concepts of customer confidence measured by surveys.
2	Measurement error	Differences between the intended measurement and obtained response.	Information in ads is often inaccurate (e.g. rounded), either intentionally (false information provided) or for lack of knowledge (e.g. exact year of construction); text mining methods may introduce additional measurement errors related to algorithms they are based on.
3	Processing error	Differences between the response and edited response.	Property ads often contain long text descriptions that can be used to verify/edit values provided for variables of interest. During this process (data cleaning and editing) true values may be replaced with false or erroneous information.
4	Relevance error	The extent of disagreement between the target concept and the harmonized measure (proxy).	For instance the offer price may not be available in the statistical system, while the transaction price is; what is measured is, in fact, often unknown, due to the lack of auxiliary information for comparison.
5	Mapping error	Connected with the re-classification of variables ensure consistency with definitions used in official statistics.	Advertisements may refer to the same building with different descriptions (brokers may not be sure about construction details of a given building); floor numbering may differ (for instance the ground floor may be labaled as either 0 or 1).

Continued on the next page

Table 2.3 – *continued from previous page*

Nr	Component	Source of error	Examples
6	Comparability error	Differences between the re-classified and adjusted measures (consistency of various data)	For instance, IDSs integrated with survey or register data during the micro-integration process may result in inconsistency across these data sources.

Note: own elaboration based on Zhang (2012b).

Missing data and redundancy is an important problem in new data sources. Even if the number of records is large, item and unit non-response is likely to be present. For instance, IDS owners prepare forms for data input and decide what kind of information is required. The decision whether to provide information for optional fields in the form is left to the respondent. The missing data mechanism can be Missing Completely at Random (MCAR), Missing at Random (MAR) or Not Missing at Random (MNAR) in the sense of Rubin (1976). For instance, MCAR in an IDS can be connected with interruptions in the data collection (web-scraping) process or software/hardware failures on the data owner's side. MAR is often connected with non-response or self-selection that is associated with other than target variables. In an IDS, given the non-probabilistic character of data, MNAR is more likely. In that case missing data are the result of informative self-selection. MNAR in an IDS can involve the offer price, which means that specific flats (for instance those highly priced) are not presented online. For instance, brokers may limit the number of offers placed online, house owners may not be interested in providing information about certain characteristics of their properties, information about the (natural or legal) person responsible for the property may be missing in order to avoid paying additional fees to brokers. Moreover, the missing pattern may also vary depending geographical area. Another aspect of IDSs is redundant information. For instance, photos, text descriptions or other para-data may be outside the scope of a survey. However, assessment of such data should be done prior to their removal. For instance, text descriptions or recorded behaviour can be used for imputation of missing data. However, the removal of these data can substantially reduce the size of the final dataset.

Integrated data sources – representation *Coverage error* can be assessed after multiple data sources have been integrated (Zhang, 2012b). During the integration of IDSs it is possible to detect false objects (e.g. advertisements of flats that do not exist) or objects that do not refer to the target population. For instance, when the secondary residential market is the target population, it may happen that a given sampling frame contains IDS units from the primary market or the commercial property market. Moreover, the integration process can help to determine the degree of overlap between sources. For instance, it may not be possible to correctly link objects from several IDSs. Furthermore, assessment of the coverage error can be the first step in determining the representativeness of IDS data.

IDS objects (actions or units) are rarely statistical units. Therefore, identification of objects that can be transformed into statistical units is not free from *identification error*. For instance, an online advertisement (object) may refer to one or multiple residential properties (flats, houses) including a garage and additional storage. Another example are flats located in a specific building. This is what Zhang (2012b) refers to as *composite units*, which are made up of *base units* (objects). Taking the above into account, it may happen that an advertisement (object) contains references to different statistical units, which should be treated as *composite units*.

Another type of error which can have a substantial effect on estimates is *unit error*. Unit error is present when statistical units are derived from objects. For instance, in the real estate market multiple advertisements (objects) often refer to the

same statistical unit (e.g. flat, house). This case is mainly due to the behaviour of natural and legal persons that use many different web services to reach to a wider audience. In fact, these persons rarely use only one service, which result in multiple occurrences of the same objects in different IDSs. However, the problems listed above can also be connected with the specificity of the Polish real estate market. For instance, in the Netherlands the main association of brokers (Nederlandse Vereniging van Makelaars, NVM, www.nvm.nl) provides Statistics Netherlands with a database of all properties put up for sale, which contains statistical units (e.g. flats, houses), not objects. Nonetheless, the derivation of statistical units can be the most challenging aspect of using IDSs. *Unit error* is the last type of error associated with *representation* and now we will discuss *measurement* errors found in integrated data sources.

2.5.2 Measurement errors

One data source – measurement *Validity error* is the first type of measurement error. It refers to situations when the concept measured in an IDS is verified against concepts used in statistics. This error is committed at the conceptual level and does not directly refer to the response. In the case of the real estate market, concepts observed in IDSs and official statistics are the same. For instance the number of rooms, floor area or the definition of a kitchen. However, in Poland official statistics mainly uses the definition of value/purchase price in the context of *value of goods and materials*. However, for the purpose of the National Bank of Poland and CSO survey (which will be discussed later on) the definition of the offer price is also used. If we are interested in the transaction price, the offer price is a proxy variable and may be affected by validity error. Another example is related to social media or Internet search indices, where the target variable can be masured on the basis of statements expressed in natural languages, which are also prone to validity error. In contrast, customer confidence indices are based on standarized surveys, which substantially differ from those formulated in natural languages.

Another issue that should be assessed is *measurement error*. In surveys, measurement error mainly refers questions: whether they are correctly understood and whether the final measurement takes place as planned. In registers it can be connected with the difference between definitions of what is measured in official statistics (e.g. unemployment defined by ILO) and what is obtained from administrative sources (e.g. registered unemployment). Zhang (2012b) summarised it as the difference between the intended and obtained response. In IDSs measurement errors can refer to values that are false on purpose (e.g. incorrect floor number to avoid the identification of a given apartment, or to prevent a direct purchase from the owner), by mistake (e.g. misspelled street, floor area) or for lack of knowledge (e.g. year of construction, building material). Another error is connected with providing inaccurate values due to rounding. For instance, floor area may be rounded up to the nearest integer.

Processing error is connected with the process of editing the obtained measurement. This step includes errors connected with editing, imputation and the processing of natural language responses, photos or other formats that are not

commonly used in official statistics but frequently appear in IDSs. For instance, natural and legal persons often provide text descriptions about residential properties, which may contain the same, different or additional information. After processing long text descriptions can provide values for missing fields. The processing of IDS data is the most challenging aspect of data preparation for statistics. Methods used at this stage are not widely applied in official statistics and require certain skills. The challenges connected with IDS will be presented later on.

Integrated data sources – measurement *Relevance error* error is connected with the process of harmonizing measures. Concepts used in different data sources may differ from the target concept. The process of re-classification of input variables to harmonized variables involves *mapping error*. For example, different spellings of street or district names may be used, the same variables may be coded differently (number of rooms, floor number), floor numbering may start from 0 or 1 (depending on how the ground floor is counted) or the description of the kitchen can be coded either as a separate room or kitchen space connected with another room.

Finally, *comparability error* is the result of adjusting measures between data sources to be coherent. The micro-level reconciliation of inconsistency in multiple-sourced data is often referred to as micro integration (Bakker, 2010; Zhang, 2012b). For instance, in Polish IDS about the real estate market it may happen that multiple objects (ads) refer to the same statistical unit (dwelling) but have inconsistent information about the number of rooms or floor area (unrounded, rounded). On the other hand, the register of real estate market transactions can provide unbiased information on these variables. Linking with register data may help to correct inconsistencies in IDSs.

Careful study of these errors can provide ideas for possible solutions and methods that are used in official and non-official statistics. Moreover, these errors should be described in detail to provide an overall assessment of possible bias and uncertainty in using new data sources for statistics. In the coming years these data sources will play an important role in officials statistics, mainly in the face of increasing non-response and falling budgets.

2.5.3 Data source specific errors

In addition to errors discussed below there are data source specific errors. This type of error are due to the specific character of data or method of data collection. Data can be collected by the private or government agency, longitudinal surveys or data collection process (e.g. mixed-modes). For instance, in repeated/panel surveys bias connected with number of wave is observed (Bailar, 1975; Krueger et al., 2014). Brakel and Krieg (2009) take into account this bias in small area estimation models. Other problems with longitudinal studies are attrition/drop-outs of units or changing the structure of households. Lohr and Brick (2012) summarise these errors as panel conditional errors. Lohr and Brick (2012) identify the following sources of bias for the victimization surveys – sponsorship, associated with the entity that collects the data; recall period which rely on the memory of respondent; or interviewer and mode effect, which depends on the person who

collects data. For more examples Biemer and Lyberg (see 2003), Couper (2011), and Zhang (2005). In case of IDSs we can name the following sources of errors (1) popularity of given data source, (2) data source owner and market organisation and (4) profitability and costs of using given online services.

The first group refers to multiple data sources on the same topic. In such situation, there are a couple of the most popular web-portals that may (in total) cover major part of target population. For example, in Poland we have three main portals OtoDom.pl, Gratka.pl and OLX.pl that according to the non-official research are very likely to be choose by the brokers to place real estates to sale[24]. Popularity of a certain web-portal may differ in time, what in result in problems with continuity in the longitudinal studies. A change in popularity of given data source may stability of given data source in time.

The second subgroup refers to the owners of given data source and market organisation. It is common case that one company may own several web-portals devoted to the same or similar part of the market. For example, company Melog[25] own 6 web-portal[26] that are very likely to contain the same units. Usage of multiple data sources is also connected with how market is organised. Brokers use software that allow to automatically send ads to several advertisement services. Moreover, this error is also connected with the perception of company image. For example, brokers or individuals may have bad experience with or attitude towards a given company which may result in attrition to competitors.

The last subgroup are profits and costs of using certain online service. To follow the example of real estate market, brokers may be interested in the web-portals where are more likely to find potential customers. On the other hand, fees that are paid for the owner of the web-portal may be not suitable. For instance, prices changes with number of offers placed on the IDSs and this may cause that not all offers will be placed on advertisement service. In addition, the Internet companies compete not only by presenting number of users but also in costs that should be covered by they users. On the other hand, individuals are not interested in covering fees and as a result migrate to web-portals that are free to use.

To sum up, all of the errors connected with the nature of data source influence the self-selection mechanism and are not observed directly. Therefore, it is difficult to quantify the effect of these errors on probability that given unit will be present on IDSs.

2.6 Conclusions

The above section has dealt with several topics related to the quality of Internet data sources about the real estate market. The first major area were problems with defining the population with respect to the real estate market was discussed. Four different populations (IDS, Advertisements, Brokers and Transactions) were presented. Potential statistical and non-statistical data sources were discussed as

[24] See http://media.allegro.pl/pr/296617/najskuteczniejsze-serwisy-ogloszen-nieruchomosci-wyniki-badania-pbs.

[25] See http://www.melog.com.

[26] See www.oferty.net,www.domy.pl, www.komercyjne.pl, www.bezposrednie.com, www.noweinwestycje.pl, www.nieruchomosci.pl.

well as the problem of selecting IDSs for statistics. The last topic analysed in this section was the integration of IDSs into the statistical system and potential errors that can be expected in this process. The next chapter will focus on theoretical concepts that can be used to assess IDSs for real estate market statistics.

Chapter 3

The theoretical basis of Internet data sources

3.1 Basic notation and definitions

This chapter is devoted to theoretical aspects of Internet data sources. Basic notations and definitions regarding IDSs are introduced in the first section. Later, the discussion will focus on the notion and definitions of representativeness in the context of IDSs. The third section describes a two-step procedure of measuring representativeness. The chapter ends with an outline of possible measures to assess the representativeness of IDSs.

3.1.1 The notation

In the second chapter the concepts of population and sample were defined with respect to IDSs. Now, let us introduce the basic definitions and the notation concerning sample and population distributions. Let $\mathbf{\Omega}_{TP,t,d}$, $\mathbf{\Omega}_{IP,t,d}$, $\mathbf{\Omega}_{IDSsP,t,d}$, $\mathbf{\Omega}_{IDSs,t,d}$, $\mathbf{\Omega}_{AP,t,d}$, $\mathbf{\Omega}_{RP,t,d}$ be populations defined in Chapter 2 and observed in period $t \in \{1, ..., T\}$, where T denotes the number of periods and domain $d \in \{1, ..., D\}$, where D denotes the number of domains. In addition, let $\mathbf{\Omega}_{IDSs,k,t,d}$ be an IDS population observed in data source $k \in \{1, ..., K\}$, where K denotes the number of data sources. Let $s_{IDSs,t,d}$ be the observed sample size in period t for domain d and let r_{IDSs} be statistical units for which the target variable is not missing in period t for domain d.

Let y be the target variable of interest (continuous, binary, ordinal etc.), which has a definition in the statistical system. In the case of the real estate market this definition can refer to the offer or transaction price of a residential property. Now, let us assume that v is a proxy variable that is observed in the IDS population ($\mathbf{\Omega}_{IDSs}$) or the register population ($\mathbf{\Omega}_{RP}$). In the case of the real estate market variable v can be defined as the transaction price, which is observed in $\mathbf{\Omega}_{RP}$. Definitions used in IDSs about the real estate market may be consistent with those specified in the statistical system, in which case $v = y$. It should be noted, however, that is by no means common, for instance the statistical system may only use a definition of the transaction price, while IDSs only make use of the offer price. In Poland both offer and transaction prices are surveyed by NBP/CSO. Therefore, for the sake of simplicity it is assumed that there is no difference between definitions of v and y in the real estate market in Poland and y will be used.

Let us assume that $\boldsymbol{x}$ refers to auxiliary variables associated with the statistical unit, such as floor area, the number of rooms or location. In addition, let us assume that $\boldsymbol{z}$ denotes variables that are indirectly associated with the statistical unit and are responsible for the selection mechanism. Such variables can be taken from Internet paradata (e.g. cookies), specific IDSs or refer to certain characteristics of a related population (e.g. broker's characteristics).

Let θ be a target characteristic of variable y, in addition let $\theta_{t,d}$ denote θ in given period t for domain d. Since different data sources are used to provide estimates of θ, the following notation is proposed. In addition, the proposal is consistent with Fosen and Zhang (2011) and Zhang (2012a), which serves as the basis for the approach to estimating bias, which will be presented later.

- Sample survey based estimates: $\hat{\theta}, \hat{\theta}_t, \hat{\theta}_d, \hat{\theta}_{t,d}$ – is the estimator of θ for variable y observed in sample survey s,
- Register-based estimates: $\tilde{\theta}, \tilde{\theta}_t, \tilde{\theta}_d, \tilde{\theta}_{t,d}$ – is the estimator of θ for variable v observed in $\boldsymbol{\Omega}_{RP}$,
- IDSs-based estimates: $\breve{\theta}, \breve{\theta}_t, \breve{\theta}_d, \breve{\theta}_{t,d}$ – is the estimator of θ for variable y observed in $\boldsymbol{\Omega}_{IDSs}$. In addition, if different data sources are used, then subscript k denoting k IDSs should be added, for instance $\breve{\theta}_{k,t,d}$.

3.1.2 Response propensity

In survey methodology, the terms non-informative or informative sampling are often used (Pfeffermann, 2011). Non-informative sampling refers to a situation when the probability of inclusion in the sample, denoted by I_i, is associated only with $\boldsymbol{x}$ and/or $\boldsymbol{z}$, that is $Pr(I_i = 1 \mid \boldsymbol{x}) = \pi_i$ or $Pr(I_i = 1 \mid \boldsymbol{x}, \boldsymbol{z}) = \pi_i$, where π_i denotes the probability of inclusion. For instance, $\boldsymbol{z}$ can refer to variables associated with primary sampling units (PSUs) or, in general, the sampling scheme (e.g. the number of houses in a given area, the size of a municipality), while $\boldsymbol{x}$ are variables associated with the statistical unit (e.g. gender, age).

In contrast, informative sampling occurs when the probability of inclusion is also associated with target variable y. It means that $Pr(I_i = 1 \mid \boldsymbol{x}) \neq Pr(I_i = 1 \mid \boldsymbol{x}, y)$ or $Pr(I_i = 1 \mid \boldsymbol{x}, \boldsymbol{z}) \neq Pr(I_i = 1 \mid \boldsymbol{x}, \boldsymbol{z}, y)$. This may be due to including means or variances of the target variable into the sampling scheme or sample allocation.

Due to non-probability / self-selectivity of IDSs, the notion of informativeness or non-informativeness cannot be applied directly. IDSs or $\boldsymbol{\Omega}_{IDSsP}$ are generally not created as a result of probability sampling or on the basis of a probability sampling scheme. Figure 3.1 presents the selection mechanism that can be observed in IDSs. In the first phase, the population of property owners and real estate brokers is limited to those using the Internet. However, because the actual target population of interest are properties, no indicator variable is available to distinguish between owners and brokers that use the Internet and those who don't. Next, in the second phase, owners and brokers decide whether to use the Internet to sell properties. $\boldsymbol{\Omega}_{IDSsP}$ is a subset of $\boldsymbol{\Omega}_{IP}$, hence each statistical unit i has an indicator variable denoted I_i which is $I_i = 1$ if $i \in \boldsymbol{\Omega}_{IDSsP}$ and $I_i = 0$ otherwise.

Finally, in the third phase, not all statistical units have information about the target variable. As a result, the observed sample shrinks to r_{IDSs}, for which y is available. The r_{IDSs} was previously defined and described in Figure 2.1. Let R_i be the response indicator that is equal to 1 when $i \in r_{IDSs}$ and 0 otherwise. The size of sample r_{IDSs} is equal to $m_{IDSs} = \sum_{i=1}^{N_{TP}} R_i$. Because it is not possible to estimate π_i, instead we need to estimate *response probability* or the *propensity score*.

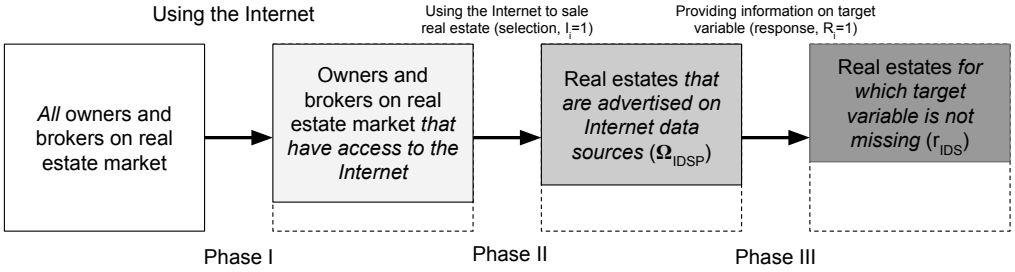


FIGURE 3.1: The self-selection mechanism underlying Internet data sources about the secondary real estate market

The propensity score was introduced in the context of missing data patterns by Rubin (1976). Rubin (1976) discusses three missing data patterns: (1) missing completely at random (MCAR), (2) missing at random (MAR) and (3) not missing at random (NMAR). The MCAR pattern occurs when missingness is due to random events such as system failures, interruptions in the data collection process or, in general, is not associated with $\boldsymbol{x}, \boldsymbol{z}$ or y. The MAR pattern is connected with $\boldsymbol{x}$ and/or $\boldsymbol{z}$, but not y itself. MAR is also the result of the sample selection process. In terms of response probability (response propensity, propensity score) MAR can be defined as:

Definition 3.1. Response probability under *Missing at Random* is the expectation of the response indicator variable conditional on auxiliary variables, but not on the target variable itself.

$$E(R_i) = \rho_i = Pr(R_i = 1|\boldsymbol{x}, \boldsymbol{z}, y, I_i = 1) = Pr(R_i = 1|\boldsymbol{x}, \boldsymbol{z}, I_i = 1). \quad (3.1)$$

The last group of missing data patterns is *Not Missing at Random*. In the MNAR pattern, missingness is not only related to auxiliary variables but also to the target variable. Ignoring the MNAR pattern may result in large biases and erroneous inference (Pfeffermann, 2011; Rubin, 1976; Szreder, 2010). The definition of response propensity under MNAR is given below.

Definition 3.2. Response probability under *Not Missing at Random* is the expectation of the response indicator variable conditional on auxiliary variables and the target variable itself.

$$E(R_i) = \rho_i = Pr(R_i = 1|\boldsymbol{x}, \boldsymbol{z}, y, I_i = 1) \neq Pr(R_i = 1|\boldsymbol{x}, \boldsymbol{z}, I_i = 1). \quad (3.2)$$

In practice, identification of all units that are members of $\mathbf{\Omega}_{IP}$ or $\mathbf{\Omega}_{IDSsP}$ is unlikely. It is more realistic to assume that information on all units of $\mathbf{\Omega}_{TP}$ and r_{IDSs} is available. Another case is when a probabilistic sample s_{TP} of $\mathbf{\Omega}_{TP}$ or a population related to $\mathbf{\Omega}_{TP}$ (for simplicity $\mathbf{\Omega}_{RP}$) and r_{IDSs} is available. In the first case, we reduce the number of conditions in (3.1) and (3.2) by eliminating I_i, which results in the expectation of R_i, which for $\forall_{i \in r_{IDSs}} R_i = 1$ and $\forall_{i \notin r_{IDSs}} R_i = 0$. Expectations of R_i for the case when only $\mathbf{\Omega}_{TP}$ and r_{IDSs} is available are given by equation (3.3) and (3.4):

$$E(R_i) = \rho_i = Pr(R_i = 1|\boldsymbol{x}, \boldsymbol{z}, y, I_i = 1) = Pr(R_i = 1|\boldsymbol{x}, \boldsymbol{z}), \tag{3.3}$$

and

$$\begin{aligned} E(R_i) = &\rho_i = Pr(R_i = 1|\boldsymbol{x}, \boldsymbol{z}, y, I_i = 1) = \\ &Pr(R_i = 1|\boldsymbol{x}, \boldsymbol{z}, y) \neq Pr(R_i = 1|\boldsymbol{x}, \boldsymbol{z}). \end{aligned} \tag{3.4}$$

When only r_{IDSs} and s_{TP} or $\mathbf{\Omega}_{RP}$ are available, response propensities are expressed in a similar way as in (3.3) and (3.4). However, the definition of R_i is different and is given by equation (3.5):

$$R_i = \begin{cases} 1 & \text{if } i \in r_{IDSs}, \\ 0 & \text{if } i \in s_{TP} \setminus r_{IDSs} \text{ or } i \in \mathbf{\Omega}_{RP} \setminus r_{IDSs}. \end{cases} \tag{3.5}$$

In (3.5) one assumes, realistically, that $r_{IDSs} \cap s_{TP} \neq \emptyset$ and $r_{IDSs} \cap \mathbf{\Omega}_{RP} \neq \emptyset$. This assumption means that the response propensity is calculated in reference to all units observed online ($R_i = 1$), while $R_i = 0$ is calculated only for units observed outside of r_{IDSs}. The approach whereby response propensity is accounted for by an auxiliary sample (s_{TP}) is used for self-selection web panels (see Lee, 2006) or in propensity score matching in observational studies (see Rosenbaum and Rubin, 1983).

Because ρ_i is unknown it should be estimated. There are several models that can be used to estimate ρ_i (Bethlehem and Biffignandi, 2011, ch. 11.2.4):

$$log(\frac{\rho_i(\boldsymbol{x}_i)}{1 - \rho_i(\boldsymbol{x}_i)}) = \boldsymbol{\beta}'\boldsymbol{x}_i + \epsilon_i, \tag{3.6}$$

where equation (3.6) is a logit model, $\boldsymbol{\beta}$ is a vector of regression parameters and ϵ_i is random error. Another model is a probit model

$$\Phi^{-1}(\rho_i(\boldsymbol{x}_i)) = \boldsymbol{\beta}'\boldsymbol{x}_i + \epsilon_i, \tag{3.7}$$

a generalized linear model

$$g(\rho_i(\boldsymbol{x}_i)) = \boldsymbol{\beta}'\boldsymbol{x}_i + \epsilon_i, \tag{3.8}$$

where g is a link function that should be specified beforehand, or classification and regression trees, CART (Breiman et al., 1984). These models can also be extended by using $\boldsymbol{z}$, which can influence the propensity score. These variables can refer to

brokers' or owners' characteristics, as well as paradata taken from IDSs. Models (3.6), (3.7) and (3.8) can also be extended by using the mixed model approach (cf. Bethlehem and Biffignandi, 2011, ch. 11.2.4).

3.1.3 Methods to reduce bias

The use of response propensity to correct for the self-selection bias has become a popular solution in web surveys Bethlehem (2010), Cobben (2009), Schonlau et al. (2009), and Szreder (2011). The inverse of estimated response propensity is used as a weight to reduce bias. This approach will be presented in Section 3.1.4. However, in general three methods to reduce bias can be distinguished – imputation, reweighting procedures or the model-based approach. Imputation is a procedure whereby missing values for one or more study variables are "filled in" with substitutes. These substitutes can be constructed according to a rule, or they can be observed values but for elements other than non-respondents. In this sense imputed values are artificial and can therefore contain error (Särndal and Lundström, 2005). There are several techniques that are used to impute missing values (cf. Lohr, 2009, ch. 8.6):

- deductive imputation,
- cell mean imputation,
- hot-deck imputation,
- regression imputation,
- cold-deck imputation,
- multiple imputation.

In addition to those presented above *mass imputation* can be found in the literature, particularly with reference to registers (Houbiers, 2004). Mass imputation is a procedure of imputing values for all units of the population based on sample surveys. As a result, all units receive values for the target variables. The use of imputation with respect to IDSs can resemble mass imputation, in other words imputation is made for all units from the target population based on values from IDSs. However, if the response mechanism underlying IDSs is follows the MNAR pattern, imputation can be biased. For more information on imputation see Fay (1996), Rao (1996), Rubin (1987, 1996), Särndal and Lundström (2005), and Shao et al. (2003).

Another approach used for bias reduction is reweighting. It is based on the assumption that population totals are known (e.g. demographic characteristics) and a vector of numeric values to match the known totals is applied. Initially, weighting is associated with sample surveys, where weights d_i are the inverse of probability π_i that a given unit will be sampled ($d_i = 1/\pi_i$). However, when a vector of initial weights $\boldsymbol{d}$ is applied to the final sample, it may happen that the sum of these weights ($\sum_i \boldsymbol{d}_i$) will not be equal to known population totals denoted by $\boldsymbol{X}$. Therefore, a weighting procedure should be applied in order to meet the

criteria. The formal background for calibration, which is the generalization of reweighting approaches, was proposed by Deville and Särndal (1992) and is given below. Special cases of calibration are known as raking and post-stratification. Let $\boldsymbol{d} = (d_i, ..., d_n)^T$ be a vector of initial sampling weights, $\boldsymbol{w} = (w_i, ..., w_n)^T$ a vector of final weights, x_j the auxiliary variable for which totals are known $\boldsymbol{X}_j = \sum_{i=1}^{N} x_{ij}$, where N is the population size and n is a sample size. Therefore, in order to get $\boldsymbol{w}$ we need to resolve the following set of equations

$$D(\boldsymbol{w}, \boldsymbol{d}) = \sum_{i=1}^{n} d_i G(\frac{w_i}{d_i}) \rightarrow min, \tag{3.9a}$$

$$\sum_{i=1}^{n} w_i x_{ij} = \boldsymbol{X}_j, \text{ j = 1, ..., k.} \tag{3.9b}$$

Equation (3.9b) is known as a calibration equation, which should be met in order to ensure equality of survey-based and population totals. A set of equations (3.9a) - (3.9b) has an additional restriction on weights given by (3.10)

$$L \leq \frac{w_i}{d_i} \leq U, \text{ where } L < 1 \text{ and } U > 1, i = 1, ..., n. \tag{3.10}$$

Equation (3.9a) is a minimization of distance function G between known weights $\boldsymbol{d}$ and unknown weights $\boldsymbol{w}$. Distance functions that are discussed in the literature are listed below (Särndal, 2007; Szymkowiak, 2009). The distance given by (3.11e) is used by Eurostat in the EU-SILC.

$$G_1(x) = \frac{1}{2}(x-1)^2, \tag{3.11a}$$

$$G_2(x) = \frac{(x-1)^2}{x}, \tag{3.11b}$$

$$G_3(x) = x(\log x - 1) + 1, \tag{3.11c}$$

$$G_4(x) = 2x - 4\sqrt{x} + 2, \tag{3.11d}$$

$$G_5(x) = \frac{1}{2\alpha} \int_1^x sinh[\alpha(t - \frac{1}{t})]dt. \tag{3.11e}$$

Initially, calibration did not take into account non-response. Lundström and Särndal (1999) and Särndal and Lundström (2005) extended calibration to the case when item or unit non-response is present. The main idea behind this approach is to get a set of weights that account both for non-response and auxiliary variables. However, the suitability of the calibration approach under non-response depends on auxiliary variables. Calibration will reduce bias when auxiliary variables are related to the non-response mechanism. When, a set of variables $\mathbf{x}$ does not account for that mechanism or the MNAR pattern is present, calibration will not eliminate bias.

Calibration is also widely used in registers, for example Wallgren and Wallgren (2014, ch. 11) provide an overview of how reweighting is applied in administrative sources. In that case an artificial vector of initial weights $\boldsymbol{w} = \mathbf{1}$ is created

and calibrated to match known population totals. Another approach is repeated weighting where all register-based census tables are calibrated to have the same population totals (Renssen et al., 2001).

As in the case of imputation, there are drawbacks of using weighting procedures. Gelman (2007) argues that weighting *is a mess* and notes that the application of weighting procedures is a very complicated procedure. As an alternative he presents a model-based Bayesian approach to modelling the target variable in the first phase and suggests reweighting estimates using post-stratification. Gelman (2007) proposes that models should contain all variables that were used for sample selection (that were used to create weights) and other variables that may influence the modelled outcome. Brick (2013) maintains that powerful auxiliary variables are required in order to reduce bias in sample surveys that suffer from non-response. However, some of the commentators disagree with Brick (2013) and Gelman (2007) arguing that weighting procedures provide consistent estimates and are easier to apply than model-based approaches.

The last method is discussed in Andrew Gelman's papers (cf. Shirley, 2015; Wang et al., 2015). These papers cover both non-probability and probability samples. For instance, Wang et al. (2015) applied a Bayesian hierarchical model with post-stratification to a non-representative poll of Xbox users. Wang et al. (2015) considers all possible combinations of sex (2 categories), race (4 categories), age (4 categories), education (4 categories), state (51 categories), party ID (3 categories), ideology (3 categories) and 2008 vote (3 categories) in the model. His results indicate small differences in the share of Obama voters estimated by the proposed approach based on Xbox data on the day before the election and estimates obtained from the 2012 national exit poll. Reilly et al. (2001) studied the situation when population totals are unknown, but a proxy variable under a dynamic model was used. The model-based approach is also applied in the case of small sample size domains (small area estimation) and has gained wide acceptance in such cases. Nonetheless, the main argument against the model-based approach is the need for strong auxiliary variables that explain the target variable and account for the selection mechanism.

3.1.4 Estimation

Finally, the expected outcome of probability and non-probability samples are estimates. Let us assume that the mean is the parameter of interest ($\bar{y}$), then the sample-based Horvitz-Thompson estimator is given by:

$$E(\theta) = \frac{\sum_i^{n_s} y_i d_i}{\sum_i^{n_s} d_i}, \tag{3.12}$$

where y_i denotes the offer price, $d_i = 1/\pi_i$ is a sampling weight assigned to each unit i and n_s denotes the sample size. For this simple example, it is assumed that non-response is not present and no calibration is performed.

When the register is used and contains a proxy variable v, the estimator of θ is a simple random sample mean:

$$E(\theta) = \frac{\sum_i^{N_{RP}} v_i}{N_{RP}}, \tag{3.13}$$

where, in the Polish context, v_i denotes the transaction price (proxy variable) and N_{RP} denotes the size of the register population (the number of transactions). Here, we assume that there are no missing data in the register (for instance, no weighting is applied).

Now, we focus on IDS-based estimation, continuing the example we are interested in $\bar{y}$ based on advertisement services, we should consider the following estimator based on Bethlehem (1988):

$$E(\theta) \approx \frac{1}{\bar{\rho} N_{IP}} \sum_{i=1}^{N_{TP}} \rho_i I_i y_i, \tag{3.14}$$

where I_i denotes the indicator variable (presence in $\mathbf{\Omega}_{IP}$), ρ_i denotes the response probability/propensity score, y_i is defined as before. When weighting adjustment is applied, then (3.14) is given by:

$$E(\theta) \approx \frac{1}{\bar{\rho} \sum_{i=1}^{m_{IDSs}} w_i} \sum_{i=1}^{N_{TP}} \rho_i w_i I_i y_i, \tag{3.15}$$

where w_i are weights obtained from post-stratification or calibration. Bethlehem and Biffignandi (2011) demonstrated that the variance of (3.14) is given by equation (3.16):

$$V(\theta) \approx \frac{1}{\bar{\rho} N_{TP}} \sum_{i=1}^{N_{TP}} \rho_i (1 - \rho_i)(y_i - E(\bar{y}))^2. \tag{3.16}$$

Bethlehem (1988) and Bethlehem and Biffignandi (2011) note that mean response probability $\bar{\rho}$ is not equal to the population mean, hence estimator (3.14) is biased. This bias can be quantified by the following equation:

$$Bias(\theta) = \frac{S_{\rho y}}{\bar{\rho}} = \frac{cor_{\rho y} sd_\rho sd_y}{\bar{\rho}}, \tag{3.17}$$

where $S_{\rho y}$ is the covariance between the values of the target variable and the response probabilities (ρ), $cor_{\rho y}$ is the corresponding correlation coefficient, sd_ρ is the standard deviation of ρ and sd_y is the standard deviation of y. Bias depends on the correlation between ρ and y. If this correlation is high, bias is large. In addition, if sd_ρ or sd_y is high, then bias is also large. Bethlehem (2010) showed that given $\bar{\rho}$, maximum standard deviation sd_ρ cannot exceed:

$$sd_\rho \leq \sqrt{\bar{\rho}(1 - \bar{\rho})}. \tag{3.18}$$

In addition, Bethlehem (2010) demonstrates that the maximum bias of (3.17) is equal to:

$$|max(Bias(\theta))| = sd_y\sqrt{\frac{1}{\bar{\rho}} - 1}. \tag{3.19}$$

To account for bias, Bethlehem and Biffignandi (2011) introduces the bias-adjusted estimator given by the following equation:

$$E(\theta)^* = \frac{\bar{\rho}}{m_{IDSs}} \sum_{i=1}^{N} \frac{R_i y_i}{\rho_i}, \tag{3.20}$$

Equation (3.20) reduces to equation (3.21) because we sum over $R_i = 1$:

$$E(\theta)^* = \frac{\bar{\rho}}{m_{IDSs}} \sum_{i=1}^{m_{IDSs}} \frac{y_i}{\rho_i}. \tag{3.21}$$

If weighting adjustments have been made (for instance, post-stratification, calibration), equation (3.21) is extended by including weights denoted by w_i:

$$E(\theta)^* = \frac{\bar{\rho}}{m_{IDSs}} \sum_{i=1}^{m_{IDSs}} \frac{y_i w_i}{\rho_i}. \tag{3.22}$$

After estimating ρ_i equation (3.22) is given by:

$$E(\theta)^* = \frac{\hat{\bar{\rho}}}{m_{IDSs}} \sum_{i=1}^{m_{IDSs}} \frac{y_i w_i}{\hat{\rho}_i}. \tag{3.23}$$

This section has discussed the basic setting for IDSs and possible solutions that can lead to unbiased estimates. However, in order to estimate response propensities or apply weighting procedures, it is necessary to examine the underlying selection mechanism of IDSs. This mechanism can be detected by measuring representativeness. Various concepts covered by this term will be studied in the next section.

3.2 The notion and definitions of representativeness

Representativeness is a concept widely discussed in survey methodology, particularly in official statistics. There is, however, no straightforward definition of representativeness, which was already noted by (Kruskal and Mosteller, 1979a,b,c), who provides a list of concepts used in statistical and non-statistical literature of that time:

- general, unjustified acclaim for the data,
- absence of selective forces,
- mirror or miniature of the population,
- typical or ideal case(s),
- coverage of the population,

- a vague term to be made precise,
- representative sampling as a specific sampling method,
- representative sampling as permitting good estimation,
- representative sampling as good enough for a particular purpose.

In addition to the above denotations, Bethlehem (2009) provides two definitions of a representative survey sample.

Definition 3.3. A survey data set is defined to be *representative with respect to variable(s)* $\boldsymbol{x}$ if the distribution of $\boldsymbol{x}$ in the data set is equal to the distribution of this variable in the population (Ω):

$$f_s(\boldsymbol{x}, I_i = 1) = f_\Omega(\boldsymbol{x}), \tag{3.24}$$

where f denotes probability density function (PDF) and $I_i = 1$ is an indicator variable as previously defined. A vector of auxiliary variables $\boldsymbol{x}$ refers to characteristics of the population, particularly demographic variables, such as sex, age, education or marital status. In the case of the real estate market $\boldsymbol{x}$ may refer to property type, floor area or the number of rooms.

The second definition formulated by Bethlehem (2009) refers to a weighted sample adjusted to the known population marginal distribution of $\boldsymbol{x}$.

Definition 3.4. A survey data set is defined to be *representative with respect to auxiliary variables* $\boldsymbol{x}$ if the distribution of $\boldsymbol{x}$ is adjusted to the known distribution marginal distribution of $\boldsymbol{x}$ in population (Ω):

$$f_s(\boldsymbol{x}|w_i, I_i = 1) = f_\Omega(\boldsymbol{x}), \tag{3.25}$$

where w_i denotes adjusted weights for each unit i. w_i can represent the inverse of the probability of inclusion in the sample $w_i = d_i = 1/\pi_i$ or corrected weights (e.g. through calibration, post-stratification $w_i = \lambda_i d_i$). The definitions provided by Bethlehem (2009) can be understood as cases representing a *miniature of the population* proposed by Kruskal and Mosteller (1979a,b,c). In other words, a sample should have the same characteristics as the target population. Nonetheless, Bethlehem's definitions refer to the case when a weighting scheme must be applied in order to make sample and population distributions equal.

Recently, Schouten et al. (2009) has proposed two definitions of representativeness with respect to survey response – strong (given in Definition 3.5) and weak (given in Definition 3.6). The proposed approach assumes that a probabilistic sample has been drawn, but the final outcome is ultimately determined by the selection mechanism. Information about sampled units and their participation is required in order to assess the representativeness of the response.

Definition 3.5. (*strong*) A response subset is representative with respect to the sample if response propensities ρ_i are the same for all units in the population:

$$\forall_i \; E(R_i) = \rho_i = P(R_i = 1 \mid I_i = 1) = \rho \tag{3.26}$$

and if the response of a unit is independent of the response of all other units (Schouten et al., 2009),

where R_i denotes the response of unit i and I_i is an indicator showing whether a unit took part in the survey. Schouten et al. (2009) notes that strong representativeness corresponds to the MCAR pattern for every target variable Y. It means that non-response does not cause estimators to be biased. Although the definition is appealing, its validity can never be tested in practice. To solve this problem a weaker definition of representativeness was introduced by Schouten et al. (2009).

Definition 3.6. (*weak*) A response subset is representative of categorical variable $\boldsymbol{x}$ with H categories if the average response propensity over the categories is constant:

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} \rho_{ih} = \rho, \text{for } h = 1, 2, ..., H, \tag{3.27}$$

where N_h is the population size of category h, ρ_{ih} is the response propensity of unit i in class h and summation is over all units in this category (Schouten et al., 2009).

The weak definition corresponds to the MCAR pattern with respect to $\boldsymbol{x}$, since MCAR means that it is not possible to distinguish respondents from non-respondents based on the knowledge of $\boldsymbol{x}$. The definitions proposed by Schouten et al. (2009) can be identified with *the absence of selective forces* and *representative sampling as permitting good estimation* proposed by Kruskal and Mosteller (1979a,b,c).

Finally, Pfeffermann (2011) discusses complex surveys and the MNAR mechanism in the context of the modelling process. Pfeffermann (2011) does not invoke the term *representativeness* directly, but emphasizes that under informative non-response the conditional distribution of target variable y in the sample and the population is not equal. Hence, after rewriting the concepts provided by Pfeffermann (2011), a *representative model* can be defined as:

Definition 3.7. A model is *representative* only when the conditional distribution of y given $\boldsymbol{x}$ is equal in the sample and in the population. $f_s(y_i|\boldsymbol{x}_i) = f_\Omega(y_i|\boldsymbol{x}_i)$ only if

$$Pr(R_i = 1 \mid \boldsymbol{x}_i, y_i, I_i = 1) = Pr(R_i = 1 \mid \boldsymbol{x}_i, I_i = 1), \tag{3.28}$$

because the conditional distribution of y given $\boldsymbol{x}$ can be expressed as:

$$\begin{aligned} f_s(y_i|\boldsymbol{x}_i) = f(y_i|\boldsymbol{x}_i, I_i = 1, R_i = 1) = \\ \frac{Pr(R_i = 1 \mid \boldsymbol{x}_i, y_i, I_i = 1) f_\Omega(y_i|\boldsymbol{x}_i)}{Pr(R_i = 1|\boldsymbol{x}_i, I_i = 1)}, \end{aligned} \tag{3.29}$$

where f denotes probability density function (PDF), $f_s(y_i|\boldsymbol{x}_i)$ refers to sample conditional PDF, $f_\Omega(y_i|\boldsymbol{x}_i)$ and I_i, R_i are defined as previously. In view of the

above, the definition based on Pfeffermann (2011) can be matched to *representative sampling as permitting good estimation* proposed by Kruskal and Mosteller (1979a,b,c).

Given the above concepts and definitions, let us discuss the possibility applying them to IDSs. First of all, for the purpose of clarity the following denotations listed by Kruskal and Mosteller (1979a,b,c) will not be studied in detail: *general acclaim for the data*, *typical or ideal case(s)*, *a vague term to be made precise* and *representative sampling good enough for a particular purpose*. The motivation is that these definitions do not directly refer to the representative sample in the context of survey methodology, but are rather general statements about how representativeness is perceived.

The concept of *the absence of selective forces* indicates that all units of the population are selected via probability sampling and no self-selection mechanism influences the probability of inclusion in the sample. In the case of IDSs, we do observe self-selection mechanisms. Units of $\mathbf{\Omega}_{TP}$ independently decide whether to use the Internet and register on a certain web portal. The self-selection mechanism can be associated with observed ($\boldsymbol{x}$), non-observed ($\boldsymbol{z}$) or target variables (y). Hence, understanding representativeness as the *absence of selective forces* is not suitable with respect to IDSs.

The second denotation considers a representative sample as a *miniature of the population*. The main assumption is that marginal distributions of auxiliary variables for the population are known, so they can be compared with sample distributions (Bethlehem, 2009). In the theory of survey methodology, a sample is drawn from a population sampling frame and, therefore, reference distributions are known in advance. A sampling frame may be a list of addresses or a population or business register. However, in the case of IDSs it may happen that data about the target population are unavailable. For example, Twitter, Facebook or Google Trends data may be used to identify the target population as people living in a given country. Therefore, we may have reference marginal distributions that come from this population. On the other hand, we may be interested in studying a population for which we have limited or no information about the distribution of auxiliary variables. This is precisely the case with the secondary real estate market: while information about sold properties is available, there are only limited, survey-based, data about apartments put up for sale.

Moreover, sample characteristics observed in IDSs are often unknown a priori and no aggregate data are available for assessment. In fact, these data sources are 'living organisms', which means that the distribution of $\boldsymbol{x}$ is likely to vary over time. For example, Google Trends provide information about the popularity of certain keywords in time and space but without demographic background information on those who made the searches. Facebook, in contrast, publishes demographic information about users and provides limited access to these data. In the case of IDSs about the real estate market only a limited number of characteristics are published, including the mean price per square meter classified by the flat size or the number of rooms, the fraction of flats by type and categorized according to floor area or the number of rooms. The scope of aggregate data provided by different IDS owners may vary and may not be consistent with available official statistics (different categories, definitions). Moreover, information provided on

websites may contain errors, which is why access to individual data is crucial for assessing whether distributions are consistent with population data.

Another denotation of representativeness mentioned by Kruskal and Mosteller (1979a,b,c) is *coverage of the target population*. Administrative sources, i.e. population registers, like sampling frames, are considered to fully cover the target population, mainly due to their obligatory character. However, there is evidence suggesting this may not always be the case Baffour et al. (2013), Coleman (2013), Gołata (2012), Wallgren and Wallgren (2014), and Zhang (2012b). For instance, the VAT register for short-term statistics has been assessed in this respect. For example Dehnel (2009) and Dehnel and Gołata (2012) studied the possibility of using register data for Polish business statistics and Dehnel (2015) discussed the possible use of the Social Insurance Institution register (pl. *Zakład Ubezpieczeń Społecznych*, ZUS) for business statistics. Ouwehand and Schouten (2014) and Scholtus et al. (2015) studied measurement error in the VAT register through linkage with survey data. Pavlopoulos and Vermunt (2015) evaluated measurement error in register data from the Dutch Institute for Employee Insurance with a linked LFS survey using the hidden markov model. Gołata and Dehnel (2013) and Józefowski and Rynarzewska-Pietrzak (2010) presented the possibility of using the PESEL register (Universal Electronic System for Registration of the Population, Pol. *Powszechny Elektroniczny System Ewidencji Ludności*) for population statistics. On the other hand, IDSs may also not fully cover the target population. Continuing the example of the real estate market, the question is how many flats that are put up for sale are presented online. However, to answer this question it is necessary to precisely define market participants and how they operate in the secondary real estate market.

In general, properties can be put up for sale by several market participants – local governments, housing cooperatives, brokers (individual or associated) and individuals (owners). Local government agencies are legally obliged to announce the sale of properties at auctions through the Public Information Bulletin website (PIB, Pol. *Biuletyn Informacji Publicznej*[1]). Therefore, there is an official online record of all properties offered by these entities. On the other hand, housing cooperatives are not required to use PIB, and rely on other channels to advertise properties for sale. Apartments can be sold between members of a housing cooperative or offered to a wider audience, in which case, the Internet is a popular option . Nonetheless, there is no official or non-official research on this matter and the Register of Real Estate Prices and Values does not cover transactions made by these units.

Another group of market participants are individual owners, who directly and independently offer properties in the market. The law does not regulate whether owners should inform the authorities about properties for sale, so there is no register of such activities. Moreover, owners can use several channels simultaneously to advertise their properties, including the Internet and advertisement portals. The decision is entirely up to the owner but depends on the owner's ability to use new technologies and the Internet, so the reliance on the Internet is highly correlated with the owner's age, which means that certain sub-populations may not be present online.

[1] See `http://www.bip.gov.pl`.

Brokers are the last and by far the biggest group of market participants. They work on the behalf of owners and can be classified into two groups – self-employed independent individuals and brokers employed by associations or enterprises (e.g. Home Broker). Moreover, in Poland there are two kinds of brokerage agreements between the owner and the broker – exclusive and open. Under an exclusive agreement only one broker or one association of brokers may put up properties for sale, while in the case of an open agreement, the same property for sale can be handled by any number of brokers. These two types of arrangements determine the choice of advertising channels and the use of the Internet for advertising purposes.

In Poland, unlike the Netherlands[2], there is no single leading association for real estate brokers. Consequently, there is no database or official register containing properties for sale. In addition, there is no information about the fraction of properties presented online by brokers and owners. However, it should be noted that this percentage is likely to be highly correlated with the Internet coverage, market liquidity and demand. On the other hand, specific types of properties (e.g. expensive, big) may not be offered to a wider group of customers but rather to a narrower group of potential buyers.

Finally, Internet coverage and use plays a crucial role in assessing the suitability of IDSs for statistics. Internet coverage is highly associated with the size of the city[3] and age groups. Thus, the Internet may have sufficient coverage in big cities and among young people while it is likely to be low in small towns and among older people. In addition, Internet use may vary between market participants: for instance local government agencies are obligated to place information online, while other participants are free to make their own decisions.

Representativeness is also understood as *a specific sampling method*, particularly when it is based on probabilistic assumptions. According to the survey theory, valid inferences about the target population can only be made only if a probabilistic sample is selected (all units have non-zero $\pi_i > 0$) . However, probabilistic sample surveys tend to have non-sampling errors (e.g. non-response) and the final sample consists of units that have decided to participate. For instance, in 2014 non-response in the Polish Household Budget Survey was over 54%, in the Polish EU-SILC nearly 26% and in the Labour Force Survey over 30%.

In the case of IDSs certain units may have $\pi_i = 0$ and, therefore, may not be included in the sample. The problem can be substantial when units with $\pi_i = 0$ differ from units for which $\pi_i > 0$. For this reason these units may never be present in the IDSs sample. However, units with $\pi_i = 0$ should be thoroughly investigated. In the case of the real estate market there is a lack of research on this matter. Moreover, given the absence of population frames, inclusion probabilities π_i are unknown in advance. One possible way to obtain π_i is to reweight the sample to known population totals suing the calibration approach to obtain weights w_i. A similar solution is used to obtain register-based statistics when register totals are not equal to population totals. It should be noted that these weights will be constant in subgroups, so it is assumed that each unit has the same probability of inclusion.

[2]For example see www.funda.nl.

[3]For instance, in Poland the rural-urban classification is used and towns are classified into 5 size groups – below 20K, 20-100K, 100-200K, 200-500K, over 500K citizens.

The last denotation of representativeness refers to the question of whether a sample/data source *permits good estimation*. This case is associated with definition 3.7 based on Pfeffermann (2011) and with the using of response propensity ρ, which can account for the self-selection mechanism. Hence, the following issues should be considered:

- information about other data sources may be required,
- strong auxiliary variables $\boldsymbol{x}$ that explain y should be available,
- auxiliary variables ($\boldsymbol{x}$, $\boldsymbol{z}$) that explain the selection and response mechanism should be available,
- the relationship between the selection / response mechanism and the target variable y should be checked,
- population totals or means for auxiliary variables (or proxies, as shown in Reilly et al. (2001)) should be known.

To sum up, in the case of IDSs the following concepts of representativeness should be considered in the assessment process. (1) Coverage of the target population and (2) comparison of sample and population distributions to determine what categories of units are observed online; (3) assessment of the selection mechanisms and estimation. The procedure to assess the representativeness of IDSs will be proposed in the next section.

3.3 A proposed two-step procedure to measure representativeness

3.3.1 A two-step procedure to measure representativeness

The suggested procedure to measure representativeness consists of two steps. Flow diagram 3.1 illustrates the general idea of the procedure. It starts with a more general question: *Step I: Is the Internet useful for providing statistics?* This question refers to the overall suitability of the Internet (or, more generally, big data) as a data source for statistics. The first step can lead on to Step II or the procedure ends with *END: Search for other source than the Internet*.

The second step *Step II: Internet data source(s)* focuses on a given data source and its representativeness with respect to existing statistical and non-statistical data sources. It starts with one IDS (or multiple integrated IDSs). The second step ends with the measurement of representativeness of IDSs in box *END: Measure representativeness* or with a suggestion regarding an existing data source *END: conduct/modify a survey or search for adm. data source*. This alternative outcome suggests that in order to assess representativeness: (1) an additional survey should be conducted, (2) additional questions concerning IDSs should be added to existing sources or (3) other administrative sources should be found (e.g. it is not available at a given time).

Diagram 3.1 consists of different figures to distinguish the steps of the procedure. Squares denote the start and the three possible outcomes of the measurement process. Diamonds denote questions and provide directions to further steps, letters A and B refer to different levels of the reference data source and numbers denote the order of steps. Blue diamonds marked with the letter A refer to statistical (census, sample surveys, reporting, statistical registers) and administrative sources (non-statistical data sources). These sources should be used directly or transformed to be used for statistical purposes (by NSIs). It is assumed that registers contain both statistical (e.g. persons) and non-statistical units (e.g. transactions, legal units), which can be transformed into statistical units. Red diamonds marked with the letter B refer to statistical and non-statistical sources that are only available as aggregate data at the domain level.

The term *reference data* will be used with respect to statistical or non-statistical data sources used for statistics to underline that these sources be used as a reference for comparisons with IDSs. The term *domain* will be used to describe the available level of reference data. The term *Individual data* refers to object-level (non-statistical unit) or unit-level (statistical) data. For simplicity, the term object-level is used to describe both actions (e.g. transactions) and objects (e.g. advertisement, units). Data processing and cleaning is not included in Diagram 3.1. However, it should be noted that this is a crucial stage of the procedure that affects the derivation of units or estimates based on new data sources (Daas et al., 2015).

The proposed procedure takes into account that the Internet (or, more generally, big data) may not be suitable or necessary to derive statistics (answer *no* to the question *Is the Internet useful for providing statistics?*). The proposed diagram also takes into account the fact that representativeness can be measured with respect to different populations. For instance, IDSs for the real estate market may contain information on properties as well as brokers. Thus, representativeness may be measured with respect to brokers (what fraction of brokers can be observed in IDSs), properties (what fraction of properties is offered online) or both (what fraction of properties is offered by brokers online). The second step of the procedure was partially inspired by the work done by Buelens et al. (2014). Buelens et al. (2014) provide a flow diagram to assess the selectivity of big data sources, which was adapted and modified to take into account several aspects that were not originally included in the working paper by Buelens et al. (2014) .

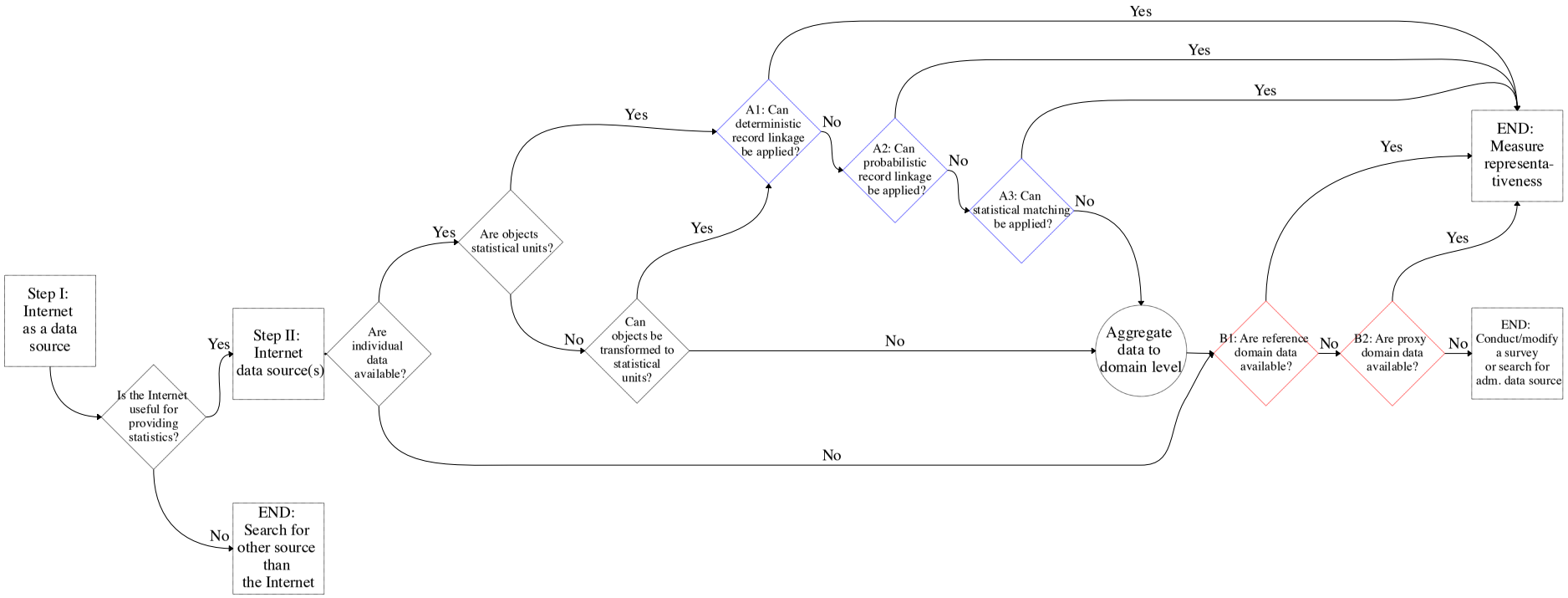


DIAGRAM 3.1: The two-step procedure to measure representativeness

Step I: Usefulness of the Internet for statistics

the first question *Is the Internet useful for providing statistics?* is used to determine whether the Internet can be used for statistics, particularly for official statistics. To answer this question, the following topics should be examined:

- What are the problems with existing statistical and non-statistical data sources?
- What kind of information do users of official statistics need and is it possible to obtain this information online?
- What is Internet coverage and use in the target population? Is the Internet a relevant source of information about the target population?
- For what purposes is the Internet used and is it important for the target population?

First, one should first identify problems with existing data sources. As far as surveys are concerned, the key issues include increasing unit non-response, panel attrition and high respondent burden (Brick, 2013; Groves, 2006). These problems should be further studied to find out whether the use of available online data can improve surveys. Another problem is whether the Internet can be used to obtain similar or new information. For instance, real estate market statistics in Poland are limited to quarterly and annual information with a very narrow scope. Advertisement services may be used to provide more timely and detailed information, but the quality of these data is unknown a priori.

The most important part of the first step are the last two questions. Internet coverage and use are crucial in determining whether IDSs can be considered an important data source. These questions may limit further work on the assessment of IDSs. For instance, data from countries with low or moderate Internet coverage or those where only specific units use the Internet are likely to be unsuitable for this approach.

In order to answer the question stated in the first step of the procedure data sources should be identified. Most NSIs conduct surveys or maintain administrative data sources which can be analysed to answer the questions. However, it should also be remembered that IDSs may contain data that are not yet available in official statistics. In such a situation data sources with proxy variables should be considered.

Data sources that can be used to assess the representativeness of IDSs NSIs of countries in the European Union are obligated to conduct the Information and Communication Technologies (ICT) survey. The survey is standardized and coordinated by Eurostat to produce coherent EU statistics. Results of the survey are published as part of statistics on Information Society and cover e-business, postal services, ICT in enterprises and households (Eurostat, 2015a). Other NSIs such as the US Census Bureau, the Australian Bureau of Statistics, Statistics Canada or the World Bank conduct similar surveys in this field. Because the Central Statistical Office in Poland is obliged to conduct the ICT survey as part of the Eurostat ICT survey, the main focus will be placed on this survey.

The Information Society survey is conducted annually in all Member States, as well as in two countries of the European Free Trade Association (EFTA), and acceding and candidate countries aspiring to join the EU. The data collection is based on Regulation (EC) 808/2004 of the European Parliament and the Council. The transmission of micro data to Eurostat was voluntary until the reference year 2010 and has been mandatory since 2011 (Eurostat, 2015a).

The ICT survey gathers information on the access to and use of ICT using two separate questionnaires: one for enterprises and another for households. The survey covers households with at least one member at the age between 16 and 74 and individuals aged between 16 and 74. As for enterprises, the target group includes companies employing at least 10 persons. The activity coverage is restricted to those enterprises whose principal activity is within NACE Rev. 2 Sections C through N, excluding Section K and Division 75 but including Group 95.1[4].

Information on access to ICT, e.g. type of Internet connection, is collected at household level, whereas statistics on the use of ICT, mainly on the use of the internet is collected from persons. In the case of enterprises, the focus is on data about the use of e-commerce and e-business technologies. The survey distinguishes between annual core subjects, which are included in the survey every year, and episodic topics on various ICT phenomena, which change in different survey years. The annual core subjects are:

- Access to ICT,
- Use of computers,
- Use of the Internet,
- eGovernment,
- eCommerce,
- eSkills.

The episodic topics in different survey years include:

- eGovernment (2006, 2013),
- Skills and digital literacy (2007, 2011),
- Advanced services (2008),
- e-Commerce and trust (2009),

[4]Enterprises classified in the NACE Rev.2 sections and groups: C – manufacturing; D,E – electricity, gas and steam, water supply, sewerage and waste management; F – construction; G - wholesale and retail trade, repair of motor vehicles and motorcycles; H – transportation and storage; I – accommodation and food service activities; J – information and communication; L –real estate activities; Division 69-74 – professional, scientific and technical activities; N – administrative and support activities; S95.1 – repair of computers (since 2010); K64.19+64.92+65.1+65.2+66.12+66.19 – financial and insurance activities (optional since 2011). See `http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Community_survey_on_ICT_usage_in_enterprises`.

- Internet security (2010),
- Mobile use of the internet and ubiquitous connectivity (2012),
- Cloud computing (2014).

In order to analyse the variables about the access and use of ICT in relation to characteristics of households, persons or enterprises, a number of background variables are collected. These include household composition, income and location as well as the age, gender, educational attainment and employment situation of persons. In the case of enterprises, the key variables include total turnover, main economic activity or the average number of employees. The ICT survey in different countries is slightly different and a detailed description is provided in the Methodological Manual (Eurostat, 2015b). Details for Poland will be provided below.

In Poland the ICT survey of households is conducted as follows: a probability sampling design is used, with 2 sampling stages and explicit stratification during the first stage by region and type of residence (urban/rural). In the first stage, Area Survey Points (ASPs) are selected using proportional-to-size sampling based on the number of dwellings in a given stratum. The sampling unit at the second stage is the dwelling. All individuals in each household are interviewed. The net sample size in 2014 was equal to 17 774 with a response rate of 90.7% (Eurostat, 2015b, p. 161).

The enterprise sample is stratified according to the following variables and breakdowns: (1) Economic activity (NACE Rev.2 – at two digit level, except when another aggregation is specified); (2) The number of employees (0; 1–4; 5–9; 10–49; 50–149; 150–249; 250 and more); (3) Turnover (up to 500k EUR; 500k to 25 mln EUR, 25 mln EUR and more). In addition, according to the methodological notes, in 2014 some categories of stratification variables (economic activity and the number of employees) were broken down based on the results of a 2013 study, which focused on the main indicators, namely Internet access and the company website. This procedure is designed to reduce the heterogeneity observed in some groups of enterprises. The following requirements are made - sample size according to the Neyman-Allocation; at least five enterprises in each stratum; a priori sampling error of 5% for each economic activity stratum and for each employment size stratum; sample selection according to random sampling; priority to enterprises with reduced respondent burden; all large enterprises are selected (enterprises with 250 and more employees or enterprises with turnover over 25 mln EUR). The net sample size in 2014 was equal to 14 948 with a response rate of 79% (Eurostat, 2015b, p. 81).

Finally, analysis of selected data sources should provide the basis for answering the question asked in the first step – *Is the Internet useful for providing statistics?* If the answer to the question is *yes*, then the procedure moves on to the second step, which focuses on particular data source(s). This step assumes that access to IDSs is possible and data from these sources are available. If the answer to the question is *no*, then the procedure ends with the suggestion that sources other than the Internet should be considered.

3.3.2 Representativeness of Internet data source

Step II: Representativeness of Internet data source

The main question to be answered in the second step is *are specific IDSs representative of the target population?* In addition, this is related to a practical question: *given the available data, how to measure the representativeness of IDSs?* The diagram is separated into two parts that are distinguished by question *Are individual IDS data available?* The blue diamonds in Diagram 3.1 refer to individual data that are made available for the statistician, while the red diamonds represent aggregate data that are available or are shared with the statistician.

Case 1: only domain-level data are available The case when individual data are not available is selected when the answer to the question *Are individual IDS data available?* is *no*. The answer leads to question B1, which verifies whether reference data are available at the domain level (*B1: Are reference domain-level data available?*). To answer question B1 the following conditions should be taken into account. Data can be made available to the statistician (1) at the level of aggregation predefined by the statistician/NSI; (2) at the level defined by the data source owner.

The first case refers to the situation when the statistician/NSI collaborates with the data owner. Access to individual data is not possible due to privacy or practical aspects (e.g. volume of the data). For instance, for the purpose of the present study the author established collaboration with three data owners. The OtoDom advertisement service provided historical data in the form of predefined aggregates. These levels were more detailed than currently available official statistics. Table 3.1 presents sample rows and columns from the data file that was made available by OtoDom.pl. A detailed description of the data sources is provided in Chapter 4 and Chapter 5.

TABLE 3.1: Sample rows and columns from the dataset prepared by OtoDom.pl

City	Month	Floor area [m^2]	Rooms	Count	Average price per m^2
Poznań	2011-01	to 20	1	5	4363.2
Poznań	2011-01	20-30	1	285	6399.3
Poznań	2011-01	30-40	1	577	6370.3
Poznań	2011-01	40-50	1	56	5695.5
Poznań	2011-01	50-60	1	12	4210.3
...	...	...	...	...	...

Note: compiled using data from OtoDom.pl. City – the name of the city, Month – the reference month with the year, Floor area – floor area divided into intervals, Rooms – the number of rooms, Count – the number of objects that meet the aggregation criterion.

Another example of such aggregates are publicly available indices, for instance accessible through special services. The most known service is Google Trends, which provides information on the popularity of words and terms searched

through Google. Google Trends classify queries into groups using algorithms developed by Google and provide indices at levels defined by Google (e.g. provinces in Poland, or classification of searches in Google). Continuing with the example of the real estate market, in Poland several online advertising services provide price statistics and indexes based on ads[5]. It should be noted that domain-level data do not allow the statistician to fully assess data quality. Calculations are made by the data owner, who does not always provide information about data processing or the cleaning process or the methodology applied.

Therefore, the answer *yes* to the question *B1: Are reference domain-level data available?* represents two possible cases. Reference data are available at the same level of aggregation as IDSs or both data sources should be harmonized before they can be compared. In this case the procedure ends with *END: Measure representativeness*. Otherwise, the question *B1: Are proxy data available?* should be answered. Proxy data can refer to the same population but contain a variable with a different definition or a variable with a similar definition but for a different population. Examples of proxy data for IDSs include transactions in the real estate market or primary market characteristics (when the secondary market is considered). Methods to assess representativeness at domain level will be discussed in the Section 3.4.

In the case when proxy domain-level data are available, the answer *yes* to question B1 leads to *END: Measure representativeness*. However, it should be noted that due to the character of reference data (differences in concepts or populations) there may be differences between IDSs and official statistics. In this case, a comparison of data sources can involve visual identification of common time or spatial trends. If time series data are available, it may be possible to verify whether the structure of time series is similar (e.g. seasonality) or whether there is co-integration between IDSs and proxy data.

The measurement of representativeness at domain level is not straightforward. For instance, if a dataset was prepared by the IDS owner, the statistician/NSI has no control over the data cleaning process or calculations. This lack of control may result in unit or measurement errors. Moreover, it may happen that time series data for different periods should be compared. For instance, NSIs provide information on a monthly, quarterly or annual basis, while IDS-based statistics are available on a daily or intraday basis. This poses another problem of how to aggregate IDS data. There may also be a time shift between official data and IDSs. Another issue is that owing to the character of official data (census, survey, reporting, or statistical registers) their aggregation level may be limited to the country, regional or city level. Such a limit can reduce the variance of estimates and will not capture differences between regions or cities. Finally, and most importantly, aggregate data may be insufficient to capture the selection process. Aggregate data can only provide an approximation of this mechanism. The measurement of representativeness and possible approaches will be presented in Section 3.4.

Case 2: individual data are available This paragraph focuses on individual data that are available for IDSs and reference data. The question *Are objects*

[5]For instance, see `https://ceny.szybko.pl/ceny-nieruchomości` or `http://www.morizon.pl/ceny`

statistical units? attempts to verify whether IDSs contain statistical units (e.g. persons, properties, establishments) or non-statistical units (e.g. advertisements, legal units, transactions). The answer to this question opens to possibilities. If the answer is *no* the next question is *Can objects be transformed into statistical units?*. If the answer is *yes*, then the path with blue diamonds is selected.

In the case of IDSs, the answer to the question *Are objects statistical units?* is more likely to be *no*. For instance, advertisement services contain ads that refer to statistical units or composite units (composition of statistical units); accounts on social media might refer to individual persons or groups of people. The following question *Can objects be transformed into statistical units?* is intended to find out whether it is possible to match objects to statistical units. This part is the most challenging aspect of IDSs. For one thing, multiple records can refer to the same statistical unit. For example, in the real estate market it is common for one flat to be presented for sale a number of times. Moreover, information provided or collected by statistician s can be limited to protect privacy and is not sufficient to identify statistical units. It can be argued that the following statistical units present online can be identified quite easily: (1) products (e.g. groceries, electronics) or services, (2) vehicles (e.g. using VIN number), (3) job offers, (4) properties (but it depends on the market and country) or (5) establishements/enterprises. On the other hand, identification based on the following services can prove challanging: social media (e.g. for Facebook or LinkedIn it may be easier than for Twitter) or query data (e.g. who has made a given query).

Transforming objects into statistical units should also be done for registers. For instance, the natural choice of in the case of the real estate market in Poland is the Register of Real estate Prices and Values. The register contains data about transactions in the primary and secondary market. Table 3.2 contains selected variables from the Register of Transactions. Note that transactions can refer to one or several properties. For instance, one transaction can refer to a property with a garage and a storage room. The R egister of Transactions and other registers about the real estate market in Poland will also be discussed later as a potential source for linkage with IDSs.

It should be taken into account that transforming objects into statistical units may not be possible. In such a situation, the measurement of representativeness can be done only at the aggregated level based on non-statistical units. This leads to the answer *no* to the question about the possibility of transformation and the circle *Aggregate data to domain level* in Diagram 3.1. The aggregation process should be followed by determining possible levels of comparison with official statistics and data cleaning in order to derive characteristics of statistical units. Possible approaches to deriving statistical units from non-statistical units are presented in Zhang (2011) and Wallgren and Wallgren (2014, ch. 7). Another interesting approach is profiling (Chandrasekaran et al., 2012; Daas and Burger, 2015; Daas et al., 2015; Flekova and Gurevych, 2013). After data aggregation, the process continues along the path for domain-level data presented in Section 3.3.2.

The answer *yes* to the question *Can objects be transformed into statistical unit?* leads to the same path as the positive answer to the question *Are objects statistical units?*. In this path three methods of linking IDS with reference data

TABLE 3.2: Selected variables available in the Register of Real estate Prices and Values

Variable	Description
Date of transaction	Exact date of transaction
Transaction ID	Transaction ID which may refer to several properties (e.g. flat and garage)
Object ID	Objects that were included in the transaction
Total Price	Total transaction price (for all objects)
Object Price	Prices of objects included in the transaction
Use of Property	Intended use of the property (e.g. residential, non-residential)
Mortgage Number	The document number in the mortgage register
City	Name of the city where the property is located
Address	Precise information about location with street name, floor number, etc
Floor area	Floor area of a given property measured in square meters
Number of Rooms	The number of rooms in a given property
Transaction market	Primary, secondary market or auction
Seller	Information about the seller: legal or natural person
Buyer	Information about the buyer: legal or natural person

Note: compiled on the basis of the Register of Real Estate Prices and Values.

are considered: deterministic record linkage, probabilistic record linkage and statistical matching/data fusion.

Deterministic record linkage The question *A1: Can deterministic record linkage be applied?* is designed to find out whether data can be linked using common identifiers. This type of linkage considers the case when IDSs and reference data contain the same units and identifiers are present in both sources. For instance, when a legal or natural person registers business activity, they receive a REGON identifier (ang. *National Official Business Register*, Pol. *Rejestr Gospodarki Narodowe*). Thus, hypothetically, it should be possible to link units from IDSs and the REGON register using the REGON ID. However, in the case of properties, no common identifiers that cab be used for linkage purposes are available in the statistical system. Advertisement services use different IDs that can be used for deterministic linkage. Examples of such codes used in IDSs about the real estate market are listed in Table 3.3.

The positive answer to question A1 leads to the measurement of representativeness. Possible measures for individual data are presented in Section 3.4. Nonetheless, owing to privacy restrictions or the policy of IDS owners this type of linkage is rarely possible in practice, which results in the more likely answer

TABLE 3.3: Example ID codes used in IDSs

Data source	Example ID	Comment
OtoDom.pl	41534981	8 digits code (possibly an additional ID given by broker)
Dom.Gratka.pl	gratkaBZMMS13039	gratka and additional codes with alphanumeric characters
Domiporta.pl	475597; 141688242	Code assigned by a broker or 9 digits code in the link
Morizon.pl	morizon1/W; mzn2019470302	Code assigned by a broker or alphanumeric character code in the link
Domy.pl	domy1515291128; dol1717311330	"domy-" with 10-digit code and alphanumeric character code in the link
Nieruchomosci-Online.pl	ms_241111 or 14988720	"ms_" with 6-digit or 8-digit code

no to question A1 and is followed by the question *A2: Can probabilistic record linkage be applied?*.

Probabilistic record linkage Question A2 refers to probabilistic methods for linking records from IDSs and reference data. This type of linkage does not require common identifiers but a set of common variables that are present in both data sources. The theoretical foundations of the method are presented in Fellegi and Sunter (1969) and are still being developed ((cf. Winkler, 2006)). The general idea consists in assigning a linkage weight that refers to the probability that two units from datset $\boldsymbol{A}$ (for example IDS) and dataset $\boldsymbol{B}$ (for example census) refer to the same unit. Details of the method can be found in Fellegi and Sunter (1969) and Lavallée and Caron (2001) and R implementation can be found in *RecordLinkage* package Borg and Sariyar (2015) and Sariyar and Borg (2010).

Probabilistic record linkage is used in official statistics. For instance, when the system of registers is used for statistical purposes (Chambers, 2009; Christen, 2012; Consiglio and Tuoto, 2015; Roszka, 2013; Sariyar et al., 2012; Statistics Finland, 2004; Yildiz and Smith, 2015; Zhang, 2012b) or census and survey data are linked (Knottnerus and Duin, 2006; Lavallée and Caron, 2001; Renssen et al., 2001). However, the underlying theory for methods based on probabilistically linked data is currently the subject of ongoing research (cf. Consiglio and Tuoto, 2015; Kim and Chambers, 2012; Samart, 2011; Samart and Chambers, 2014).

In the case of the real estate market, probabilistic record linkage can be used as follows: NBP/CSO conduct a survey of brokers, which covers both the primary and secondary market. Brokers report information about properties offered for sale, but the questionnaire does not include a question about which properties are offered online. Assuming that brokers provide full information about their offers it may be possible to link IDS data with (1) brokers to check which brokers use

the Internet to publish information about properties, or (2) properties with those reported by brokers.

Moreover, in Poland there are several registers that could be used for purposes of data linkage. There is already the Integrated Cadastral Register (ICR, Pol. *Zintegrowany system informacji o nieruchomościach*), which consists of the Register of Real estate Prices and Values (Pol. *Rejestr cen i wartości nieruchomości*), Mortgage Registers (Pol. *Księgi Wieczyste*) and the Land and Buildings register (Pol. *Ewidencja Gruntów i Budynków*). In addition, the following registers may be considered: Tax registers (e.g. VAT or NIP, Pol. *Numer Identyfikacji Podatkowej*) or business registers (CEIDG, Pol. *Centralna Ewidencja i Informacja o Działalności Gospodarczej*). On 13^{th} June 2013 several professions (including brokers) were deregulated(The Real Estate Management Act, 1997). Under the new regulations no professional licence is required to conduct brokerage activity. Thus, since 2013 there has been no official register of brokers in Poland. In this situation the statistical register REGON could be used instead. REGON covers: legal persons, organisational units without the status of a legal person, natural persons conducting economic activity. The scope of the REGON register makes it possible to link brokers and real estate agencies that use advertisement services. This can be done by using the company or person's name and address.

Taking into account the above mentioned registers, the Register of Real Estate Prices and Values seems to be the most suitable one. The register records transactions made both in the primary and secondary market as well as those made by local governments at auctions. There are several reasons for choosing the Register of Transactions: (1) advertisement services devoted to the real estate market contain information about properties put up for sale; (2) the register covers all transactions involving fully-owned properties (properties owned by housing cooperatives are excluded) and linkage may help to identify units that are not observed in IDSs; (3) information in the register can be used to identify objects; (4) data stored in the register are unbiased and free of measurement error (information is taken from the Mortgage Register); and (5) most importantly, due to the new legislation that came into effect on 12^{th} of July 2014 researchers form Polish Universities can acquire register data for research and teaching purposes free of charge (The Geodetic and Cartographic Act, 1989, art. 40a par. 2 pt 2). Table 3.2 presents possible variables that could be used for linkage.

To sum up, the Register of Real Estate Prices and Values is the most suitable data source for linkage with IDSs. Objects that cannot be linked to the Register of Transactions can provide information about the characteristics of properties that are not presented online. This approach can help to identify the selection mechanism, because the Register of Transaction contains transaction prices. Hence, it may be prove useful in determining whether the response mechanism follows the MNAR pattern. Finally, if probabilistic linkage is not possible, the question *A2: Can probabilistic record linkage be applied?* is answered negatively, which leads to the question *A3: Can statistical matching be applied?*.

Statistical matching/data fusion Statistical matching/data fusion (D'Orazio et al., 2006; Moriarity and Scheuren, 2001; Rässler, 2012; Rubin, 1986) assumes that there are two (or more) data sources containing statistical units from the same

target population. Following the definitions provided by D'Orazio et al. (2006), there are two data sources denoted by $\boldsymbol{A}$ and $\boldsymbol{B}$, which share a set of variables $\boldsymbol{x}$, and variable $\boldsymbol{y}$ is available only in $\boldsymbol{A}$ and variable $\boldsymbol{z}$ is only present in $\boldsymbol{B}$. Variables $\boldsymbol{x}$ are shared by both data sources, while variables $\boldsymbol{y}$ and $\boldsymbol{z}$ are not observed jointly.

Statistical matching (SM) consists in investigating the relationship between $\boldsymbol{z}$ and $\boldsymbol{y}$ at "micro" or "macro" level. At micro level, the aim of SM is to create a "synthetic" data source in which all the variables, $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$, are available (usually $\boldsymbol{A} \cup \boldsymbol{B}$ with all the missing values filled in or simply $\boldsymbol{A}$ filled in with the values of $\boldsymbol{z}$). At macro level, the data sources are integrated to derive an estimate of the parameter of interest, e.g. the correlation coefficient between $\boldsymbol{y}$ and $\boldsymbol{z}$ or the contingency table $\boldsymbol{y} \times \boldsymbol{z}$. SM methods assume *conditional independence* of $\boldsymbol{y}$ and $\boldsymbol{z}$ given $\boldsymbol{x}$, which, as D'Orazio et al. (2006) notes, is strong and seldom holds in practice.

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = f(\boldsymbol{y} \mid \boldsymbol{x}) \times f(\boldsymbol{z} \mid \boldsymbol{y}) \times f(\boldsymbol{x}). \tag{3.30}$$

As a result of the matching procedure, correlation and contingency tables can be used to compare distributions between $\boldsymbol{y}$ and $\boldsymbol{z}$ when jointly observed. The implementation of statistical matching in the R statistical software is described in `StatMatch` package (D'Orazio, 2015; D'Orazio et al., 2006). The basic statistical matching/data fusion technique aims at obtaining a complete dataset based on the input datasets.

However, for purposes of measuring representativeness, the idea of using statistical matching should be considered in two situations: (1) create a complete dataset to provide information about the differences between $\boldsymbol{y}$ presented online (IDSs) and $\boldsymbol{z}$ that comes from reference data; (2) link units between sources using non-parametric measures and evaluate linkage quality (e.g. matching by propensity score, (Rässler, 2012, chap. 2.3-2.4), using distance functions in the KNN approach, (D'Orazio et al., 2006, chap. 2-3)). To link units Gower (1971) one can consider dissimilarity measures. The final dissimilarity between the i-th and j-th unit is obtained as a weighted sum of dissimilarities for each variable and is given by the following equation:

$$d(i,j) = \sum_k (\delta_{ijk} \times d_{ijk}) / \sum_k (\delta_{ijk}), \tag{3.31}$$

where d_{ijk} represents the distance between the i-th and j-th unit computed considering the k-th variable and $\delta_{ijk} = 0$ when $x_{ik} = NA$ or the variable is asymmetrically binary and $\delta_{ijk} = 1$ otherwise. The measure depends on the nature of the variable – logical, categorical, ordered or numeric. Then, for each unit from the statistical and Internet source we obtain a matrix of distances and compute the overall distance given by (3.31). Possible measures based on different approaches will be presented later in the chapter. When the use of deterministic or probabilistic record linkage or statistical matching/data fusion is not possible (negative answers to A1, A2 and A3, then aggregation to domain level is the recommended solution.

The final outcome of the proposed procedure is the measurement of representativeness. In order to assess the representativeness of IDSs the following questions should be investigated:

- **the coverage of the target population** – analysis should involve the evaluation of coverage, this can be done by using existing data sources on Internet access. This part of analysis refers to representativeness considered as *coverage of the target population* (Kruskal and Mosteller, 1979a,b,c).

- **comparison of distributions** – assessment should aim at identifying differences in the distribution of $\boldsymbol{x}$ variables. This aspect refers to representativeness understood as *a miniature of the population* (Kruskal and Mosteller, 1979a,b,c).

- **estimation of bias in** y – assessment of bias in the target variable can help to determine whether weighting procedures are able to account for the response mechanism or whether bias is associated with the target variable itself. This corresponds to representativeness perceived as *representative sampling as permitting good estimation* (Kruskal and Mosteller, 1979a,b,c).

- **detection of the selection/response mechanism** – this task refers to the following denotations *the absence of selective forces* (Kruskal and Mosteller, 1979a,b,c), *representative model* (Pfeffermann, 2011) and *representative response* (Schouten et al., 2009).

The above areas of investigation should enable the statistican to choose the method of estimation, either using calibration or other weighting schemes, or perhaps opt for the modelling of the response mechanism involving the MNAR pattern.

3.3.3 Summary and concluding remarks

A two-step procedure to measure representativeness has been proposed and discussed. The second step of the procedure has two possible outcomes – the measurement of the representativeness of IDSs or the need for new data to assess IDSs. The second case can occur in the absence of reference data for comparison at a given time. A possible solution is to conduct a new survey or extend existing surveys by adding new questions that focus on IDSs. Table 3.4 contains a summary of advantages and disadvantages of using individual and domain data for the measurement of representativeness.

NSIs are aware of the significance of new data sources. For instance, the ICT survey covers the use of social media, ownership of websites (question C9, C10, form SSI-01[6]) in enterprises and households (CSO, 2014a, 2015b). The Household Budget Survey (Pol. *Badanie Budżetów Gospodarstw Domowych*, BBGD) collects data from households which record every expense made and should indicate whether a given product was bought via the Internet or in the traditional way (BR-01[7]) (CSO, 2014b).

[6] See http://form.stat.gov.pl/formularze/2015/passive/SSI-01.pdf

[7] See http://lodz.stat.gov.pl/gfx/lodz/userfiles/_public/pliki/inne/201408_d_br01.pdf

TABLE 3.4: Advantages and disadvantages of measuring representativeness using individual and aggregate data

Aggregation level	Advantages	Disadvantages
Individual data	1. Measuring representativeness at any level	1. Limited access to individual data
	2. Detection of the selection mechanism	2. Time consuming data cleaning process
	3. Control over data processing and cleaning	3. Linkage uncertainty
	4. Linkage with units or objects from statistical and non-statistical data sources	4. Linkage may be legally prohibited or impossible
	5. Asessment of uncertainty of estimates	
Domain data	1. Alternative when individual data are not available	1. Limited possibilities of measuring representativeness and the selection mechanism
	2. Overall information about consistency with official and non-official data	2. Requires historical time series data for comparison
	3. May provide a general overview of the data without time consuming data cleaning process	3. Requires harmonization with available domain-level data
	4. May indicate whether the use of such data is possible for official statistics	4. Lack of uncertainty of estimates (also for comparison)

Taking the above into account, the NBP/CSO survey of brokers (NBP, 2014a,b) could be extended by adding two questions: (1) is a given property being advertised online?; (2) where are advertisements published (own web page, advertisement service)? The first quesiton would be used to determine the fraction of properties presented online, the second would show the use of online services. In addition, the questionnaire about properties could be extended by adding a column for the ID assigned by the broker[8].

Another possible solution presented in Diagram 3.1 is to search for new administrative data sources that were not available at a given time. However, in the Polish context, when the Register of Real Estate Prices and Values is already available, there is no need to look for additional information.

To assess the representativeness of new data sources (including IDSs) time

[8]Questionnaires are available online, but only in Polish. Accessed 15.11.2015: `http://www.nbp.pl/home.aspx?f=/publikacje/rynek_nieruchomosci/ankieta.html`

series of historical data should be considered. A comparison over time may reveal a similarity (in level or trend) to existing data sources. For instance, Daas et al., 2015, p. 255 conducted a monthly comparison between social media sentiments and Dutch consumer confidence between 2010 and 2012, Beręsewicz and Szymkowiak (2015) compare weekly Google Trends, monthly registered unemployment rate and the rate of monthly average paid employment in enterprises in Poland between 2005 and 2015. Possible measures based on different levels of available data will be studied in the next section.

3.4 The measurement of representativeness

In this section a number of possible measures of representativeness are listed and discussed. The measures have been selected taking into account the definitions of representativeness that should be considered for IDS. Two groups are discussed – measures based on individual and domain-level data.

3.4.1 Measures based on individual data

Naive coverage measures Let us assume that an IDS is of size $n_{IDS,t}$ in period t and reference register data cover all units of the population of size $N_{REG,t}$ in period t. It is assumed that the only market participants are brokers or there is a list of properties offered for sale. In such a simple setting a naive measure of representativeness with respect to coverage is given by equation (3.32).

$$p_{IDSs,t} = \frac{\sum_{j=1}^{N_{r,t}} L_j}{N_{r,t}}, \tag{3.32}$$

where L_j is an indicator, which is equal to 1 when there is a link between j-th unit from the register and i-th unit from IDSs; or is 0 otherwise. This measure denotes the fraction of brokers/properties that are observed online - representativeness is treated as the coverage of the target population. When information for H strata is required, the measure is given by equation (3.33).

$$p_{IDSs,h,t} = \frac{\sum_{j=1}^{N_{r,h,t}} L_j}{N_{r,h,t}}. \tag{3.33}$$

When the IDSs is deterministically linked to survey data $N_{r,t}$ in equation (3.32) and $N_{r,h,t}$ in (3.33) are replaced with the Horvitz-Thompson estimator given by $\hat{N}_{s,t} = \sum_{j=1}^{n_{s,t}} w_j$ and $\sum_{j=1}^{N_{r,t}} L_j$ is replaced with a weighted sum $\sum_{j=1}^{n_{s,t}} L_j w_j$. In the above, it is assumed the survey or the register fully covers the target population. In the case of the register, only brokers are covered, although other participants may also operate in the market. Hence, $N_{IDSs,t}$ should be estimated and one possible solution is to use Capture-Recapture methods, particularly the Petersen estimator.

The assumption underlying the classic Petersen estimator is that we have two independent data sources that can be linked without error. If deterministic linkage is possible, there are no linkage errors and the two data sources are assumed to

be free of errors. Hence, the Petersen estimator can be suitable in this case and is given by equation (3.34)

$$\hat{N}_t = \frac{n_{IDSs,t} n_{r,t}}{n_{IDSs,r,t}} \tag{3.34}$$

where $n_{IDSs,t}$ denotes the number of units observed in IDSs in period t, $n_{r,t}$ is the number of units in register data in period t and $n_{IDSs,r,t} = \sum_{j=1}^{n_{r,t}} L_j$ the number of linked units. Thus, $N_{r,t}$ in equation (3.32) is replaced with $\hat{N}_t$. When IDS are linked directly with survey data $n_{s,t}$ is replaced with $\hat{N}_t = \sum_{j=1}^{n_{s,t}} w_j$.

The measures presented above, particularly calculated for strata, can be usd to assess the coverage of IDSs. The use of measures based on stratification could provide additional information about the selection mechanism. However, to fully use linked sources other methods should be considered.

Measures based on propensity scores The purpose of using IDSs is to provide reliable statistics that account for the self-selection error. Linkage of IDSs and register or survey data may provide information about the response model, particularly whether it follows the MNAR pattern. For that purpose one can use R-indicators. In general, R-indicators are used to monitor the data collection process and modes (Schouten et al., 2011, 2012). However, there are also applications in short-term business statistics, for instance Ouwehand and Schouten (2014) used propensity scores as a measure of probability that the value of $\boldsymbol{X}$ observed in the register is equal to the value observed in a short-term business survey. The variance of this indicator can be estimated using bootstrap, a detailed description can be found in Shlomo and Schouten (2013) and Shlomo et al. (2012).

Let L_j denote whether unit j from the reference data source was linked with j-unit from the IDSs. Let $\boldsymbol{x}$ denote a matrix of auxiliary variables (linkage variables), for instance demographic characteristics of the broker or characteristics of the flat. Therefore, in such a setting the R-indicator is given by:

$$R(L_j = 1 \mid \boldsymbol{x}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i}^{N}(\rho_i - \bar{\boldsymbol{\rho}})^2}, \tag{3.35}$$

where ρ_i denotes the probability that unit j from the reference source was linked. Moreover, if the register contains target variable $\boldsymbol{y}$ or proxy variable $\boldsymbol{v}$ then the R-indicator is given by the following equation:

$$R(L_j = 1 \mid \boldsymbol{x}, \boldsymbol{y}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i}^{N}(\rho_i - \bar{\boldsymbol{\rho}})^2}, \tag{3.36a}$$

$$R(L_j = 1 \mid \boldsymbol{x}, \boldsymbol{v}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i}^{N}(\rho_i - \bar{\boldsymbol{\rho}})^2}. \tag{3.36b}$$

Estimation of ρ was discussed in Section 3.1.2. R indicators presented above can be used to determine whether the self-selection follows the MNAR pattern with respect to variables y or v.

Another possible approach to measuring representativeness is propensity score matching (Rässler, 2012; Rosenbaum and Rubin, 1983). The idea of using propensity score matching to measure representativeness can be explained as follows: two data sources are distinguished - IDSs and reference data - and they refer to the same population. Let L_j denote an indicator variable that is equal to 1 if a given unit j is observed in IDS and 0 if a given unit is observed in reference dataset. The main goal of using propensity score matching to measure representativeness to estimate the probability that a given unit is observed in the reference dataset $P(I_j = 1 \mid \boldsymbol{x})$. Table 3.5 presents an example concatenation of reference and IDS data. It is necessary to harmonize common variables, such as floor area (in m^2), the number of rooms or the floor number.

TABLE 3.5: IDSs and refrence data combined for propensity score weighting

ID	Price	Floor area	Rooms	Floor number	L
1	200 000	85	4	11	1
2	340 000	65	3	5	1
3	555 000	47	2	1	1
4	289 000	24	1	0	1
5	251 000	55	3	2	1
...	...	...	...	...	...
848		81	5	10	0
849		37	2	14	0
850		99	7	1	0
851		20	1	2	0
852		45	2	5	0
...	...	...	...	...	...

The measurement of representativeness will be correct if IDSs and reference data refer to the same population. However, it is rarely possible to link data from exactly the same population. For instance, in the Polish context the obvious choice from the point of view of representativeness would be the Register of Transactions. The target population in this case includes fully-owned properties that have been sold, while IDSs contain properties that can (but may eventually not) be sold. Naturally, there will be differences between these populations – depending on the market and location, offer prices will be higher than transaction prices, the market structure is likely to be different due to the overrepresentation of certain flats.

However, if we link IDSs and the Register of Transactions, we will be able to see what kind of flats that have been sold are not observed online. Let us now consider whether this outcome will be suitable for purposes of measuring representativeness. Firstly, the outcome will be limited only to the register population and the relation to the target population (all properties sold) should be verified. Secondly, linkage will show what flats are not observed online and it will be the

outcome that we are looking for. If the register covers all transactions and a fraction cannot be linked with IDSs, then we may be able to check whether these units differ from the ones that were linked. Thirdly, one usually considers one-to-one linkage , but IDSs contain multiple occurrences of the same property in one or more sources.

The above description focuses on deterministic linkage, but it is more likely that probabilistic methods will have to be applied. As a result, model parameters ρ_i will be biased due linkage errors (Chambers, 2009). In such a setting a correction for linkage errors should be made. Possible solutions are presented in Chambers (2009), Kim and Chambers (2012), Samart (2011), and Samart and Chambers (2014).

Measures based on fused data The last group of measures are based on statistical matching/data fusion. Let $\boldsymbol{D}$ be distance matrix calculated for each unit i-th unit from the reference data and j from the IDSs. The matrix $\boldsymbol{D}$ is of size $n_r \times n_{IDS}$, where n_r refers to the observations in reference official/non-official data source, and n_{IDS} to number of observations in IDSs, is given by (3.37). Each element of $\boldsymbol{D}$ matrix is defined by the distance function (for instance, Gower (1971) distance or other used in statistical matching) that in $[0, 1]$ range. Calculation of $\boldsymbol{D}$ is the first step for the representativeness measurement under assumption that the reference official/non-official data contain all possible units from the target population.

$$D_{n_r,n_{IDSs}} = \begin{pmatrix} d(1,1) & d(1,2) & \cdots & d(1,n_{IDSs}) \\ d(2,1) & d(2,2) & \cdots & d(2,n_{IDSs}) \\ \vdots & \vdots & \ddots & \vdots \\ d(n_r,1) & d(n_r,2) & \cdots & d(n_r,n_{IDSs}). \end{pmatrix} \tag{3.37}$$

In the next step, for each i-th unit given all units from IDSs. Thus, following vector is created according to (3.38), with restriction that only one observation form IDSs could be matched to reference (register, sample) data.

$$\boldsymbol{d} = \{d_i = \max_{1 \leq j \leq n_{IDSs}} d(i,j)\}. \tag{3.38}$$

Other approaches to restricted matching could be found in (D'Orazio et al., 2006, chap. 4.5). Measure (3.38) further could be used to assess the quality of matching between reference data source and IDSs. For instance, records that are close to 0 are present in IDSs, while those with distance near 1 may denote units that are rarely.

3.4.2 Measures based on domain data

The last group of measures are based on domain data. The main idea is to compare point estimates between IDSs and reference data. Two scenarios should be considered: (1) only point estimates are available for both sources and (2) point estimates contain information about precision (for instance sampling variance). This problem is crucial in the context of domains that are characterised with small sample size. For instance, most survey-based research have high precision at country

level, while IDSs data are often available for more detail domains. For these domains direct estimate based on surveys might be unreliable (Rao, 2003).

Estimation of variance in surveys was widely studied in the literature (see Wolter, 2007). However, measurement of uncertainty in new data sources, was not studied before. The reason for that is due to several issues: (1) non-probability character of data, (2) lack of knowledge how the data is generated, (3) unknown characteristics of data, (4) unknown errors, and (5) massive character of the data (in comparison to samples). Uncertainty of IDSs estimates should be not neglected due to sample character.

Unfortunately, in practice, only point estimates are available (e.g. from registers). Reference data could be obtained from several statistical/non-statistical data sources. For instance, Central Statistical Office in Poland provide data from the survey or register through special platform Local Data Bank[9] on yearly and quarterly estimates. In addition, several other statistics are available on CSO wepage in Excel format. Eurostat provide access to data through online database[10], which could be also obtained by `eurostat`[11] package (Lahti et al., 2014).

Representativeness measurement depends on availability of (historical) time series data. Comparison of IDSs and reference data in time allows to detect similarities. For instance, characteristics of time series data (e.g. trends, seasonality) could be compared to detect whether systematic bias is present.

To compare distributions the following measures could be used (D'Orazio, 2015; D'Orazio et al., 2006). *Dissimilarity index* or *total variation distance* is given by equation (3.39)

$$D = (1/2) \times \sum_j \mid p_{1,j} - p_{2,j} \mid, \tag{3.39}$$

where $p_{s,j}$ are the relative frequencies of j-th class level ($p_{s,j} \in [0,1]$), $s \in \{1,2\}$ denote sample sample membership. For instance $s = 1$ might denote one of $IDSs$, while $s = 2$ denote reference official or non-official domain data. The dissimilarity index ranges from 0 (minimum dissimilarity) to 1. It can be interpreted as the smallest fraction of units that need to be reclassified in order to make the distributions equal.

Another measure is *overlap between two distributions* given by equation (3.40)

$$O = \sum_j min\{p_{1,j}, p_{2,j}\}, \tag{3.40}$$

where $p_{s,j}$ are defined as in (3.39). Overlap measure ranges from 0 to 1 (the distributions are equal). It should be noted that $O = 1 - D$.

Third proposed measure is Bhattacharyya coefficient given by equation (3.41)

$$B = \sum_j \sqrt{p_{1,j} \times p_{2,j}}, \tag{3.41}$$

[9] See `http://stat.gov.pl/bdlen`.

[10] See `http://ec.europa.eu/eurostat/data/database`.

[11] Community developed, rOpenGov, `http://ropengov.github.io/about/`.

where $p_{s,j}$ are defined as in (3.39) and ranges from 0 to 1 (the distributions are equal).

Penultimate measure is Hellinger's distance given by (3.42). It is a dissimilarity measure which ranges from 0 (distributions are equal) to 1 (max dissimilarity). It satisfies all the properties of a distance measure ($d_H \in [0, 1]$; symmetry and triangle inequality).

$$d_H = \sqrt{1 - B}, \tag{3.42}$$

where B is defined by (3.41). It should be noted that relation between (3.42) and (3.41) can be expressed as $d_H^2 \leq D \leq d_H \times \sqrt{2}$. Finally, the last measure is based on Pearson's χ^2 and is given by equation (dist-comp-chi2)

$$\chi_P^2 = n_1 \times \sum_j \frac{(p_{1,j} - p_{2,j})^2}{p_{2,j}}, \tag{3.43}$$

where n_1 is sample size of first data source and $p_{1,j}, p_{2,j}$ is defined the same as previously. χ^2 value can be used to test the hypothesis that two distributions are equal ($df = J - 1$), where J denotes number of levels of given variable. To determine acceptance of the null hypothesis (equality of distributions) in the case of alpha=0.05 $\chi_{P/d}^2 \leq \chi_{J-1,\alpha=0.05}^2$ should be checked. D'Orazio (2015) and D'Orazio et al. (2006) noted that using this measure for two complex surveys might be not straightforward and reuire calculation of generalized design effect for both surveys. However, in case of IDSs where no sampling is done, this measure is acceptable.

Measures presented above are applicable to separately for domains and time period. However, when assessment of representativeness an important factor is coherence estimates based on two data sources in time. In such setting, the following steps should be considered to assess representativeness:

1. Visual comparison of point estimates for target variable on time series plot – this inspection allows for compare distributions of target variable in IDS and reference data, determination of bias or different behaviour in time.

2. Comparison of estimated trends and (if applicable) seasonal component – reference data are often based on sample surveys. To estimate trend LOWESS (locally weighted scatterplot smoothing) algorithm proposed by Cleveland (1979) and further developed by Cleveland et al. (1990) could be used. Name LOESS (local regression) is also used for this method. In this approach the following model is assumed

$$y_i = f(x_i) + \epsilon_i, \tag{3.44}$$

where $f\dot{()}$ is unknown function and $\epsilon_i \sim N(0, \sigma^2)$. Method do not take any assumption about $f\dot{()}$ and should be approximated using polynomial.

Statistical packages, such as R, use weighted local regression to obtain estimates of $f()$ minimizing

$$\sum_{i=1}^{n} w_i^j (y_i - \beta_0^j - \beta_1^j x_i), \tag{3.45}$$

where w_i is defined by

$$W(x) = \begin{cases} (1 - |x|^3)^3 & \text{if } |x| < 1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.46}$$

β_0^j and β_1^j are the intercept and slope of the linear relation between x and y in the neighbourhood of x_j (Hollander et al., 2013, p. 664-665). In case when time series is characterised with higher frequency methods such as X12-ARIMA, TRAMO-SEATS or X13-ARIMA-SEATS to estimate trend and seasonal component could be considered.

3. Co-integration tests could be applied in order to verify whether two or more data sources are behaving in the same way; Moreover, time series clustering could be also applied in order to detect similarities between IDSs and official statistics.

Measures that are presented in this section answer different questions regarding representativeness. Methods based on individual data might provide information about selection mechanism.

3.5 Conclusions

The above chapter has dealt with the theoretical basis of Internet data sources, focusing on the concept of representativeness and its measurement. The basic notation has been introduced along with selected definitions from the literature have been presented and discussed in the context of IDSs. In particular the chapter has examined the concept of representativeness as presented in the literature and the way it can be applied to IDSs. Based on these theoretical consideration a two-step procedure to measure representativeness has been proposed. The method is a general approach to assess the quality of IDSs. It takes into account different data sources (e.g. surveys, register data) and aggregation levels (unit and domain). The procedure is not limited to Internet data sources but could easily be applied to other types of new data sources. Moreover, the approach has been illustrated with practical examples and is not limited to the real estate market. The next chapter of the dissertation will be devoted to the application of the two-step procedure to assess IDSs about the real estate market in Poland.

Chapter 4

Empirical assessment of the representativeness of Internet data sources

This chapter presents an empirical assessment of the representativeness of IDSs for the secondary real estate market in 12 selected cities according to the procedure proposed in Chapter 2. The first step involves an overall assessment of the extent to which the target population is represented in the Internet. Because there is no survey to determine the number of properties published online, information about companies operating in the real estate market will be discussed in detail. The second step is a comparison of the aggregated IDS data with the Register of Transactions and the NBP/CSO survey of brokers. Due to the availability of register data, the comparison will be limited to the city of Poznań.

4.1 Internet access in real estate market enterprises in Poland

The first step in measuring the representativeness of IDSs is to answer the question whether the Internet is useful as a source of statistics. In the case of the secondary real estate market in Poland the answer is not straightforward. There are no official or non-official data on how many properties are presented online, nor is there a list of all properties offered for sale in the secondary real estate market. In this situation one can only rely on information about brokers. Due to the lack of access to the REGON register, the author decided to use the Information and Communication Technologies (ICT) survey to obtain information about the use of the Internet in real estate market companies.

In 2002 the European Commission established annual Information Society surveys benchmarking the ICT-driven development in enterprises and by individuals. Eurostat coordinates two surveys which are carried out at the national level, one on "ICT usage and e-Commerce in enterprises" and one on "ICT usage in households and by individuals". The surveys are developed in close collaboration with Member States and OECD and are adapted to the changing needs of users and policy makers. These surveys are based on Model Questionnaires and

the accompanying methodological guidelines for their implementation. The methodology applied in the ICT survey ensures harmonized data for all EU-28 Member States. The target population for the enterprise ICT survey are companies with over 10 employees[1].

Table 4.1 contains information on Internet access in enterprises that were classified into the Real Estate Activities (REA) group based on the ICT survey. Data refer to the period from 2010 and 2015 and were limited to countries for which information for all selected years was available. Internet access in REA enterprises varied between 58% and 100%. In nearly all countries Internet access was over 90% between 2010 and 2015. However, in Romania and the United Kingdom reported Internet access was lower. The overall level of access in the European Union (28 countries) is stable and on average equals 94.5%. In the case of Poland the number of REA enterprises is higher than the EU average and was equal to 100% in 2010 and 97% in 2015.

TABLE 4.1: Enterprises (Real estate activities, with 10 or more employees) with Internet access between 2010-2015 in selected countries [%]

Country	2010	2011	2012	2013	2014	2015
Austria	93	93	99	100	98	100
Bulgaria	85	90	91	96	96	100
Cyprus	92	95	94	100	100	93
Czech Republic	95	94	98	98	97	98
Estonia	100	93	95	94	100	97
France	100	98	100	100	100	100
Germany	88	95	100	93	96	100
Hungary	90	91	95	88	84	90
Ireland	94	92	98	96	95	97
Italy	95	93	96	94	97	94
Latvia	91	94	94	95	97	98
Lithuania	100	100	99	99	100	100
Netherlands	99	100	100	100	100	100
Poland	100	97	99	99	96	97
Portugal	99	100	100	100	100	100
Romania	88	62	58	97	85	94
Slovakia	100	100	100	100	95	98
Slovenia	100	100	100	100	100	100
Spain	98	98	96	100	98	98
Sweden	100	100	100	97	97	99
the UK	88	88	86	90	81	92
European Union (28 countries)	94	93	95	95	93	97

Note: based on Eurostat ICT survey data (indicator code: isoc_ci_in_en2). Data were downloaded from Eurostat using **eurostat** (Lahti et al., 2014) package. The R code for the table above is available in Appendix A.4.1.

[1]The methodology for the ICT surveys can be found in the Methodological Manual, see `http://ec.europa.eu/eurostat/web/information-society/methodology`.

However, overall Internet access presented in Table 4.1 can be misleading for the following reasons. The ICT survey only covers enterprises with over 10 employees, with the exclusion of micro enterprises (with fewer than 10 employees) or self-employed persons. In addition, the category of Real Estate Activities (NACE GROUP L) includes several groups of enterprises that are not directly connected with the process of buying or selling real estate. The REA group includes the following enterprises: (1) Buying and selling of own real estate, (2) Rental and operating of own or leased real estate and (3) Real estate activities on a fee or contract basis[2]. Enterprises within the scope of the present study belong to the third group and according to NACE rev.2 are assigned code 68.31. Nonetheless, the information in Table 4.1 indicates the level of Internet access and usage by real estate activity companies.

More detailed information about Internet use can be found in annual reports prepared by each Member State. Table 4.2 contains information about web pages and Internet use by enterprises in Poland, particularly in companies classified as REA. Between 2010 and 2015 a stable fraction of 65 % of enterprises in Poland had their own websites. In the case of REA enterprises, the share is higher and ranges from 67% to 73%. In other words, the average percentage of REA enterprises with websites is higher than the corresponding average for the general enterprise population in Poland.

TABLE 4.2: Ownership and use of web pages by enterprises in Poland between 2010–2015 [%]

Specification	2010	2011	2012	2013	2014	2015
Enterprises owning a website or homepage as percentage of all enterprises in the group						
Poland	65.5	64.7	67.6	66.0	65.3	65.4
Real estate activities	67.0	63.3	70.9	74.9	73.9	72.9
Product catalogues or price lists as percentage of all enterprises in the group						
Poland	48.8	46.9	51.4	51.5	60.4	60.3
Real estate activities	26.8	25.0	32.8	35.7	51.4	55.6

Note: based on data from the Polish ICT survey conducted by the Central Statistical Office.

In addition, there is a rapid increase in the use of websites for presenting products in REA enterprises from 26.8% in 2010 to 55.6% in 2015. Growth can also be observed for all enterprises in Poland, however it is not as high as in the REA category and was only 10 percentage points between 2010 and 2015. The actual question in the ICT survey concerning the web page was formulated as follows: Did your enterprise have a website in January YYYY?[3] The ICT Survey question

[2]See http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2&StrLanguageCode=EN&IntPcKey=18516584&StrLayoutCode=.

[3]Where YYYY denotes a year. The Polish version of the question was *Czy w styczniu YYYY r. przedsiębiorstwo posiadało własną stronę internetową (WWW)?*

concerning components of enterprise websites was: Does your company website have the following functions: product catalogues or price lists?[4].

In this case one faces similar problems as before (different kinds of enterprises, lack of micro enterprises); however, in the ICT survey the question concerns only the ownership of a web page, not the use of advertisement services. Such services do not require the creation of a seperate website for a real estate broker/agency in order to present properties for sale. In addition, brokers use special software (for example software in Poland: Galactica Virgo, Agencja5000 or MLS RealNet), which is mainly based on XML formats, to communicate with multiple advertisement services and do not need to maintain a dedicated web page.

Despite the above drawbacks of the surveys on Internet access in RAE companies, it should be underlined that the Internet is an important data source for real estate market statistics. Hence, the answer to the question "Is the Internet useful for providing statistics?" the answer is *yes*. This leads to the second step of the measurement of representativeness, namely empirical assessment of available IDS data.

4.2 Selected reference data sources about the real estate market

Two reference sources were used in the empirical study: a survey conducted by NBP/CSO and the Register of Transactions. The first data source was available at the city level (domain) and on quarterly basis. Data obtained from the NBP/CSO survey concern 16 cities – Białystok, Bydgoszcz, Gdańsk, Katowice, Kielce, Kraków, Lublin, Łódź, Olsztyn, Opole, Poznań, Rzeszów, Szczecin, Warszawa, Wrocław and Zielona Góra[5]. The NBP/CSO survey includes a number of characteristics of the secondary real estate market, including the average price per m^2, fractions of flats by the number of rooms (divided into four groups: 1, 2, 3 and $4+$) and floor area (categorized into four groups: below 40 m^2, $(40, 60]$ m^2, $(60, 80]$ m^2 and $80+$ m^2)[6]. Table 4.3 presents the distribution of the NBP/CSO survey quarterly sample size in the secondary real estate market between the 1st quarter of 2012 to the 4th quarter of 2014[7]. The table contains information about flats offered for sale (offers) and sold (transactions). As we can see, in the case of Warszawa, Gdańsk, Olsztyn and Białystok the sample size is the highest (relative to the average), while for Opole, Bydgoszcz, Katowice, Kielce and Zielona Góra it is the lowest. Interestingly, in the case of Opole the sample size for transactions is almost five times as high as that for offers; however, there is no information in the NBP/CSO survey to accont for this difference.

[4]The Polish version of the question was: *Czy w styczniu 2011 r. strona internetowa przedsiębiorstwa spełniała następujące funkcje: prezentacja katalogów wyrobów lub cenników*.

[5]The NBP/CSO survey also collects data about Gdynia. However, no data about the secondary real estate market in Gdynia was available. Please refer to Figure A.3, which presents locations of cities covered by the NBP/CSO survey.

[6]For more information please refer to NBP reports, see `http://www.nbp.pl/homen.aspx?f=/en/publikacje/inne/real_estate_market_a.html`.

[7]Notation 2012Q1, 2012Q2, ..., 2014Q4 will be used hereafter to denote years and quarters.

TABLE 4.3: Sample size between 1st quarter of 2012 and 4th quarter of 2014 in 16 cities

City	Type	Min	Q_1	Average	Median	Q_3	Max
Białystok	Offers	1031	1071	1109	1112	1141	1193
Białystok	Transactions	79	100	185	195	240	294
Bydgoszcz	Offers	50	62	97	76	88	315
Bydgoszcz	Transactions	109	132	205	152	264	437
Gdańsk	Offers	1769	1919	2306	2256	2566	3213
Gdańsk	Transactions	82	117	204	188	288	366
Katowice	Offers	117	152	271	173	289	693
Katowice	Transactions	97	130	215	224	298	336
Kielce	Offers	140	267	295	322	342	375
Kielce	Transactions	102	118	129	128	132	164
Kraków	Offers	527	592	736	733	777	1332
Kraków	Transactions	178	378	454	441	528	771
Łódź	Offers	130	155	975	1274	1429	1633
Łódź	Transactions	22	35	86	106	117	128
Lublin	Offers	500	600	1109	1161	1382	1970
Lublin	Transactions	52	101	212	145	330	437
Olsztyn	Offers	728	1138	1319	1397	1442	1692
Olsztyn	Transactions	247	272	304	293	334	396
Opole	Offers	25	27	34	34	38	49
Opole	Transactions	74	130	156	164	192	218
Poznań	Offers	221	355	993	532	1565	2383
Poznań	Transactions	195	326	403	369	454	740
Rzeszów	Offers	230	434	456	494	524	601
Rzeszów	Transactions	70	83	102	101	110	144
Szczecin	Offers	305	318	378	382	417	510
Szczecin	Transactions	159	223	284	296	326	418
Warszawa	Offers	2955	5001	5472	5535	6396	6896
Warszawa	Transactions	261	371	1179	1180	1885	2506
Wrocław	Offers	252	281	322	330	349	401
Wrocław	Transactions	108	204	289	292	381	467
Zielona Góra	Offers	177	252	273	272	305	351
Zielona Góra	Transactions	46	73	89	80	107	134

Note: own elaboration based on NBP/CSO survey data.

In the case of the Register of Transactions only data for Poznań was available during the work on the dissertation. The data were obtained from The Board of Geodesy and Municipal Cadastre GEOPOZ[8] based on new regulations introduced in 2014 in The Geodetic and Cartographic Act (1989). Since 01.01.2014 these data can be obtained from the Register of Transactions free of charge for research and educational purposes.

[8]pol. *Zarząd Geodezji i Katastru Miejskiego GEOPOZ*, for more information see `http://www.geopoz.pl/portal/index.php?t=200&id=1236`.

The data were delivered by GEOPOZ in a MS Excel file and contained information about transactions involving apartments between 2008 and 2014. However, the data were not prepared by statisticians or for statistical purposes and their quality needed to be verified. The Register lists transactions that took place in the primary and the secondary market. Hence, to ensure consistency with the target population (the secondary market, residential properties) the register was cleaned in the following steps:

- Only transactions involving residential properties were taken into account. Garages were excluded.
- Transactions (objects) were transformed into units (residential properties). Each row contained the transaction number and the property ID that was the subject of the transaction. The property ID was taken from the Mortgage Register and is unique for each property. ID contains a 4-character code assigned by the Department of Land Registry of the District Court in a given city, 8 digits from the mortgage register assigned to a given property and a control digit, for instance PO1P/XXXXXXXX/1. After transformation, the dataset consisted of:
  1. Connections objects-to-units: one-to-one: 11 257, one-to-many: 2 081. If we take into account the unit-error theory proposed by Zhang (2011), there were 10 types of blocks within the allocation matrix (objects to statistical units).
  2. The number of times units were sold during the period: only once – 11 760, twice – 70, three times – 48, four times – 5. During this period most residential properties were sold only once, only 760 residential properties were sold multiple times (5.7% of all residential properties).
- Market type (primary, secondary) was corrected. For instance, apartments that were sold in the secondary market appeared in the primary market.
- Records with missing data in floor area (# missing = 9), number of rooms (# missing = 279) and total price (# missing = 424) were removed from the comparison. As a result, 703 records were removed and the final dataset contained 13 338 rows (residential properties).

Table 4.4 presents descriptive statistics for total price (in PLN), floor area (in m^2) and the number of rooms of residential properties sold between 2012Q1 and 2014Q4. The smallest residential property was 12.80 m^2 and the biggest over 226 m^2 and the average floor area was equal to 52.45 m^2. The lowest total price was 1500 PLN[9] and the highest over PLN 3.1 mln and on average an apartment was valued at 253 886.65 PLN (about 60k EUR). There are several outliers in the cleaned dataset: for instance the cheapest flat was 53.3 m^2 and had 3 rooms, or there were flats with one room and over 100 m^2 of floor area.

Figure 4.1 presents the distribution of sample size for Poznań in the two reference data sources (NBP/CSO survey, Register) and the three selected IDSs. For

[9]The average EUR-to-PLN exchange rate stood at 4.3 PLN at the end of 2015.

TABLE 4.4: Descriptive statistics of selected characteristics from the Register of Real Estate Prices and Values between 2012Q1 and 2014Q4

Statistic	Total Price	Floor Area	Number of Rooms
Min	1 500.00	12.80	1.00
Percentile 0.1%	28 153.02	17.60	1.00
Q_1	185 000.00	37.80	2.00
Median	235 000.00	48.49	3.00
Mean	253 886.65	52.45	3.00
Q_3	297 000.00	61.80	4.00
Percentile 99.9%	1 259 930.00	172.77	6.99
Max	3 175 000.00	226.86	9.00

Note: Number of records = 13 338.

comparison, the data were log-transformed (using the common logarithm). Table 4.5 presents (untransformed) basic statistics for sample size for all the data sources. The differences between the data sources are substantial: in the case of the register on average 519 transactions of residential properties were recorded, which is close to the number of residential properties surveyed by NBP/CSO. Moreover, the NBP/CSO survey saw a decrease in sample size over time, from over 2000 in 2012Q1 to nearly 500 in 2014Q4.

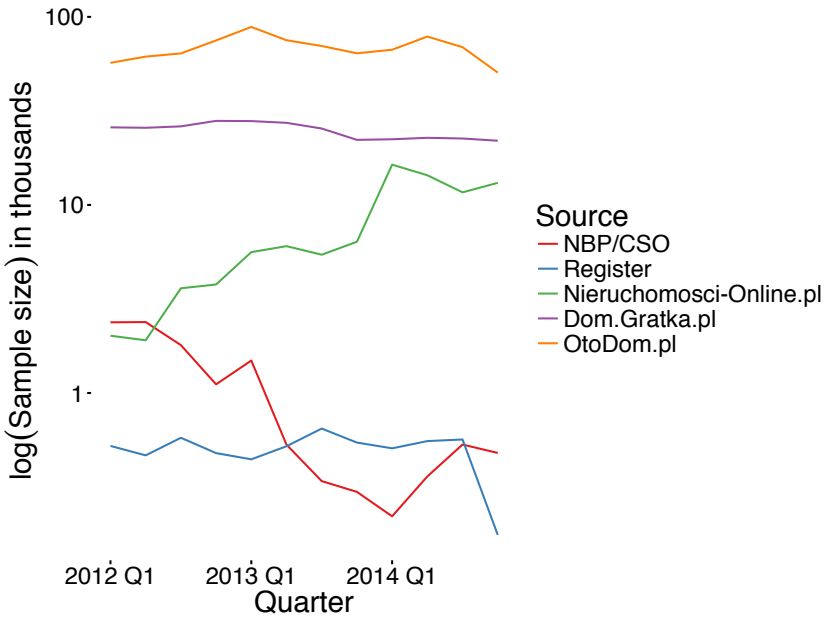


FIGURE 4.1: Sample size in NBP/CSO survey, Register of Transactions, Dom.Gratka.pl, Nieruchomosci-Online.pl and OtoDom.pl between 2012Q1 and 2014Q4 in Poznań

The decline in the number of transactions in the last quarter of 2014 in the register is connected with a delay in reporting transactions. Notaries have one month to inform the local government about transactions, and data administrators

have one month to enter the information in the register. Hence, the maximum delay between a transaction and its appearance in the register can be up to three months. As a result, complete information about all transactions in 2014Q4 will be available at the end of 2015Q1. However, the number of transactions in the secondary residential market in Poznań was stable over time.

The biggest differences in sample size can be observed between the IDSs. In the case of Dom.Gratka.pl and OtoDom.pl the average number of residential properties is stable between 2012Q1 and 2014Q4 and is equal to 24 850 and 68 310 respectively. In the case of Nieruchomosci-online.pl the sample size increases over the reference period to reach 13 104 in 2014Q4. Nonetheless, these numbers can be misleading for two reasons: (1) the presence of duplicates in these sources; (2) the data collection process.

The first problem is strictly connected with the character of the data sources – advertisement services – and the market organization. As noted earlier, multiple advertisements can refer to the same residential property: on the one hand, this is the result of how individual brokers operate in the market; on the other hand, it is the consequence of open agreements. Open agreements denote arrangements whereby multiple brokers promote the same property. The second issue has to do with the data collection process. Dom.Gratka.pl and OtoDom.pl provided aggregated data, which were not de-duplicated before delivery, thus the number of observations may be severely biased. However, in both data sources a decrease over time is observed.

TABLE 4.5: Descriptive statistics of sample size in the NBP/CSO survey, the Register of Transactions, Dom.Gratka.pl, Nieruchomosci-Online.pl and OtoDom.pl between 2012Q1 and 2014Q4 in Poznań

Source	NBP/CSO	Register	Nieruchomosci-Online.pl	Dom.Gratka.pl	OtoDom.pl
Min	221	184	1907	21940	50550
Q_1	355	492	3733	22500	63310
Mean	993	519	7523	24850	68310
Median	532	549	5824	25600	67930
Q_3	1565	576	12030	26460	74980
Max	2383	670	16370	27980	88380

In the case of Nieruchomosci-Online.pl object-level data were web-scraped in 2014. The dataset contained current and historical data with information about the time when a given object (advertisement) was placed and updated. The updated date was used to identify the quarter. Therefore, the number advertisements placed in 2012-2013 may be underrepresented.

4.3 Selected Internet data sources in the secondary real estate market

Three data sources were used in the analysis - Nieruchomosci-Online.pl, Dom.-Gratka.pl and OtoDom.pl. For Nieruchomosci-Online.pl a special script was developed to scrape current and archived advertisements. As a result, an object-level

dataset was obtained, which was cleaned and transformed to unit-level data. In the case of Dom.Gratka.pl and OtoDom.pl aggregated data were made available by the owners of these web portals.

4.3.1 Nieruchomosci-online.pl

To obtain data from Nieruchomosci-Online.pl an R and Python program was developed to scrape information directly from Nieruchomosci-Online.pl web pages. **XML** (Lang, 2013), **RCurl** (Lang, 2014) and **httr** (Wickham, 2015) packages were used for this purpose. Algorithm 1 presents the pseudo-code for web-scraping.

Data: Web pages, N - the number of search result pages (i), n - the number of results on the search page (j)
Result: A text file with scraped data
Send a query through the form on the web page and save the link to results;
Set cookies for the 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 the search result page and write it to a text file;
        Enter j page from the search result page;
        Scrape all text data from j page and write it to a text file;
    end
end

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

In total, 685 670 records containing individual data were obtained. This dataset contains information not only about Poznań but also about several other cities. In total data on 12 cities were scraped. Table 4.6 shows the number of records (properties offered for sale) stored on Nieruchomosci-online.pl. This is compared with information about the total number of transactions in the real estate market and the number of residents in 2014 for each city. The number of offers is correlated with the number of citizens and the number of transactions, the highest number is observed for Warszawa and the lowest for Olsztyn. However, there is one exception, in 2014 Łódź ranked 3rd in terms of population, but only 6th in terms of the number of residential properties offered for sale.

The scraped data were not ideal, variables contained multiple missing values and inconsistencies. Table 4.7 contains information about the percentage of missing data for 12 selected variables[10]. The majority of missing data are in variables connected with coordinates (exact location). There are two reasons for it: firstly,

[10]Data obtained from Nieruchomosci-online.pl contained 25 variables: city name, price currency, market type, floor number, maximum number of floors, number of rooms, last updated,

TABLE 4.6: The number of observations scraped from Nieruchomosci-Online.pl, the number of transactions and the number of residents in 12 cities between 2012Q1 and 2014Q4

City	# Records	# Records [%]	# Transactions	# Residents
Warszawa	225 332	32.9	13 363	1 735 442
Kraków	150 477	21.9	10 687	761 873
Poznań	94 466	13.8	3 874	545 680
Wrocław	60 488	8.8	3 668	634 487
Gdańsk	46 787	6.8	5 792	461 489
Łódź	25 387	3.7	5 068	706 004
Szczecin	23 593	3.4	1 435	407 180
Białystok	15 819	2.3	1 906	295 459
Katowice	15 159	2.2	413	301 834
Lublin	14 025	2.0	1 329	341 722
Opole	8 361	1.2	795	119 574
Olsztyn	5 776	0.8	1 841	173 831

Note: Symbol # denotes "number of".

brokers/owners may not want to reveal the exact location of a flat for sale, secondly, it is the result of the data collection process. Google Maps limit the number of queries to 2 500, which posed the main problem during web-scraping. The remaining variables related to bathroom, kitchen, market type or who placed the advertisement. Another reason for the occurrence of missing values is the fact that the relevant information is placed in the longer text description and not included in the appropriate fields. Interestingly, there is almost a complete lack of missing data for the number of rooms, the floor area and the total price. However, even if data is not missing, the variables are not free from errors (e.g. measurement error). This problem will be discussed later in this section.

The percentage of missing data was compared across cities in order to detect any missing data patterns. Two types of missing data mechanisms (MAR or MNAR) should be considered. Table 4.8 presents information about missing data categorized by cities. The fraction of missing data in geographical coordinates is similar across the cities; the biggest differences are visible in other variables. For instance, the fraction of missing data about the person that placed an ad varies from 25% in Opole (characterised by the smallest number of transactions) to 89% in Lublin. For the variable defining the kitchen type the percentage of missing values ranges from 18% (in Szczecin) to 74 %; for the year of construction missing values account for from 12% to 92% of all observations.

Table 4.8 reveals a correlation between variable *Who placed advertisement* (column Who) and *overall average offer price per* m^2 (column Price m^2). The Pearson correlation coefficient is equal to 0.59 and Spearman's rho is equal to 0.64.

longitude, latitude, city district, year of construction, kitchen type, bathroom type, number of levels, who placed the ad, building type, total price, price per m^2, floor area, usable floor area, street name, link to offer, NUTS2 level, LAU1 level and LAU2 level.

TABLE 4.7: Missing data for selected variables from Nieruchomosci-Online.pl between 2012Q1 and 2014Q4

Variable	Percentage of missing data [%]
Longitude	92.43
Latitude	92.43
Who placed ad	75.15
Bathroom	74.74
Kitchen	60.31
Market	49.40
Year of build	33.36
No. of Floors	3.94
Floor Number	3.25
No. of Rooms	0.68
Floor Area	0.00
Total Price	0.00

Note: N = 685 670.

The higher the average offer price is, the more likely it is for an ad to lack information about who placed it (legal person - broker, or natural person). For example, 25% of offers in Opole and 43% in Katowice lacked this information (cities with the lowest overall average price), compared with 85% in Warszawa and 95% in Wrocław (cities with the highest overall average price). This may be connected with brokers' fee that is not as high as in cities where the price is two or three times higher.

TABLE 4.8: A comparison between the average offer price per m^2 and the percentage of missing values in selected variables from Nieruchomosci-Online.pl for 12 cities between 2012Q1 and 2014Q4

City	Price m^2	Who	Bath	Kitchen	Market	Year	No. Fl	Fl	Rooms	Area	Price
Warszawa	8200	87	75	55	54	12	2	1	0	0	0
Kraków	6789	71	68	65	53	42	7	5	1	0	0
Wrocław	5714	68	71	65	44	46	4	3	1	0	0
Gdańsk	5697	70	84	53	8	19	3	0	1	0	0
Poznań	5270	69	94	78	72	40	4	6	1	0	0
Lublin	4737	89	63	54	22	37	3	2	0	0	0
Szczecin	4154	81	37	18	24	43	4	3	1	0	0
Białystok	4051	87	44	40	40	44	6	3	0	0	0
Olsztyn	3908	68	78	71	40	61	5	4	1	0	0
Łódź	3863	57	81	62	41	72	6	7	1	0	0
Katowice	3734	43	87	74	51	76	3	7	1	0	0
Opole	2289	25	86	63	60	92	5	2	0	0	0

Column names: Price m^2 – overall average offer price per m^2 (for the whole period, computed as $\sum price_i / \sum floorarea_i$), Who – Who placed advertisement, Year – year of build, No. Fl – Number of floors, Fl – floor, Rooms – number of rooms, Area – floor area, Price – total price.

Note: N = 685 670.

Table 4.9 contains basic descriptive statistics for selected variables. Descriptive statistics were also calculated for the number of floors, the number of rooms, the year of construction in order to present the range of values observed for these variables. In addition to the number of missing values (last column of Table 4.9), the dataset also contains erroneous values. For instance, the minimum number of floors is -3, which is not a correct value for this variable; the maximum number of floors is 65 535, which is also erroneous. Geographical coordinates also contain several values outside the interval between 14 and 24 degrees E (longitude) and from 49 to 55 degrees N (latitude), which roughly covers the area of Poland. The minimum value for the year of construction is 1 and the maximum is 32767, obvious errors in these variables. The year of construction should be expressed as a four digit number. The most important variables for the analysis (floor area and total price) also contain invalid values and they had to be corrected in the first place. Therefore, for the purpose of data correction the following criteria were applied: (1) the total price should lie between 44 000 and 3 300 000 (0.1% and 99.9% percentile); (2) the number of rooms should be lower than 20; (3) the number of floors must be equal to or greater than 0 (0 = ground level). As a result, 647 969 observations were obtained, accounting for 94.5% of the initial number of records.

TABLE 4.9: Basic descriptive statistics for selected variables in Nieruchomosci-Online.pl for 12 cities between 2012Q1 and 2014Q4

Variables	Min	Q_1	Mean	Median	Q_3	Max
Floor area	1	44	61.97	55	70	197000
Total price	1	245 000	399 197.28	329 000	453 000	40 0000 000
Floor number	-3	1	2.88	2	4	20
# Floors	1	3	5.31	4	7	20
# Rooms	1	2	2.64	2	3	65535
Lat	0.00	51.12	51.89	52.21	52.28	54.46
Long	-83.50	18.99	19.90	20.92	21.03	23.56
# Bathrooms	1	1	1.26	1	1	270
Year of build	1	1965	1982	1988	2004	32767

Another step of data cleaning was the transformation of objects into units based on the following variables (excluding the total price, which may differ over the reference period): City name, Street name, Floor number, Number of floors, Number of rooms, Year of construction, Bathroom, Kitchen, Number of levels, Floor area. After transformation the effective number of records was reduced to 347 797 from the initial 685 617 (50% of the initial number of records). The number of unique statistical units offered for sale between 2012Q1 and 2014Q4 was 294 360. Figure 4.2 shows the number of units observed on Nieruchomosci-Online.pl between 2012Q1 and 2014Q4. To improve readability, square root transformation was applied. The order of city name labels reflects the number of observations in the last quarter. However, it should be remembered that properties were assigned to specific quarters on the basis of the date when the ads were last updated. Another problem is that the data were collected in 2014, which can explain the peaks in the counts. This is especially visible for Warszawa, where

the number of observations skyrocketted after 2014Q3 to over 33000, the highest number of properties offered in the secondary market, while in other cities the increase was much smaller. Three group of cities can be distinguished: (1) Warszawa, (2) Kraków, Poznań, Gdańsk and Wrocław and (3) Szczecin, Łódź, Białystok, Lublin, Katowice, Opole and Olsztyn. These groups are not only homogenous with respect to the number of observations but also in terms of the average offer price presented online.

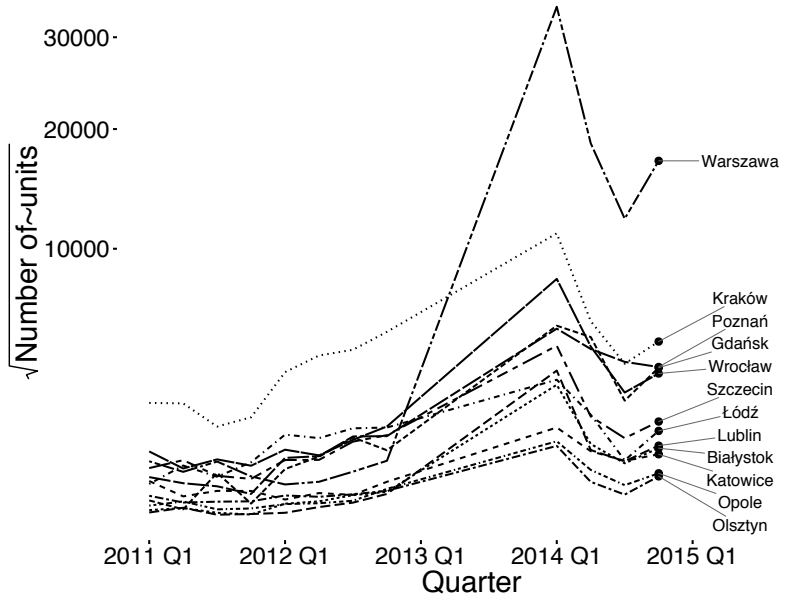


FIGURE 4.2: Number of units observed at Nieruchomosci-Online.pl between 2012Q1 and 2014Q4

4.3.2 Dom.Gratka.pl

Data from Dom.Gratka.pl were obtained based on a collaboration established between the Department of Statistics, Poznań University of Economics and Business and Polska Presse Group (the owner of Dom.Gratka.pl). This co-operation opened access to aggregated data but only for domains specified by the owner. The dataset prepared by the owner contained monthly historical data on the average total price, average price per m^2 and the floor area for a period from January 2012 to March 2015 and the following domains: NUTS2 level, city, district, floor area, number of rooms and year of construction. The floor area was classified into four groups: $[10, 35]$, $(35, 50]$, $(50, 70]$, over 70. The number of rooms was classified into 1, 2, 3 and $4+$ and the year of construction was converted into a two-valued variable, for properties built between 1900 and 2000, and after 2000.

Polska Presse Group prepared 10 datasets, which are presented in the list below. For each dataset there is information about available domains, the number of records and the total number of advertisements. Each data set contains a prespecified combination of variables, for instance the first dataset contains data aggregated for year, month and NUTS2 level[11] (Year $\times$ Month $\times$ NUTS2 level).

[11] In Poland there are 16 NUTS2 level units.

The number of records (domains created by Year × Month × NUTS2 level) is equal to 621, while the total number of advertisements in all domains was 16 580 958. Each dataset contains information about the overall average offer price, the average mean price per m^2 and the number of records for each domain.

Despite the aggregated character of the data, the total number of advertisements for each dataset can be used to estimate the number of missing data. For instance, a total of 15 600 757 ads contained the city name (Dataset 2), while only 13 067 677 ads contained the district name (83.76%, Dataset 3). Furthermore, only 10 268 263 (65.82%, Dataset 4) ads included information about the floor area, the number of rooms and the year of construction, while 9 227 479 (59.15%, Dataset 8) ads contained information about the district, the number of rooms and the year of construction. The difference in the total number of advertisements indicates that there is a substantial amount of missing values for the district name, while for the number of rooms and the floor area missing values may be at the same level as in data from Nieruchomosci-Online.pl.

1. Dataset 1:
    - Domains: Year × Month × NUTS2 level,
    - Number of records (domains): 621,
    - Total number of advertisements = 16 580 958,
2. Dataset 2:
    - Domains: Year × Month × NUTS2 level × City,
    - Number of records (domains): 18 590,
    - Total number of advertisements = 15 600 757,
3. Dataset 3:
    - Domains: Year × Month × NUTS2 level × City × district,
    - Number of records (domains): 24 324,
    - Total number of advertisements = 13 067 677,
4. Dataset 4:
    - Domains: Year × Month × NUTS2 level × Floor area × Build year,
    - Number of records (domains): 4 968,
    - Total number of advertisements = 10 930 734,
5. Dataset 5:
    - Domains: Year × Month × NUTS2 level × Number of rooms × Build year,
    - Number of records (domains): 4 968,
    - Total number of advertisements = 10 514 769,
6. Dataset 6:

- Domains: Year × Month × NUTS2 level × Floor area × Number of rooms × Build year,
- Number of records (domains): 19 872,
- Total number of advertisements = 10 874 412,

7. Dataset 7:

- Domains: Year × Month × NUTS2 level × City × Floor area × Number of rooms × Build year,
- Number of records (domains): 594 880,
- Total number of advertisements = 10 268 263,

8. Dataset 8:

- Domains: Year × Month × NUTS2 level × City × District × Number of rooms × Build year,
- Number of records (domains): 194 592,
- Total number of advertisements = 9 227 479,

9. Dataset 9:

- Domains: Year × Month × NUTS2 level × City × Number of rooms × Build year,
- Number of records (domains): 148 720,
- Total number of advertisements = 9 926 897,

10. Dataset 10:

- Domains: Year × Month × NUTS2 level × City × Floor area × Build year,
- Number of records (domains): 148 720,
- Total number of advertisements = 10 320 764.

Table 4.10 lists 10 sample rows from dataset 7, which was used later in the study. This dataset contains information about the joint distribution of the floor area, the number of rooms, the average total offer price and the average mean offer price per m^2. Initially, the set contained information about 370 cities; however, for comparative purposes it was limited to 12 cities (Białystok, Gdańsk, Katowice, Kraków, Łódź, Lublin, Olsztyn, Opole, Poznań, Szczecin, Warszawa, Wrocław).

In the selected dataset the mimimum average price per m^2 was equal to 0, the maximum was 39640 and 64 rows contained missing data. The zero and NA values were due to the lack of observations in the domains provided in dataset 7. The following restriction concerning the average total price was applied to dataset 7: prices with zero or missing values were excluded; in addition, as was done with data from Nieruchomosci-Online.pl, rows ranging from 0.1th (3 501) to 99.9th (11 433) percentile of the total offer price were used for further analysis. Thus, as a result, a total of 2 778 rows were excluded from the analysis.

TABLE 4.10: 10 sample rows from the selected dataset containing information on Dom.Gratka.pl

City	Floor Area m^2	# Rooms	Mean Price m^2	# Records	Quarter
Wrocław	to 35	1	7632	800	2012Q1
Lublin	to 35	1	5439	155	2012Q1
Łódź	to 35	1	3910	376	2012Q1
Kraków	to 35	1	7682	1960	2012Q1
Warszawa	to 35	1	9428	6503	2012Q1
...	...	...	...	...	...
Gdańsk	70+	4+	6385	705	2014Q4
Katowice	70+	4+	5794	76	2014Q4
Olsztyn	70+	4+	4176	84	2014Q4
Poznań	70+	4+	6143	249	2014Q4
Szczecin	70+	4+	4612	257	2014Q4

Table 4.11 contains the total number of advertisements posted in Dom.Gratka.pl. The city with the biggest number of ads were those for properties located in Warszawa (3 503 369), while the smallest number of ads were posted about properties in Opole (18 654). 46.03% of all advertisements referred to properties offered in Warszawa (compared with 32.9% in Nieruchomosci-Online.pl); however, in comparison with Nieruchomosci-Online.pl 15.54 times more ads were obseved. The relation between the number of advertisements presented on Dom.-Gratka.pl and the number of statistical units identified in Nieruchomosci-Online.pl can be used to estimate the number of advertisements that refer to one property. For instance, in Warszawa it would be 15.6 (= 3 503 369 / 225 332), in Kraków 8.1 (= 1 216 571 / 150 477) or in Poznań 3.4 (= 318 707 / 94 466).

TABLE 4.11: The number of advertisements in Dom.Gratka.pl for 12 cities between 2012Q1 and 2014Q4

City	# Advertisements	# Advertisements [%]
Warszawa	3 503 369	46.03
Kraków	1 216 571	15.98
Wrocław	936 972	12.31
Gdańsk	494 390	6.50
Poznań	318 707	4.19
Szczecin	316 294	4.16
Łódź	272 559	3.58
Lublin	202 090	2.65
Białystok	119 078	1.56
Olsztyn	113 083	1.49
Katowice	100 004	1.31
Opole	18 654	0.25

Figure 4.3 shows the final number of advertisements observed in Dom.Gratka.pl web portal. The order of the cities is almost the same as in Nieruchomosci-Online.pl (Poznań is ranked lower); however, they cannot be grouped as easily

as in the case of Nieruchomosci-Online.pl. In comparison to Nieruchomosci-online.pl, the number of advertisements posted on Dom.Gratka.pl is more stable over time, as are the differences between the cities. The number of properties posted on Nieruchomosci-online.pl increases over the reference period. The biggest difference can be observed for Warszawa: the number of advertisements posted on Dom.Gratka.pl does not increase as it does on Nieruchomosci-Online.pl. On the other hand, there is an interesting pattern for Warszawa and Kraków: a decline in 2013Q2 followed by a rise in 2013Q3 (no such pattern is present for the other cities).

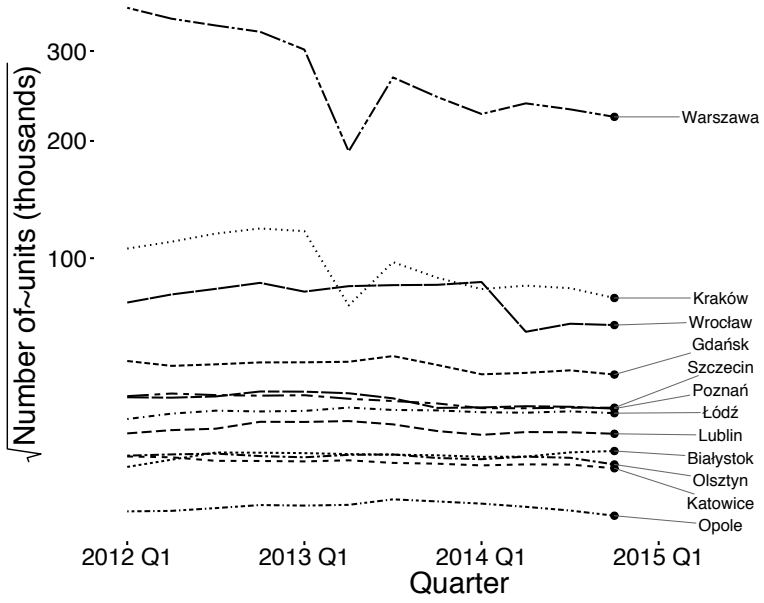


FIGURE 4.3: Number of units observed at Dom.Gratka.pl between 2012Q1 and 2014Q4

Owing to the aggregated character of the data, the variable containing information about the number of rows in each domain was used as a weight for further analysis. Table 4.12 contains unweighted and weighted descriptive statistics for the average price per m^2. The main differences can be observed in the mean and median, where the differences are approximately 1400 and 1800 PLN/m^2 respectively.

Figure 4.4 shows unweighted (solid) and weighted (dashed) density plots of the average price per m^2 on Dom.Gratka.pl. Substantial differences can be seen between these two distributions. First of all, the weighted plot has more peaks (multimodality). This can be the result of counting multiple times properties that contained the same information. In addition, the application of weights has made the distribution less skewed, which implies shrinkage.

4.3.3 OtoDom.pl

Another data source that was made available for research purposes was prepared by Allegro.pl, the owner of the advertising service OtoDom.pl. As in the case of Dom.Gratka.pl, the data was made available in aggregated form; however,

TABLE 4.12: Weighted and unweighted average offer price per m^2 between 2012Q1 and 2014Q4 on Dom.Gratka.pl

Statistic	Unweighted Price m^2	Weighted Price m^2	Weight
Min	975	975	1
Percentile 5%	3733	4039	16.0
Percentile 10%	3902	4707	20.0
Q_1	4657	5867	40.0
Mean	5761	7152	419.1
Median	5421	7231	120.0
Q_3	6647	8617	396.0
Percentile 90%	7869	9357	978.8
Percentile 95%	9003	9541	1783.0
Max	39 645	39 645	16812

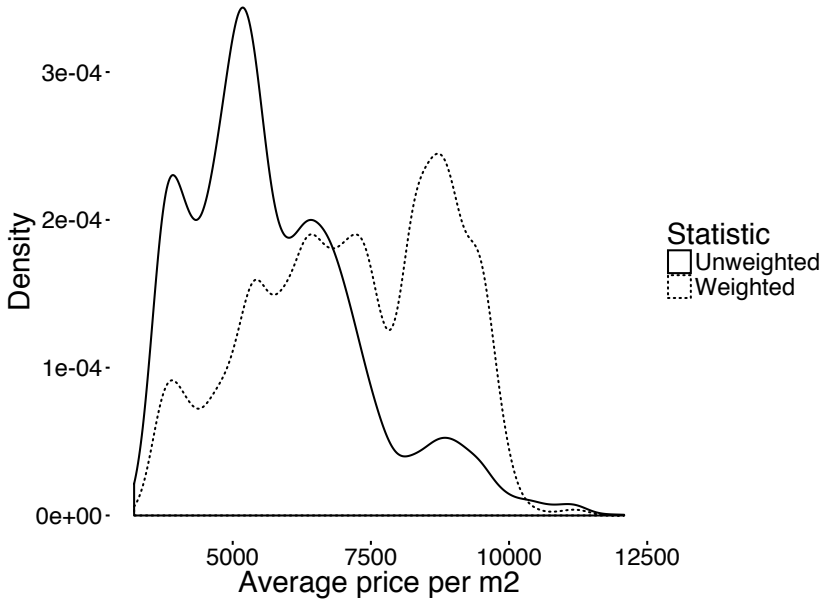


FIGURE 4.4: Unweighted and weighted density plot of average price per m2 between 2012Q1 and 2014Q4 on Dom.Gratka.pl

the selection of relevant domains (cross-classifications) had been agreed upon with Allegro.pl staff beforehand. The resulting domains are more refined: they are defined as Year $\times$ Month $\times$ City $\times$ Number of rooms $\times$ Floor area. The floor area is categorized into groups with a 10 m^2 interval, the number of rooms is as specified by the person who placed advertisement. The variable Number of rooms has 71 levels and Floor area has 51 levels. For each domain there is information about the average and median price, the average price per m^2, the floor area group, and the number of advertisements. The original dataset contained information about 20 cities but was limited to 12 to ensure comparability. Table 4.13 presents 10 sample rows from the dataset that were made available by Allegro.pl (OtoDom.pl).

Table 4.14 shows the distribution of the number of rooms, the average floor area, the average price per m^2 and the average and median price. OtoDom.pl

TABLE 4.13: 10 sample rows from the dataset provided by OtoDom.pl

City	# Rooms	Floor Area	Mean Price	Mean Price m^2	Quarter	Weight
Poznań	1	20-30	171593.90	6399.27	2012Q1	285
Poznań	1	30-40	220159.39	6370.34	2012Q1	577
Poznań	1	40-50	255231.41	5695.53	2012Q1	56
Poznań	1	50-60	230675.00	4210.30	2012Q1	12
Poznań	2	20-30	194695.16	6809.17	2012Q1	31
...	...	...	...	...	...	...
Poznań	5	80-90	325181.82	3811.61	2014Q2	11
Poznań	5	120-130	514083.33	4056.90	2014Q2	12
Poznań	5	130-140	567141.67	4173.71	2014Q2	24
Poznań	5	140-150	569691.50	3902.30	2014Q2	12
Poznań	5	150-160	631257.14	4029.94	2014Q2	14

TABLE 4.14: Basic descriptive statistics for selected variables in OtoDom.pl data for 12 cities between 2012Q1 and 2014Q4

Variable	Min	Q_1	Mean	Median	Q_3	Max
Number of rooms	0	2	4	4	5	255
Average floor area	0.10	73.50	135.01	120	178.50	500
Average Price m^2	0	3 996.02	7 158.53	5 122.46	7 170.73	3 235 190
Median Price	1	322 500	854 456.98	530 000	937 000	22 000 000
Average Price	1	333 414.97	879 481.33	550 000	980 000	22 000 000

chose the number 0 to denote missing values. However, the upper limit was not restricted. The maximum number of rooms indicated in advertisements on OtoDom.pl was 255, while the largest floor area was equal or over 500 m^2. Variables concerning prices also contained false information. For instance, the average price per m^2 equal to 0 PLN/m2 or the price of one property equal to 1 PLN. Some maximum values were also implausible: for example, the average price per m^2 equal to 3 mln PLN. Hence, the data were cleaned using the following conditions:

- the number of rooms was restricted to be within the range from 1 to 9,
- the floor area was restricted to range between 10 m^2 and 200 m^2,
- the median price was set to be lower than 10 mln PLN,
- cross-classifications with fewer than 10 ads were excluded,
- the average price was restricted to lie within the range from 0.01th to 99.9th percentile.

This procedure reduced the initial 100 910 records to 24 542, resulting in a total of 14 661 903 ads, comprised with the initial number of 17 935 280 ads. Table 4.15 contains information about the number of advertisements in 12 cities between 2012Q1 and 2014Q4. The biggest number of advertisements was recorded

for Warszawa, the lowest for Opole. In comparison to Dom.Gratka.pl, the number of ads for Warszawa was over one million higher and, in relation to Nieruchomosci-Online.pl, was 20 times higher (in comparison to 15.6 on OtoDom.pl). For Kraków the number of ads was similar to the number posted on Dom.Gratka.pl. In the case of Poznań, the number of ads posted on OtoDom.pl was over twice as high as on Dom.Gratka.pl.

TABLE 4.15: The number of advertisements posted on OtoDom.pl for 12 cities between 2012Q1 and 2014Q4

City	# Advertisements	# Advertisements [%]
Warszawa	4 503 973	43.61
Kraków	1 535 955	14.87
Wrocław	1 159 789	11.23
Poznań	685 199	6.63
Gdańsk	545 972	5.29
Szczecin	496 267	4.81
Łódź	395 033	3.83
Białystok	283 787	2.75
Katowice	253 814	2.46
Lublin	220 261	2.13
Olsztyn	140 921	1.36
Opole	106 133	1.03

Figure 4.5 shows how the number of advertisements for 12 cities changed between 2012Q1 and 2014Q4. To improve readability, the square root of the number of ads is used. As in the case of Dom.Gratka.pl, for most of the period the number of ads in each city remains relatively stable. However, in 2014 there is a decrease for all cities. Based on Figure 4.5 three groups of cities can be distinguished: (1) Warszawa, with the highest number of ads, (2) Kraków i Wrocław, like on Dom.Gratka.pl, and (3) the remaining cities.

Figure 4.6 presents the distribution of the average price per m^2 on OtoDom.pl and Table 4.16 shows basic statistics about the distribution. The same process of assigning weights was applied as in the case of Dom.Gratka.pl. The weighted distribution has become multimodal but the peaks are higher and more leptocurtic, suggesting a higher number of advertisements that refer to the same units. After applying weights the distribution has become less skewed and has shrunk towards the average.

The average weighted price per m^2 on OtoDom.pl was equal to 6 959 PLN/m^2, compared to 7 152 PLN/m^2 on Dom.Gratka.pl. The weighted median price was equal to 6 946 PLN/m^2 and was lower than on Dom.Gratka.pl (7 231 PLN/m^2). Prices observed on OtoDom.pl were higher than those on Dom.Gratka.pl. More differences between the two data sources and a comparison to the reference data from the NBP/CSO survey will be provided in the next section.

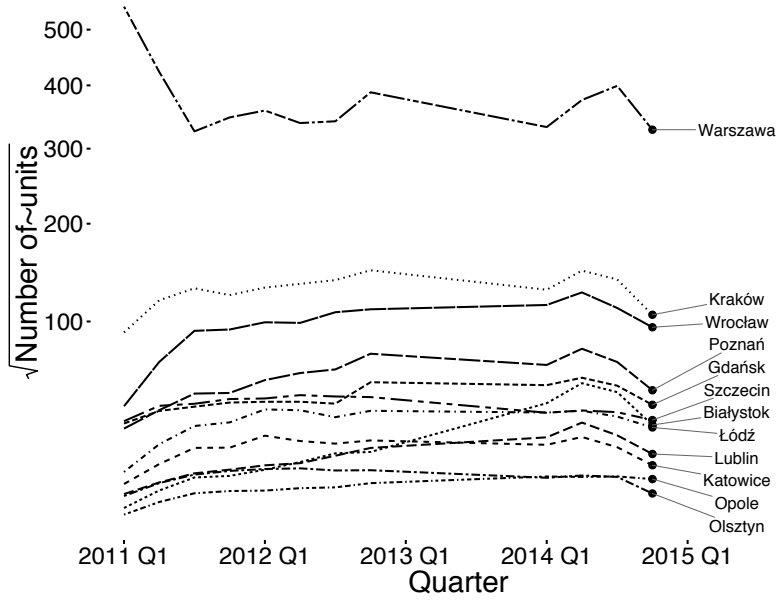


FIGURE 4.5: Number of units observed in OtoDom.pl between 2012Q1 and 2014Q4

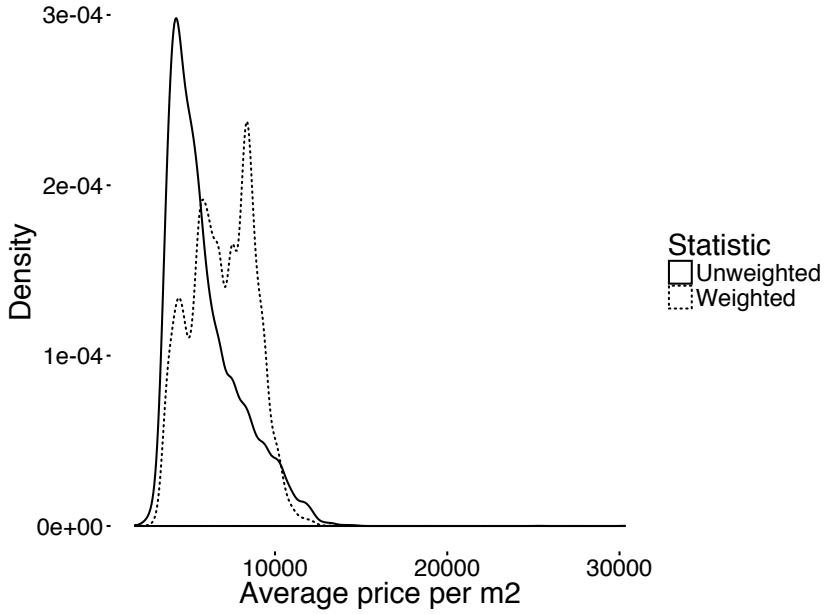


FIGURE 4.6: Unweighted and weighted density plot of average price per m^2 between 2012Q1 and 2014Q4 on OtoDom.pl

4.4 Empirical assessment of representativeness

The first section of this chapter provided an overall assessment of online presence of real estate companies. It was the first step in the process of measuring representativeness, which was proposed in Chapter 3. This section describes the second step this procedure. Given the aggregated nature of the data and the kind of cross-classifications in the sources, a comparison of distributions will be conducted according to categories found in the NBP/ CSO survey. For that purpose,

TABLE 4.16: Weighted and unweighted average offer price per m^2 between 2012Q1 and 2014Q4 on OtoDom.pl

Statistic	Unweighted Price m^2	Weighted Price m^2	Weight
Min	1707.39	1707.39	11
Percentile 5%	3704	3998	13
Percentile 10%	3943	4392	16
Q_1	4446	5596	31
Mean	5996	6959	597.4
Median	5444	6946	94
Q_3	7143	8419	425
Percentile 90%	9218	9357	1337
Percentile 95%	9719	9541	2848
Max	30322.61	30322.61	28972

the following variables for 12 cities in period 2012Q1 and 2014Q4 will be analysed and compared:

- a comparison of distributions of the number of rooms – reference data: the NBP/ CSO survey,
- a comparison of distributions of the floor area – reference data: NBP/ CSO survey,
- a comparison of distributions of the offer and transaction price only for Poznań - reference data: the Register of Real Estate Prices and Values.

The variable Number of rooms was split into four groups, following the grouping adopted in the NBP/ CSO survey – 1, 2, 3 and 4+ rooms. The floor area was also divided into four groups, in line with the NBP/ CSO survey: up to 40 m^2, from 40 to 60 m^2, from 60 to 80 m^2 and over 80 m^2.

The definition of representativeness provided by Bethlehem (2009) was used in order to assess whether the distributions of the number of rooms and the floor area are similar to those recorded in the NBP/CSO survey. To verify whether the differences between distributions found in the IDSs and the NBP/ CSO survey are significant the following measures were used:

- Bias is given by equation (4.1):

$$Bias = \tilde{\theta}_{k,j,d,t} - E(\theta_{j,d,t}) = \tilde{\theta}_{k,j,d,t} - \hat{\theta}_{j,d,t}, \tag{4.1}$$

  where $\hat{\theta}_{d,t}$ is the NBP/CSO based fraction of category j of a given variable, d refers to domain, t refers to quarter and k denotes a data source,

- Absolute relative bias is given by equation (4.2):

$$ARB = \left| \frac{\tilde{\theta}_{k,j,d,t} - E(\theta_{j,d,t})}{E(\theta_{j,d,t})} \right| = \left| \frac{\tilde{\theta}_{k,j,d,t} - \hat{\theta}_{j,d,t}}{\hat{\theta}_{j,d,t}} \right|, \tag{4.2}$$

  where symbols have the same denotations as in equation (4.1),

- Hellinger's distance given by equation (3.42),
- χ^2 goodness of fit test given by equation (3.43).

Finally, in the case of the Register of Transactions values of the price, floor area and the average price per m^2 were compared using quantile-quantile plots and the goodness of fit test.

4.4.1 A comparison with the NBP/CSO survey

The distribution of the number of rooms variable

Table 4.17 provides information about the distribution of bias and absolute relative bias (ARB) given by (4.1) for 12 cities between 2012Q1 to 2014Q4. Bias was calculated separately for each level of the variable and each IDS source. Mean bias is equal to zero, which is due to the character of the variable (percentages s) and average differences between categories of this variable cancel each other out. The cities differ in the level of bias, for instance the smallest differences can be found for Kraków (the minimum is equal to -9.1 percentage points, the maximum is equal to 7.9 p.p.) and Olsztyn (minimum is equal to -9.2 p.p., maximum is equal to 9.5 p.p.). The highest bias is observed for Opole (from -20.8 to 22.9 p.p.), Łódź (from -18.3 to 14.5 p.p.) and Lublin (from -17.4 to 12.9 p.p.).

On average ARB is equal to 18%. On average the lowest ARB is observed for Warszawa (6.4%), Kraków (8.9%) and Olsztyn (9.8%), while the highest for Opole (51.7%) and Wrocław (23.2%). In the case of Opole the high level of ARB is mainly due to the small sample size observed in the NBP/CSO survey (on average 34 properties). ARB above 100% is observed for Opole (544.0%), Łódź (223.2%) and Lublin (154.7%). In 8 cities median ARB is above 10% (Gdańsk, Kraków, Łódź, Lublin, Olsztyn and Warszawa). The measures presented in Table 4.17 provided information at the city level, which can be used to determine which cities in IDSs are characterised by the highest bias in the number of rooms. The following section will focus on identifying groups that are under- or overrepresented in IDSs.

Table 4.18 provides information about the distribution of bias and absolute relative bias across the IDS sources and levels of the variable Number of rooms. On average the highest differences can be observed for Nieruchomosci-online.pl (abbreviated as N-online.pl), while the smallest for OtoDom.pl. The highest values of ARB are observed for the 4+ category, in particular 544.0% for Dom.Gratka.pl, 438.3% OtoDom.pl and 377.3 % for Nieruchomosci-online.pl. On average ARB indicates that the fraction of small (one room) and big (4 and more rooms) properties is highly biased. Median ARB for small properties is 19.3% for Dom.Gratka.pl, 21.2% for Nieruchomosci-Online.pl and 13.6% for Dom.Gratka.pl. In the case of biggest properties, ARB for Dom.Gratka.pl is equal to 12.2%, for Nieruchomosci-online.pl 22.3% and for OtoDom.pl 13.8%.

Medium-sized properties (2 and 3 rooms) are characterised by smaller ARB in comparison with two groups presented above. For 2 and 3 rooms, median ARB for Dom.Gratka.pl is equal to 6.6% and 5.7%, for Nieruchomosci-online.pl is equal to 10.4% and 8.2% and for OtoDom.pl - 7.7% and 6.1%. Results presented in Table 4.18 indicate that the distribution of medium-sized properties

TABLE 4.17: The distribution of bias and absolute relative bias (ARB) for the variable Number of rooms in 12 cities between 2012Q1 to 2014Q4

City	Min	Q_1	Median	Mean	Q_3	Max
Bias [in p.p.]						
Białystok	-11.6	-2.6	-0.2	0.0	2.2	13.9
Gdańsk	-10.9	-1.5	0.3	0.0	1.7	12.9
Katowice	-12.0	-3.0	-0.4	0.0	3.0	16.5
Kraków	-7.9	-2.1	0.1	0.0	1.6	9.1
Łódź	-14.5	-2.0	0.1	0.0	2.3	18.3
Lublin	-12.9	-1.6	-0.2	0.0	1.6	17.4
Olsztyn	-9.5	-1.4	-0.2	0.0	1.5	9.2
Opole	-22.9	-4.5	0.8	0.0	4.5	20.8
Poznań	-10.2	-2.1	0.7	0.0	2.6	7.9
Szczecin	-9.9	-3.2	0.1	0.0	3.0	8.9
Warszawa	-10.7	-1.3	-0.0	0.0	1.2	10.8
Wrocław	-7.3	-4.5	-1.1	0.0	4.0	10.4
Absolute relative bias [in %]						
Białystok	0.1	4.9	11.6	16.0	18.9	89.9
Gdańsk	0.3	3.5	7.7	11.5	13.3	91.6
Katowice	0.1	6.6	13.7	18.9	29.4	65.9
Kraków	0.1	4.2	7.7	8.9	12.8	26.1
Łódź	0.2	3.8	9.8	19.9	23.9	223.2
Lublin	0.0	3.9	7.6	11.8	13.4	154.7
Olsztyn	0.2	2.4	7.2	9.8	12.5	67.3
Opole	0.1	10.0	23.3	51.7	53.5	544.0
Poznań	0.2	4.5	10.5	14.2	19.9	76.9
Szczecin	0.1	5.5	13.6	19.6	27.4	80.2
Warszawa	0.0	2.5	6.4	10.8	15.0	92.0
Wrocław	0.4	12.3	20.6	23.2	30.2	87.2

is closer to that found in the NBP/CSO survey, while the two extreme groups (1 and 4+) have different distributions. Table 4.17 indicates that cities vary in their level of bias and ARB, which means that the differences should be further studied. For this purpose, five cities have been selected for further analysis in the dissertation, while data for the remaining cities are presented in Appendix A.2.

The 5 cities selected for further analysis are: (1) Olsztyn, which is characterised by the lowest bias, (2) Opole, characterised by the highest bias, (3) Warszawa, which has the lowest ARB, (4) Wrocław, with the second highest ARB, and (5) Poznań, which will later be compared with the Register of Transactions.

Figure 4.7 presents a comparison between the NBP/CSO survey and IDS data for the selected cities between 2012Q1 and 2014Q4. Figure 4.7 presents point estimates for each IDS and the NBP/CSO survey (red colour). Additionally, for each dataset and city LOESS-based trends are provided. In Olsztyn the biggest differences between the IDSs and the NBP/CSO survey can be seen for properties with only two rooms. Nieruchomosci-Online.pl overrepresents properties with

TABLE 4.18: The distribution of bias and absolute relative bias for the variable Number of rooms for IDSs between 2012Q1 to 2014Q4

IDS source	# Rooms	Min	Q_1	Median	Mean	Q_3	Max
Bias [in p.p.]							
Dom.Gratka.pl	1	-14.5	-2.7	-1.0	-1.4	0.7	8.7
Dom.Gratka.pl	2	-21.7	-2.5	0.4	-0.2	2.5	10.3
Dom.Gratka.pl	3	-7.9	-2.3	0.5	0.4	2.0	18.3
Dom.Gratka.pl	4+	-7.0	-0.5	0.6	1.1	2.5	15.5
N-online.pl	1	-8.9	-0.5	1.3	1.0	2.9	10.1
N-online.pl	2	-20.1	0.6	3.3	3.3	6.9	17.4
N-online.pl	3	-14.3	-4.2	-1.9	-1.7	0.4	20.8
N-online.pl	4+	-12.9	-5.2	-2.6	-2.6	-0.2	10.8
OtoDom.pl	1	-12.0	-1.8	-0.3	-0.9	0.8	4.5
OtoDom.pl	2	-22.9	-0.7	1.6	1.0	3.6	12.2
OtoDom.pl	3	-11.6	-2.2	0.2	0.4	2.2	19.3
OtoDom.pl	4+	-9.9	-2.4	-1.1	-0.6	0.6	12.5
Absolute relative bias [in %]							
Dom.Gratka.pl	1	0.2	9.2	19.3	25.9	32.2	235.8
Dom.Gratka.pl	2	0.1	3.3	6.6	8.5	10.8	37.5
Dom.Gratka.pl	3	0.1	3.0	5.7	8.8	10.9	94.9
Dom.Gratka.pl	4+	0.2	3.5	12.2	27.0	28.2	544.0
N-online.pl	1	0.1	8.9	21.2	30.2	37.8	281.0
N-online.pl	2	0.3	5.2	10.4	13.7	19.2	59.8
N-online.pl	3	0.1	4.5	8.2	10.8	13.6	108.0
N-online.pl	4+	0.1	11.0	22.3	28.7	36.5	377.3
OtoDom.pl	1	0.1	5.9	13.6	19.6	25.4	171.6
OtoDom.pl	2	0.0	3.4	7.7	9.6	12.9	44.1
OtoDom.pl	3	0.0	2.5	6.1	8.9	10.6	100.4
OtoDom.pl	4+	0.3	6.8	13.8	24.7	25.7	438.3

Note: N-online.pl is short for Nieruchomosci-online.pl.

two rooms in Olsztyn, while Dom.Gratka.pl underepresents them. The closest values to the NBP/CSO survey can be observed across all categories for OtoDom.pl.

In Opole all data sources vary in the fractions of properties. Owing to its small sample size, NBP/CSO survey data on properties with two and three rooms vary between quarters, there is also some variation for the smallest and biggest properties but it is not as high as for the middle categories. Slightly less variation can be observed for IDSs. Trends observed for properties with two, three and four and more rooms are consistent between 2012Q1 to 2013Q4, but in 2014 the differences are more pronounced. For Olsztyn all data sources indicate a similar level only for the group of smallest properties. The differences for other categories are mainly due to the small sample size in the NBP/CSO survey.

In Poznań all IDSs overrepresent properties with two and three rooms; however, for two-room properties, the differences are bigger. In the last quarter,

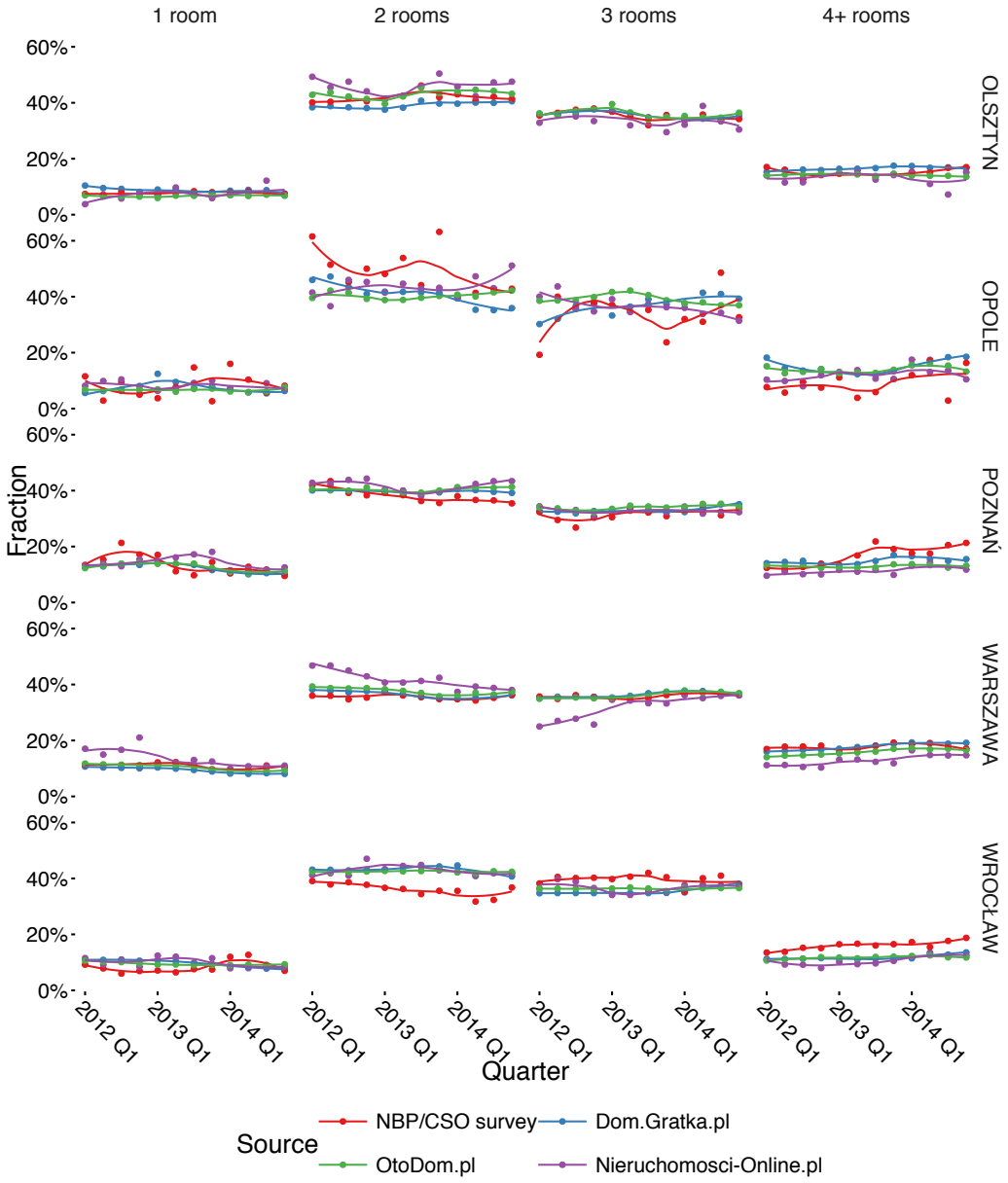


FIGURE 4.7: Comparison of number of rooms distribution in NBP/CSO survey and IDS in Olsztyn, Poznań, Opole, Wrocław and Warszawa

the differences between IDSs and the NBP/CSO survey are more visible. All IDS underrepresent properties with four or more rooms. This trends can be observed from 2012Q4, when there is an increase in the fraction of 4+ properties in the NBP/CSO survey. The most consistent group for Poznan are properties with only one room.

The NBP/CSO survey, Dom.Gratka.pl and OtoDom.pl provide similar estimates of fractions for all types of properties categorized by the number of rooms. Slight differences are visible in the category of 4+ rooms. On the other hand, Nieruchomosci-Online.pl provide biased estimates of fractions for all categories.

For instance, Nieruchomosci-Online.pl systematically underestimates the fraction of properties with 4+ rooms. The number of properties with two rooms, on the other hand, is systematically overestimated, although the bias decreases over time.

In the case of Wrocław, there is systematic bias in the representation of all types of properties shown in Figure 4.7. Bias is constant over time for each property category and is highest for two-room apartments (on average 8 p.p.). All IDSs systematically underestimate the fraction of properties with three or more rooms. In the case of apartments with 4+ rooms, bias is systematic and does not decrease over time. For three-room properties there is a decline in bias in 2014. Examination of trends presented in Figure 4.7 indicates that the IDS data contain systematic bias. Only in a few cases are the fractions of properties in different size categories at the same level as in the NBP/CSO survey.

The statistical significance of differences between the IDS-based distributions of the number of rooms and the NBP/CSO survey as the reference dataset was measured using Hellinger's distance and χ^2 goodness of fit. Table 4.19 shows the distribution of Hellinger's distance, broken down by IDS and city. Hellinger's distance is calculated independently for each IDS, city and quarter and, to improve readability, has been rescaled from [0,1] to [0,100]. Values close to 0 indicate that the distance between two distributions is equal, implying that the distributions are consistent.

TABLE 4.19: The distribution of Hellinger's distance for IDSs and cities between 2012Q1 to 2014Q4

Source/Domain	Min	Q_1	Median	Mean	Q_3	Max
Dom.Gratka.pl	0.3	2.4	3.8	5.3	7.9	20.3
Nieruchomosci-online.pl	0.3	3.9	5.5	6.7	9.1	17.6
OtoDom.pl	0.3	2.6	3.7	5.0	6.9	19.5
Białystok	1.1	2.9	3.8	5.6	8.7	17.4
Gdańsk	0.3	2.5	3.3	3.5	4.1	13.7
Katowice	1.7	3.7	7.3	7.4	10.1	13.8
Kraków	0.5	2.2	2.9	3.3	4.2	7.1
Łódź	1.1	3.3	5.1	6.2	7.2	20.3
Lublin	0.3	2.7	3.3	4.2	4.2	17.2
Olsztyn	0.3	2.3	3.0	3.4	4.0	11.7
Opole	2.2	7.6	11.2	11.3	15.2	19.5
Poznań	1.8	3.7	4.7	5.5	7.5	11.6
Szczecin	1.8	4.3	6.8	6.5	8.8	11.0
Warszawa	1.2	2.0	2.5	3.9	3.5	14.0
Wrocław	3.9	6.5	7.7	7.4	8.3	11.4

Median Hellinger's distance is the lowest for Dom.Gratka.pl (3.8) and the highest for Nieruchomosci-online.pl (5.5). The mimimum value for all data sources is the same and equals 0.3, but occurs for different cities and periods: 2014Q3 Gdańsk (Dom.Gratka.pl), Olsztyn 2013Q1 (Nieruchomosci-Online.pl) and Lublin 2013Q2 (OtoDom.pl). The highest differences between distributions of the number of rooms can be observed for Łódź 2014Q3 (Dom.Gratka.pl), Opole 2012Q1

(Nieruchomosci-online.pl and OtoDom.pl). Hellinger's distance indicates that OtoDom.pl and Dom.Gratka.pl are most consistent with the NBP/CSO survey.

TABLE 4.20: Results of χ^2 test for goodness of fit by IDS and city

Source/city	H_0 was rejected	H_0 was not rejected
Dom.Gratka.pl	142	2
Nieruchomosci-online.pl	118	26
OtoDom.pl	93	51
Białystok	29	7
Gdańsk	30	6
Katowice	27	9
Kraków	35	1
Łódź	31	5
Lublin	21	15
Olsztyn	14	22
Opole	33	3
Poznań	33	3
Szczecin	32	4
Warszawa	32	4
Wrocław	36	0

The second part of Table 4.19 shows the distribution od Hellinger's distance for the selected cities. Hellinger's distance for Warszawa (median 2.5), Kraków (median 2.9) and Olsztyn (median 3.0) is closest to the reference distribution found in the NBP/CSO survey. As was the case with the bias distribution, median Hellinger's distance for Wrocław, Katowice and Opole is the highest. Table 4.19 indicates that Kraków, Warszawa and Olsztyn are cities with the smallest differences in distributions of properties by the number of rooms between IDSs and the NBP/CSO survey.

In order to verify whether differences between these distributions are significant χ^2 test for goodness of fit was calculated. Aggregated results are presented in Table 4.20 containing two columns showing the number of times H_0 was or was not rejected. The test was conducted for each cross-classification of IDS, city and quarter. Table 4.20 shows the number of cases when the null hypothesis was rejected, implying a significant difference between distributions (column H_0 was rejected) or when it was rejected, suggesting that the distribution of the number of rooms variable was similar (H_0 was not rejected). In addition, the results are broken down by IDS and city.

Results for OtoDom.pl indicate that in 51 cases (35%) the distribution of the number of rooms variable was equal to that found in the NBP/CSO survey. However, in 93 cases the null hypothesis was rejected. The null hypothesis was not rejected for all quarters for Olsztyn and for nearly all quarters for Lublin (11 out of 12 quarters), Białystok (7 out of 12 quarters) and Katowice (5 out of 12 quarters). The highest number of cases when H0 was rejected can be observed for Dom.Gratka.pl - 142 (99%), which indicates that the of properties by the number of rooms observed in this data source is significantly different from that found in the NBP/CSO survey.

Table 4.20 also shows results for the cities. There are substantial differences between the cities in the number of times H0 is rejected. For instance, for Wrocław and Kraków none of IDSs was consistent with the NBP/CSO estimates. Poznań and Opole have the same number of H0 rejections, despite a substantial difference between the average sample size for each of them in the NBP/CSO survey. As stated previously, Olsztyn and Lublin are characterised by the lowest number of H0 rejections.

Distribution of the floor area variable

Another step of measuring representativeness involves comparing distributions of floor area. In the NBP/CSO survey this variable is divided into four groups: up to 40 m^2, $(40, 60]$ m^2, $(60, 80]$ m^2 and over 80 m^2. In order to compare distributions, floor area in Nieruchomosci-online.pl and OtoDom.pl was harmonized with the NBP/CSO survey. Data from Dom.Gratka.pl were categorized by owner into three groups: up to 35 m^2, $(35, 70]$ m^2 and over 70 m^2, which prevented a comparison with the NBP/CSO survey. Therefore, Dom.Gratka.pl is excluded from comparison for this variable.

Table 4.21 presents distributions of bias and absolute relative bias for floor area based on data from OtoDom.pl and Nieruchomosci-online.pl for 12 cities between 2012Q1 to 2014Q4. As in the case of the number of rooms, the highest bias is observed for Opole, with a minimum of -25.8 p.p. and a maximum of 14.4 p.p. ARB for Opole varies from 1.4% to 556.7%. In contrast to its results with respect to the number of rooms, Białystok is characterised by the second highest median ARB equal to 29.2%, with a maximum of 132.8%. The lowest median ARB is observed for Kraków and Warszawa (8.3% and 10.4% respectively). On average the floor area variable is characterised by higher median bias than the number of rooms. Median ARB for floor area is equal to 16.7%, compared to 12.5% for the number of rooms.

The difference between distributions of bias for floor area and for the number of rooms can be due to the character of these variables. Floor area is a continuous variable, unlike the number of rooms, which can only take on discrete values. Categorization of floor area is based on values specified by the broker or owner, which can contain measurement error (e.g. due to rounding), while the number of rooms is likely to contain less measurement error due to the character of this variable.

Table 4.22 shows bias and absolute relative bias for two IDSs and four categories of floor area between 2012Q1 and 2014Q4. Analysis of bias indicates that, on average, small properties (the first two categories) are overrepresented in comparison with the NBP/CSO survey. For Nieruchomosci-online bias for the category of up to 40 m^2 is equal to 4.3 p.p. and for (40,60] m^2 is 3.1 p.p.. OtoDom.pl data reveal a similar pattern, but bias is lower and equal to 0.8 p.p. for the first category, and 1.9 p.p. for the second. Two categories that represent the biggest properties are underrepresented in IDSs. On average, differences in the data from Nieruchomosci-online.pl are higher than those observed for OtoDom.pl for categories (60,80] m^2 and over 80 m^2. For the third category bias is equal to -3.0 p.p., compared to -1.2 p.p. for OtoDom.pl. For the fourth category bias is equal to -4.4 p.p., compared to -1.5 p.p. for OtoDom.pl. These differences can result

TABLE 4.21: The distribution of bias and absolute relative bias for floor area for 12 cities between 2012Q1 to 2014Q4

City	Min	Q_1	Median	Mean	Q_3	Max
Bias						
Białystok	-13.7	-5.9	-2.3	0.0	5.4	21.4
Gdańsk	-13.6	-3.7	-0.6	0.0	3.8	12.6
Katowice	-16.5	-3.3	-0.7	0.0	2.4	17.6
Kraków	-7.3	-1.8	-0.4	0.0	2.3	7.9
Łódź	-11.0	-3.0	-0.8	0.0	3.5	10.7
Lublin	-14.9	-4.1	-0.1	0.0	4.0	13.3
Olsztyn	-8.4	-2.6	-0.6	0.0	2.7	10.7
Opole	-25.8	-5.5	2.4	0.0	6.0	14.4
Poznań	-13.8	-3.2	0.6	0.0	3.8	10.6
Szczecin	-10.9	-3.4	0.3	0.0	3.1	11.3
Warszawa	-13.4	-2.6	-0.1	0.0	2.5	14.7
Wrocław	-11.4	-5.9	-0.8	0.0	5.5	13.1
Absolute relative bias						
Białystok	1.3	15.5	29.2	35.2	42.8	132.8
Gdańsk	0.7	9.2	16.7	21.4	27.4	126.4
Katowice	0.2	6.0	13.9	18.0	25.0	68.1
Kraków	0.4	3.4	8.3	11.7	16.6	68.6
Łódź	0.3	7.2	13.1	17.9	22.8	69.5
Lublin	0.1	6.1	14.6	20.9	29.3	107.4
Olsztyn	0.3	6.9	12.4	14.8	19.5	65.2
Opole	1.4	15.6	29.8	47.5	50.5	556.7
Poznań	0.2	4.7	15.7	17.9	28.0	55.0
Szczecin	0.8	6.6	12.4	17.1	21.9	100.2
Warszawa	0.1	4.6	10.4	16.5	19.0	81.7
Wrocław	2.0	16.5	24.1	28.0	37.2	81.3

from the popularity of a given data source. For instance OtoDom.pl has a higher number of visitors. compared to Nieruchomosci-Online.pl

In terms of ARB, as was the case with the number of rooms, the highest values can be seen for the first (up to 40 m^2) and the last category (over 80 m^2). This situation is due to a higher representation of the two middle categories, compared to the bottom and top categories. Median ARB for the first category is equal to 26.4% for Nieruchomosci-Online.pl and is twice as high as that for OtoDom.pl (13.4%). For properties with a floor area within (40,60] m^2, median ARB in data from Nieruchomosci-Online.pl is equal to 12.0%, versus 11.2% for OtoDom.pl. Median ARB for the third category is equal to 9.0% for OtoDom.pl and 14.5% for Nieruchomosci-online.pl. The biggest differences can be seen for the largest properties. For Nieruchomosci-online.pl median ARB is equal to 34.2% and 24.6% for OtoDom.pl. In addition, maximum values of ARB can also be observed for the bottom and top categories: 132.8% and 82.1% for the first category and over 377.2% and 556.7% for the last category.

Figure 4.8 presents the distribution of floor area in the NBP/CSO survey and

TABLE 4.22: Th distribution of bias and absolute relative bias for floor area for IDS between 2012Q1 to 2014Q4

Source	Floor area	Min	Q_1	Median	Mean	Q_3	Max
Bias [in p.p.]							
N-online.pl	to 40 m^2	-13.1	1.7	4.0	4.3	6.7	21.4
N-online.pl	(40,60] m^2	-21.9	0.6	4.1	3.1	6.5	17.6
N-online.pl	(60,80] m^2	-14.9	-6.1	-2.9	-3.0	-0.6	13.1
N-online.pl	over 80 m^2	-16.5	-7.6	-4.7	-4.4	-1.7	9.6
OtoDom.pl	to 40 m^2	-16.3	-0.9	1.2	0.8	3.6	7.6
OtoDom.pl	(40,60] m^2	-25.8	0.6	3.3	1.9	5.3	13.1
OtoDom.pl	(60,80] m^2	-10.0	-3.7	-1.5	-1.2	0.6	14.4
OtoDom.pl	over 80 m^2	-11.0	-4.2	-2.4	-1.5	0.4	13.9
Absolute relative bias [in %]							
N-online.pl	to 40 m^2	0.5	12.4	26.4	34.1	51.1	132.8
N-online.pl	(40,60] m^2	0.6	5.5	12.0	13.8	18.4	59.0
N-online.pl	(60,80] m^2	0.3	6.1	14.5	17.9	25.7	71.9
N-online.pl	over 80 m^2	0.5	19.6	34.2	37.4	50.4	377.2
OtoDom.pl	to 40 m^2	0.3	6.0	13.4	18.7	28.3	82.1
OtoDom.pl	(40,60] m^2	0.7	6.4	11.2	12.1	15.5	43.1
OtoDom.pl	(60,80] m^2	0.1	4.4	9.0	13.1	16.7	96.8
OtoDom.pl	over 80 m^2	0.4	13.8	24.6	30.7	32.9	556.7

Note: N-online.pl refers to Nieruchomosci-online.pl.

IDSs for five selected cities (Olsztyn, Poznań, Opole, Wrocław and Warszawa). The biggest differences are visible for Opole, particularly for the second and third category of properties. These differences are mainly due to variability of NBP/CSO survey estimates, which are based on a small sample (on average 35 observations). Estimates based on data from Nieruchomosci-Online.pl and OtoDom.pl also vary for Opole, but trends are shaped similarly.

IDS-based estimates of the distribution of floor area for Olsztyn are similar to those based on the NBP/CSO survey. Differences can be seen mainly for the second category ((40,60] m^2), where OtoDom.pl systematically overestimates the fraction of these flats, compared to the NBP/CSO survey. Data from Nieruchomosci-online.pl are characterised by more variability than those from OtoDom.pl and the NBP/CSO survey. Nonetheless, visual examination of trends observed for Olsztyn indicates consistency across all data sources.

in the case of Poznań, only the distribution of the third category of properties is similar across the 4 sources. The biggest differences can be observed for the smallest and biggest apartments. In both cases NBP/CSO data reveal a trend change starting from 2013Q1. Substantial bias is observed for the category of properties over 80 m^2, which is underestimated in IDSs by 10 p.p. Moreover, this difference can be seen in both IDS sources. In the case of apartments of up to 40 m^2 between 2012Q1 and 2013Q1 IDSs underestimate this fraction, while after 2013Q1, due to the trend change in NBP/CSO data both IDS slightly

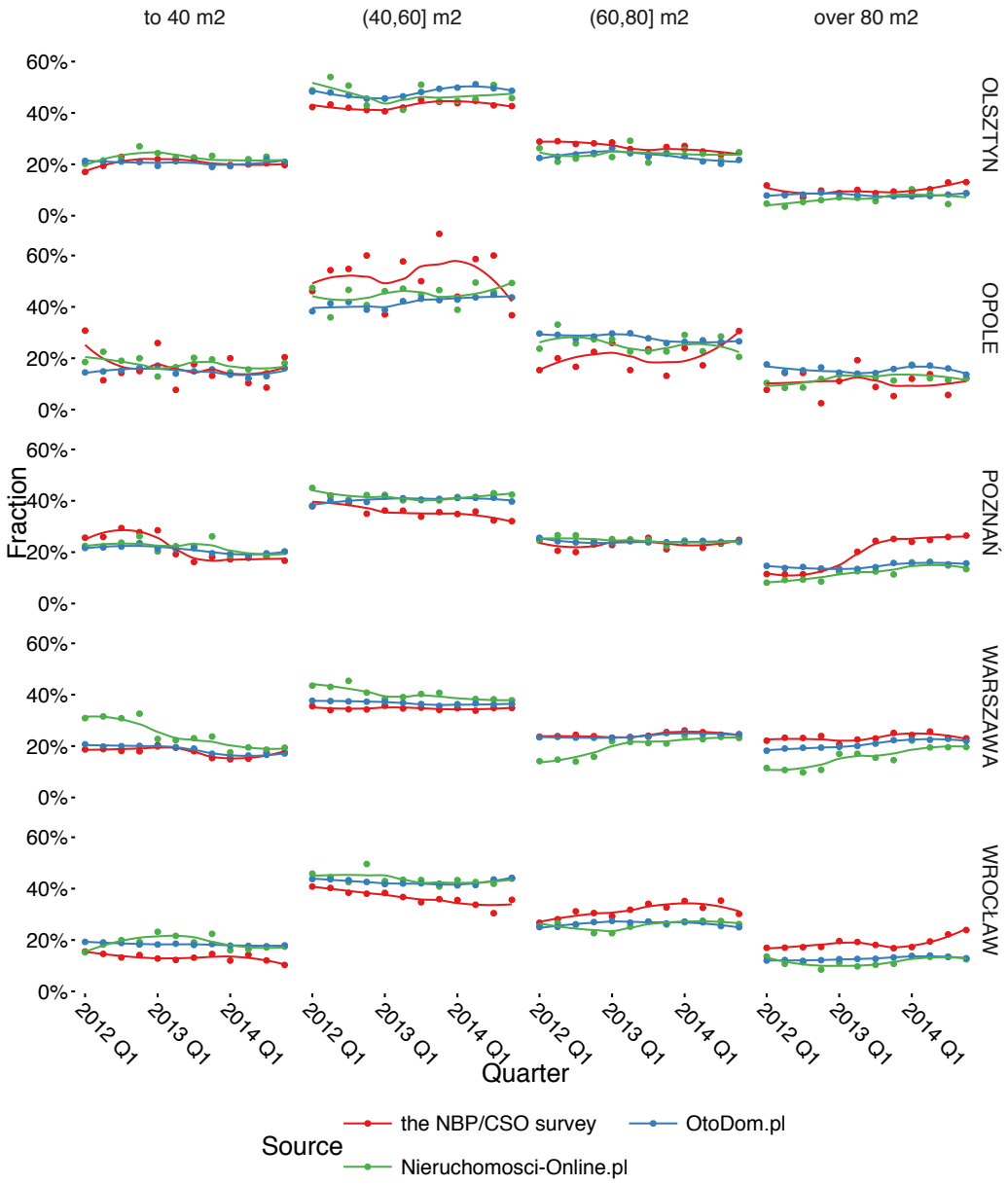


FIGURE 4.8: Comparison of floor area distribution in the NBP/CSO survey and IDS in Olsztyn, Poznań, Opole, Wrocław and Warszawa

overestimate the precentage of properties in this group. An interesting trend is revealed for the second floor area category. There is systematic overrepresentation of apartments within (40,60] m^2 category in both IDS, which increases over time. The trend estimated on the basis of NBP/CSO survey data indicates a decrease in this group, while IDS data indicate that this fraction is stable over time. Consistency between the three data sources is retained for the third category of flats in Poznań.

In terms of the distribution of the number of rooms, Warszawa is characterised by the lowest differences, but with respect to floor area, differencess between

the sources are substantial. Systematic bias can be observed for Nieruchomosci-Online.pl, while estimates based on data from OtoDom.pl are consistent, trend-wise, for all floor area categories. Data from Nieruchomosci-online.pl systematically overestimate the first two groups of properties, but this bias decreases over time and reaches its minimum level in the last quarter of 2014. a similar trend can be observed for the two top categories, where there is systematic underestimation, which also decreases over time only to reach almost zero in 2014Q4. Data from OtoDom.pl provide similar point estimates to those obtained from the NBP/CSO survey.

The last city presented in the analysis in Figure 4.7 is Wrocław. As was the case with the distribution of the number of rooms, systematic bias is also present but this time it is higher. For instance, both IDSs reveal systematic over-representation of smallest flats (up to 40 m^2 and (40,60] m^2). The difference is, on average, equal to 5 p.p. Trends observed in these groups are stable over time, but there is a slight decrease for the second category in the NBP/CSO survey. The last two categories ((60,80] m^2 and over 80 m^2) are systematically underrepresented. Data from Nieruchomosci-online.pl and OtoDom.pl provide estimates with a similar amount of bias, which increases over time for both categories.

TABLE 4.23: The distribution of Hellinger's distance for floor area by IDS and city between 2012Q1 to 2014Q4

Source/Domain	Min	Q_1	Median	Mean	Q_3	Max
Nieruchomosci-online.pl	1.2	5.2	8.2	9.0	12.5	20.9
OtoDom.pl	1.1	4.0	6.0	6.3	8.0	20.9
Białystok	6.1	7.5	9.5	11.6	15.6	20.9
Gdańsk	2.3	5.1	5.9	6.9	8.1	18.9
Katowice	1.7	4.7	6.6	7.6	9.1	18.7
Kraków	1.5	2.7	4.1	4.3	5.8	7.6
Łódź	2.0	3.5	5.8	6.8	8.9	15.2
Lublin	1.2	4.9	6.0	7.4	9.2	19.9
Olsztyn	1.5	3.9	5.1	5.5	6.7	11.5
Opole	5.3	8.2	11.3	12.2	15.8	20.9
Poznań	4.0	6.3	7.4	7.9	9.9	13.4
Szczecin	1.4	4.0	5.7	6.0	7.0	11.9
Warszawa	1.1	2.6	3.8	6.1	7.2	18.4
Wrocław	4.2	7.9	8.8	9.5	12.2	13.4

The degree of similarity between floor area distributions observed in data from OtoDom.pl, Nieruchomosci-online.pl and the NBP/CSO survey was measured using Hellinger's distance and χ^2 test for goodness of fit. The first measure was calculated for IDS and cities separately and its values are presented in Table 4.23. On average, floor area distributions based on data from OtoDom.pl are closer to those found in the NBP/CSO survey. Average Hellinger's distance is equal to 6.3 for OtoDom.pl and 9.0 for Nieruchomosci-online.pl. This measure is close to the average distance calculated for distributions of the number of rooms (5.0). Values of Hellinger's distance indicate that data from OtoDom.pl provide estimates that are closer to official estimates based on the NBP/CSO survey.

TABLE 4.24: Results of χ^2 test for goodness of fit for IDSs and cities

Source/City	H_0 was rejected	H_0 was not rejected
Nieruchomosci-online.pl	131	13
OtoDom.pl	118	26
Białystok	22	2
Gdańsk	24	0
Katowice	18	6
Kraków	23	1
Łódź	20	4
Lublin	18	6
Olsztyn	10	14
Opole	24	0
Poznań	24	0
Szczecin	21	3
Warszawa	21	3
Wrocław	24	0

Table 4.23 shows information for 12 cities. On average the highest values of Hellinger's distance were obtained for Opole (12.2) and Białystok (11.6) while the lowest for Kraków (4.3) and Olsztyn (5.5). The values of Hellinger's distance indicate that the smallest difference in estimates between IDSs and the NBP/CSO survey can be observed for Kraków (from 1.5 to 7.6), Olsztyn (from 1.5 to 11.5) and Szczecin (from 1.4 to 11.9). These results are consistent with those presented for the distribution of the number of rooms.

The second measure used to determine the significance of differences between distributions is presented in Table 4.24. As was the case with the distribution of the number of rooms, Table 4.24 presents aggregated information about the number of times the null hypothesis was rejected (two distributions are different) and the number of times it was not rejected (two distributions are similar). In 26 cases (18%) the distributions of data from OtoDom.pl and the NBP/CSO survey turn out to be similar. This is the case for selected quarters for Olsztyn (8 of 12 quarters), Katowice (4 out of 12 quarters), Łódź (4 out of 12 quarters), Lublin (4 out of 12 quarters) and three other cities. In the case of Nieruchomosci-online.pl, H0 hypothesis was rejected in over 90% of cross-classifications, which indicates that distributions observed at this portal are significantly different from those found in the NBP/CSO survey.

Table 4.24 also presents results for 12 cities. For Gdańsk, Opole, Poznań and Wrocław all IDS-based estimates are significantly different from NBP/CSO-based estimates. Previously, in the case of the number of rooms, the only city for which distributions in all quarters were different was for Wrocław. The highest degree of consistency with official statistics can be observed for Olsztyn, Lublin and Katowice. Results for other cities indicate systematic differences in floor area distributions.

4.4.2 A comparison with register data on transactions

Analysis of the distributions of the number of rooms and floor area has revealed systematic bias the IDSs. The above analysis was related to the distribution of $\boldsymbol{x}$ variables and representativeness in the sense described by Bethlehem (2009). The next step is to determine whether the selection mechanism involved is of the MNAR type, i.e. whether it related to the target variable. In order to check this property of the selection mechanism, an independent (from IDS) data source will be used. In this case the Register of Transactions is the obvious choice.

The Register of Transactions records all transactions made in the primary and secondary market. Hence, it should contain information about properties presented both online and offline. However, the register population is limited to fully-owned properties. Taking the above into account, the population of interest has been restricted to: (1) residential properties; (2) properties offered in the secondary market; (3) fully-owned property. Nieruchomosci-online.pl was the only portal to provide unit-level data, which is why this data source is used for comparison. In addition, owing to the limited access to the Register of Transactions, the following comparison is only made for Poznań.

Dittmann (2013) and Trojanek (2008) note that in Poland the offer price is, on average, higher than the transaction price, both in the primary and secondary market. This difference is natural and differences in prices are mainly due to negotiations. However, it should, on average, be smaller for cheaper flats, given the smaller price range within which to negotiate, and bigger for expensive flats. Previous research of the Polish real estate market indicates that the average percentage difference between the offer and transaction price per m^2 in Poznań is between 10% to 15% (Dittmann, 2013; Trojanek, 2008).

TABLE 4.25: The distribution of the average and median price per ^{2} on Nieruchomosci-online.pl and in the Register of Transactions and the index calculated for the prices in the period 2012Q1-2014Q3 in Poznań

	Average price m^2			Median price m^2		
Quarter	N-online (N)	Register (R)	R/N	Register (R)	N-online (N)	R/N
2012Q1	5 177	5 036	103	4 915	5 017	102
2012Q2	5 247	4 806	109	4 722	5 186	110
2012Q3	5 186	4 852	107	4 689	5 003	107
2012Q4	5 210	4 678	111	4 401	5 109	116
2013Q1	5 203	4 557	114	4 423	5 060	114
2013Q2	5 368	4 808	112	4 704	5 156	110
2013Q3	5 296	4 970	107	4 804	5 085	106
2013Q4	5 252	4 782	110	4 811	5 118	106
2014Q1	5 452	4 921	111	4 777	5 245	110
2014Q2	5 371	4 897	110	4 773	5 213	109
2014Q3	5 426	4 900	111	4 918	5 240	107

Note: N-online stands for Nieruchomosci-online.pl. N/R = Nieruchomosci-online.pl / Register and is given in %.

Table 4.25 shows the average and median price per m^2 on Nieruchomosci-online.pl and in the Register of Transactions and the index comparing the offer and transaction price between 2012Q1 and 2014Q3. Between 2012Q1 and 2013Q1 the average and median price per m^2 in the Register of transactions is on the decrease. The average price fell from 5 036 PLN/m^2 in 2012Q1 to 4 557 PLN/m^2 in 2013Q1; the median price declined from 4 915 PLN/m^2 in 2012Q1 to 4 401 PLN/m^2 in 2012Q4. Starting from 2013 the trend steadily picked up, reaching the average price of 4 900 PLN/m^2 and the median price of 4 918 PLN/m^2 in the last quarter. No such trend can be observed in the same period for prices on Nieruchomosci-online.pl. Offer prices steadily increased from the average of 5 177 PLN/m^2 in 2012Q1 to 5 426 PLN/m^2 in 2014Q3. The median price also steadily rose from 5 017 PLN/m^2 to 5 240 PLN/m^2 in the last quarter.

The offer-to-transaction price index (both for the average and the median) increased between 2012Q4 and 2013Q2, which reflects the decrease in the average and median transaction price. The average index between the median and mean price for all quarters is equal to 109%, which indicates that the offer price per m^2 is, on average, higher than transaction by 9%. The percentage difference between the average offer price per m^2 and the transaction price is higher than for the corresponding difference for the median price.

Given the differences between the offer and transaction price per m^2, a more detailed comparison between distributions can be made. This time the index is calculated separately for the actual price and the average price per m^2 across percentiles from 1st to 99th. In addition, to find any differences in price distributions between IDS and register data indexes were calculated separately for properties grouped by the number of rooms (four categories, the same as in the NBP/CSO survey) and floor area (four categories, the same as in the NBP/CSO survey).

Table 4.26 shows descriptive statistics for the offer-to-transaction price index (the actual price and the average price per m^2) calculated for each percentile. Table 4.26 is divided into three sections. The first section shows the overall index (offer / transaction price) for all percentiles, the second section contains indexes calculated for properties grouped by the number of rooms (1, 2, 3 and 4+), and the third section shows results broken down by floor area (up to 40 m^2, (40,60] m^2, (60,80] m^2 and over 80 m^2). In each sections the indexes for the actual price and the average price per m^2 are presented separately. The offer-to-transaction price index was calculated using the following formula:

$$\tau_{IDS,REG} = \frac{Q_{y,IDS,\alpha}}{Q_{y,REG,\alpha}}, \tag{4.3}$$

where $Q_{y,IDS,\alpha}$ denotes a quantile (here percentile) for variable y based on IDS, α denotes the probability level with values between $[0, 1]$ and $Q_{y,REG,\alpha}$ denotes a quantile for variable y based on the register data.

$\tau_{IDS,REG}$ indexes calculated jointly for all percentiles indicate considerable differences between offer and transaction prices. Median $\tau_{IDS,REG}$ for the average price per m^2 is equal to 110.2% and is consistent with the results presented in Table 4.25. However, $\tau_{IDS,REG}$ ranges from 105.4% to 170.1%, which means that for certain percentiles of properties on Nieruchomosci-online.pl ($Q_{y,IDS,\alpha}$) the average price per m^2 was over 70% higher than the corresponding price in the

TABLE 4.26: Descriptive statistics for the offer-to-transaction price index in Poznań between 2012Q1 and 2014Q4

Variable	Section	Min	Q_1	Median	Mean	Q_3	Max
Overall $\tau_{IDS,REG}$ [in %]							
Actual price		117.7	119.6	121.3	123.7	124.3	166.9
Price per m^2		105.4	106.9	110.2	111.7	112.2	170.1
$\tau_{IDS,REG}$ *grouped by the number of rooms* [in %]							
Actual price	1	97.8	100.1	103.4	106.2	108.1	140.0
Actual price	2	115.4	121.5	137.8	139.0	146.7	251.9
Actual price	3	133.1	137.4	139.6	144.1	143.8	218.5
Actual price	4+	136.0	142.4	148.3	150.0	156.0	237.5
Price per m^2	1	72.3	97.0	97.5	96.5	98.2	103.9
Price per m^2	2	98.7	100.6	102.9	107.0	108.4	199.8
Price per m^2	3	101.6	102.7	105.3	107.5	107.0	180.3
Price per m^2	4+	105.8	108.4	109.6	111.3	112.0	141.6
$\tau_{IDS,REG}$ *grouped by floor area* [in %]							
Actual price	to 40	103.5	108.7	111.1	115.2	113.7	193.8
Actual price	(40,60]	105.9	108.3	110.0	112.4	113.9	169.8
Actual price	(60,80]	108.3	110.5	113.1	114.3	114.5	166.8
Actual price	over 80	101.0	123.4	125.0	127.9	128.4	221.5
Price per m^2	to 40	104.1	106.3	108.8	113.6	115.4	212.8
Price per m^2	(40,60]	103.3	105.3	108.3	110.1	112.5	168.0
Price per m^2	(60,80]	106.5	109.7	112.0	113.1	112.8	160.2
Price per m^2	over 80	92.3	118.3	119.3	121.8	123.5	180.8

Register ($Q_{y,REG,\alpha}$). $\tau_{IDS,REG}$ for the actual price is higher than that for the average price per m^2. Average $\tau_{IDS,REG}$ is equal to 123.7%, which indicates that, on average, actual offer prices are over 20% higher than transaction prices. The range of $\tau_{IDS,REG}$ for the actual price is narrower than that for the average price per m^2 and ranges from 117.7% to 166.9%.

There is more diversity when properties are grouped by the number of rooms and floor area. The smallest median $\tau_{IDS,REG}$ for the actual price can be seen for properties with only one room: it is equal to 103.4, which indicates that the difference between the offer and transaction price is small, only about 3% higher. However, for certain percentiles the relation is lower than 100, which means that the offer price is lower than the transaction price. In addition, there are percentiles where the relation is as much as was 140%. $\tau_{IDS,REG}$ increases with the increasing number of rooms. For properties with two rooms median $\tau_{IDS,REG}$ is equal to 137.8%, for three-room properties it is 139.6%, and for the last group it is the highest and equals 148.3%. This relationship between $\tau_{IDS,REG}$ for the actual price and the number of rooms seems plausible: one can expect that negotiated price reductions will be higher for bigger (and more expensive) properties. Moreover, $\tau_{IDS,REG}$ is over 200% (251.9% for two-room properties, 218.5% for three-room properties and 237.5% for properties with 4 or more rooms).

As for the average price per m^2, median $\tau_{IDS,REG}$ is substantially smaller

in comparison with that for the actual price. For properties with one room median $\tau_{IDS,REG}$ is equal to 97.5, which indicates that offer prices are lower than transaction prices. This could correspond to a situation when smallest properties are underpriced or demand for such properties is higher than for others. Median $\tau_{IDS,REG}$ is equal to 102.9% for two-room properties, 105.3% for three-room properties and 109.6% for biggest properties. The relationship between $\tau_{IDS,REG}$ and the number of rooms also seems in line with expectations.

The third section presented in Table 4.26 lists offer-to-transaction price indexes for properties grouped by floor area. In comparison to the breakdown by the number of rooms, $\tau_{IDS,REG}$ index is characterised by less variability. For instance, all values of $\tau_{IDS,REG}$ are grater than 100% for the actual price, which indicates that the offer price is higher than the transaction price. Median $\tau_{IDS,REG}$ for the actual price is equal to 111.1% for properties with a floor area of up to 40 m^2, in the second group it is at a similar level and equals 110.0%. Median $\tau_{IDS,REG}$ index for the last two groups is higher and equals 113.1% and 125.0% respectively.

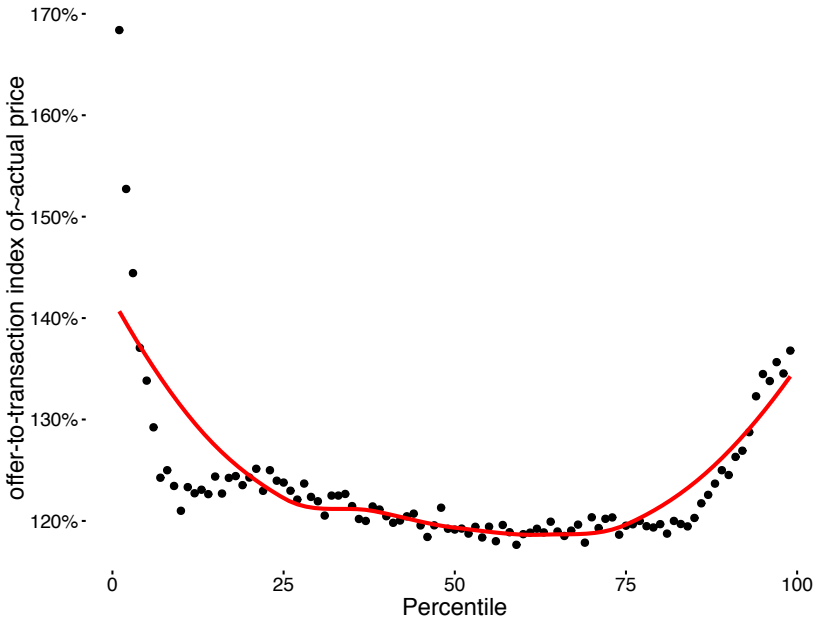


FIGURE 4.9: The distribution of the offer-to-transaction index for the actual price in Poznań between 2012Q1 and 2014Q3

A similar relationship can be observed with respect to $\tau_{IDS,REG}$ index for the average price per m^2. Median index comparing $Q_{y,IDS,\alpha}$ to $Q_{y,REG,\alpha}$ is equal to 108.8% and 108.3% for the first two groups, while for the other two - (60,80] and over 80 m^2 - it is equal to 112.0% and 119.3% respectively. The range of $\tau_{IDS,REG}$ for properties grouped by floor area is similar for both the actual price and the average price per m^2, which may imply that floor area does not affect the difference between the offer and transaction price.

Results presented in Table 4.26 do not show differences between percentiles or their corresponding indexes. Therefore, in order to specify what types of properties are not observed online plots comparing percentiles were prepared. Figure 4.9 presents the distribution of the offer-to-transaction price index for the actual price

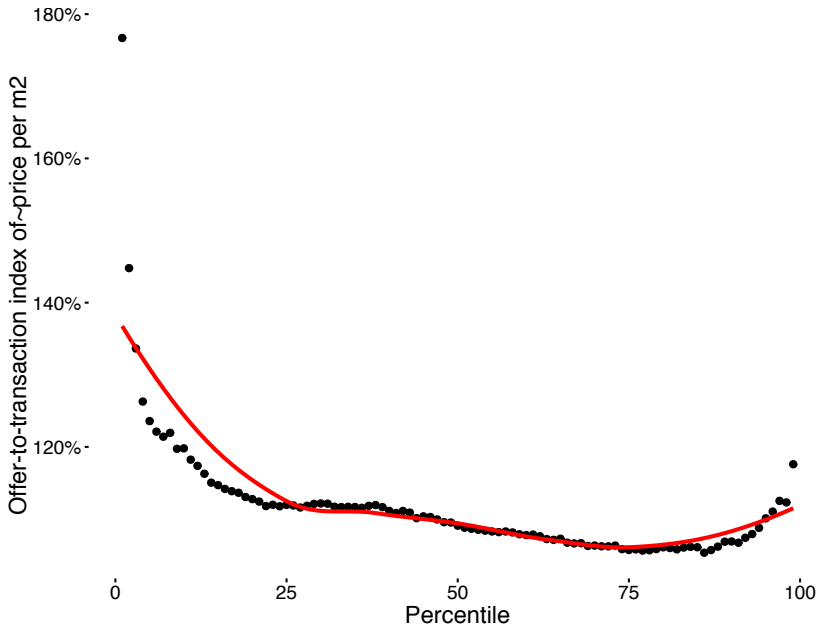


FIGURE 4.10: The distribution of the offer-to-transaction index for the average price per m^2 in Poznań between 2012Q1 and 2014Q3

for Poznań between 2012Q1 and 2014Q4. Each point refers to one percentile (from 1st to 99th) and the red line marks the estimated LOESS trend. Figure 4.9 shows a U-shaped distribution of the offer-to-transaction price index. It indicates that properties sold for a low price (lower than 130 000, 6th percentile of the actual price) were less likely to be presented online. $\tau_{IDS,REG}$ for 6th percentile and those below is significantly different from $\tau_{IDS,REG}$ values for other percentiles. In addition, certain groups of very expensive properties (over 350 000, 87th percentile) were also not published on Nieruchomosci-online.pl. $\tau_{IDS,REG}$ for 87th percentile or higher is different from other $\tau_{IDS,REG}$. The offer-to-transaction price index for 87th percentile or higher is over 120%, which is not what can be expected. For instance, a price reduction that can be obtained for more expensive property represents a smaller fraction of the actual price.

Figure 4.10 presents the distribution of the offer-to-transaction price index for the average price per m^2. This times the U-shape is slightly smoother in comparison with the figure for the actual price. On the whole, the distribution of the offer-to-transaction price index is in line with expectations, i.e. it decreases with increasing average price per m^2. However, what is more visible than in Figure 4.9 is the index distribution for low value properties. This means that certain cheap properties that were sold in Poznań between 2012Q1 and 2014Q3 were not advertised on Nieruchomosci-Online.pl. On the other hand, high value properties were more common on Nieruchomosci-online.pl than in the Register of Transactions. When it comes to the most expensive 10% of properties offered on Nieruchomosci-online.pl, their average price per m^2 is higher than the the price per calculated for the top 10% of properties found in the Register of Transaction. This relation might be due to the relatively longer period it takes to sell such properties and the fact that few buyers can afford them.

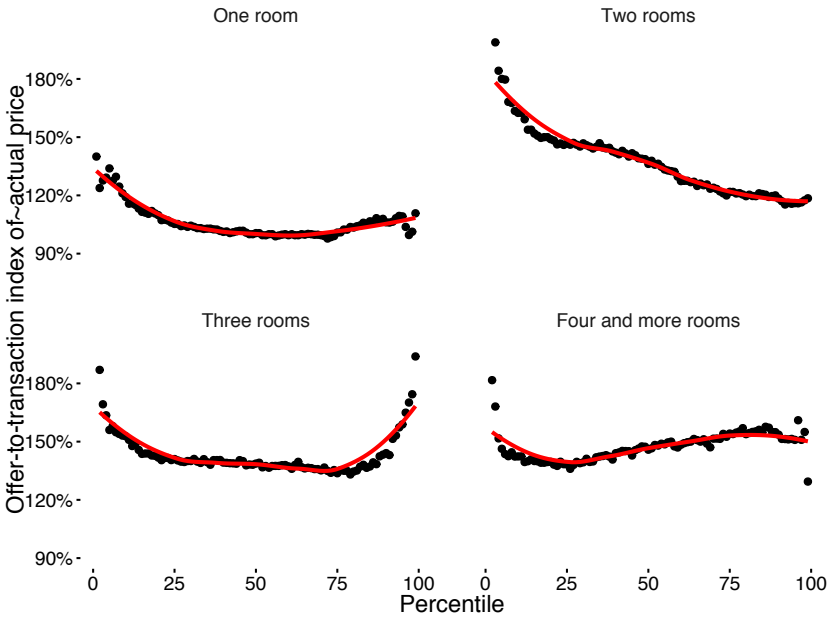


FIGURE 4.11: The distribution of the offer-to-transaction index for the actual price by the number of rooms in Poznań between 2012Q1 and 2014Q3

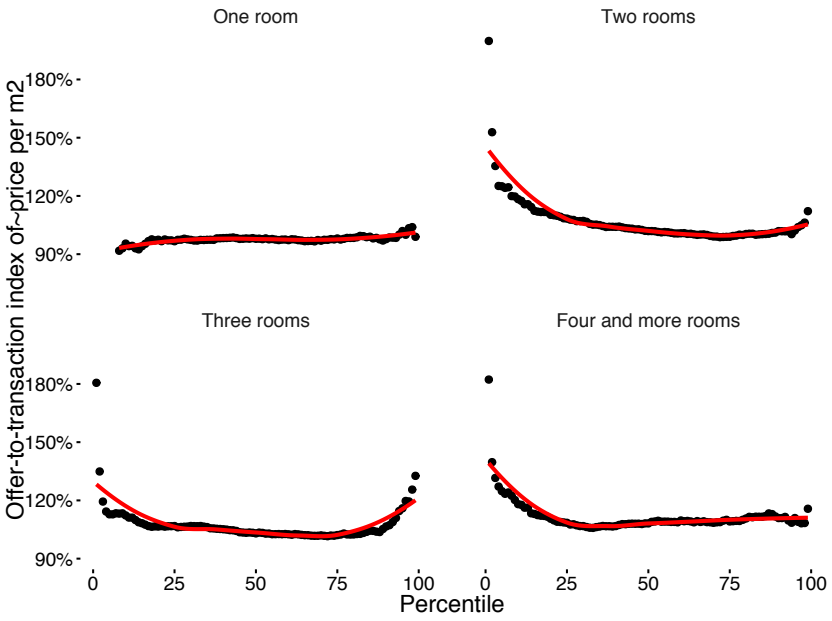


FIGURE 4.12: The distribution of the offer-to-transaction index for the average price per m^2 by the number of rooms in Poznań between 2012Q1 and 2014Q3

Figure 4.11 and Figure 4.12 present distributions of $\tau_{IDS,REG}$ index for properties grouped by the number of rooms. These distributions significantly differ between these categories of properties. The biggest differences are visible for properties with two and three rooms. In the case of properties with two rooms, the distribution of $\tau_{IDS,REG}$ decreases quite uniformly across percentiles, compared with more stable distributions in other groups of properties. The distribution observed

for properties with two rooms indicates that this group is most under-represented on Nieruchomosci-online.pl. Contrary to expactations, $\tau_{IDS,REG}$ index for cheapest properties is over 125%. The same kind of pattern can be observed for the average price per m^2, but the distribution is not as sharp.

The index for the group of properties with 3 rooms has a U-shape distribution. This means that low value properties are underrepresented online in comparison with the Register of Transactions, while high value properties are overrepresented. Underrepresentation of cheapest properties is also visible for 4+ category. In terms of $\tau_{IDS,REG}$ index only properties with one room are equally represented online and in the Register of Transactions.

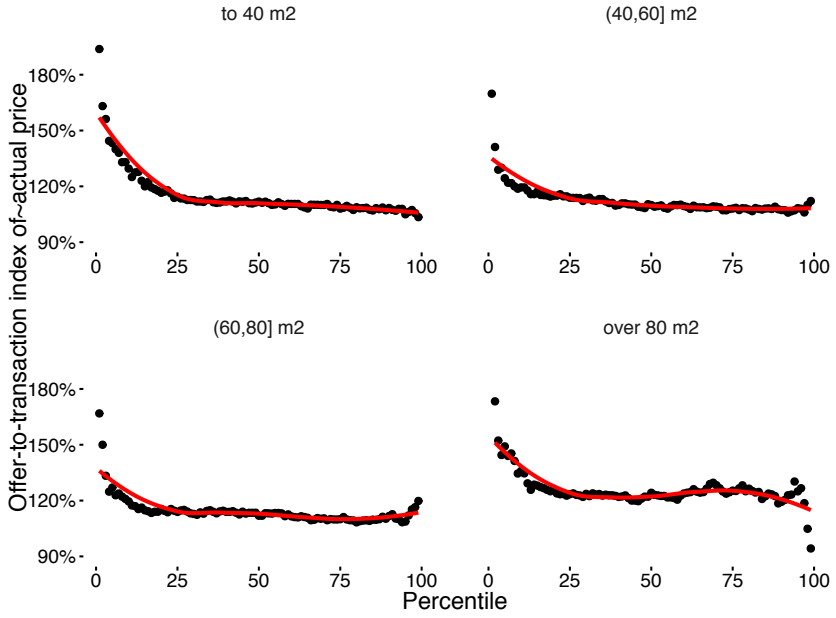


FIGURE 4.13: The distribution of the offer-to-transaction index for the actual price by floor area in Poznań between 2012Q1 and 2014Q3

Finally, Figure 4.13 and Figure 4.14 show the distribution of the offer-to-transaction index for actual price and average price per m^2 for properties grouped by floor area. Four categories are presented and the distributions of $\tau_{IDS,REG}$ index are substantially different from the previous comparisons. However, significant differences are visible for cheapest properties. In all categories of properties $\tau_{IDS,REG}$ is substantially different in the group of low value properties.

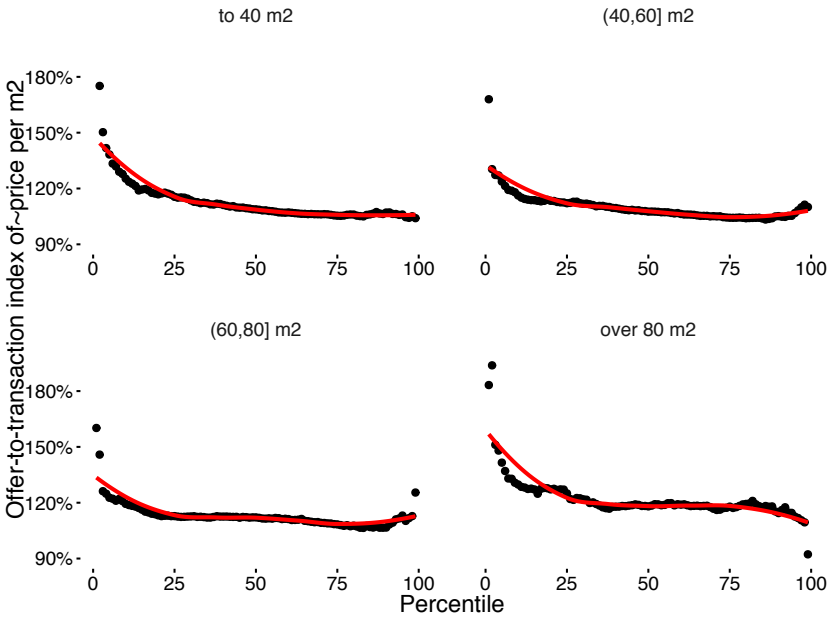


FIGURE 4.14: The distribution of the offer-to-transaction index for the average price per m^2 by floor area in Poznań between 2012Q1 and 2014Q3

4.5 Conclusions

The above chapter has dealt with the measurement of representativeness of Internet data sources about the secondary real estate market. The process of assessing representativeness was conducted following the procedure proposed in Chapter 3. The first step was the verification of the overall assessment of the usefulness of the Internet as a data source about the real estate market. This was based on the results of the ICT survey among companies classified as performing real estate activities (REA). Furthermore, data for 12 cities from three IDSs were analysed to compare the distribution of floor area and the number of rooms. Significant differences between IDSs were identified, particularly in terms of floor area. None of the IDSs was found to be fully consistent with the official survey conducted by NBP/CSO in terms of the distribution of the variables of interest: the number of rooms and floor area. The biggest differences were found for Wrocław and Opole, while the smallest for Olsztyn and Kraków. Despite these differences, estimates of the distribution of floor area and the number of rooms calculated for data from OtoDom.plwere found to be most consistent with the NBP/CSO survey. Finally, to determine whether the response model is of the MNAR type offer and transaction prices for Poznań were compared, (separately for the actual price and the price per m^2). The comparative analysis was conducted using data from Nieruchomosci-online.pl and the Register of Transactions. The results reveal substantial differences in the group of low priced properties, which are underprepresented on Nieruchomosci-online.pl. In addition, significant differences were also identified for properties with two rooms. Therefore, the overall assessment of representativeness of selected IDSs indicates that these data sources differ significantly from the NBP/CSO survey and the response model in Poznań is of the

MNAR type. Hence, the next step will consist in estimating bias in IDSs to determine whether the MNAR process can be observed for other cities and whether IDS-based estimates are significantly different from the NBP/CSO survey.

Chapter 5

Empirical assessment of bias in Internet data sources

5.1 Estimation of bias and its variance

Results presented in chapter 4 indicate that the selected IDS are not representative. There are substantial differences in distributions of floor area and the number of rooms between all IDSs and between cities. Analysis of the offer-to-transaction price index in Poznań indicates that the online presence of properties is related to the target variable. Hence, chapter 5 will focus on the estimation of bias for the average price per m^2, using data from the three IDSs. For this purpose, an extension of the approach described by Fosen and Zhang (2011) and Zhang (2012b) to estimate bias in register data will be proposed, which takes into account four random errors. The proposed approach will answer the following questions concerning bias observed in IDSs:

- Do IDS-based estimates significantly differ from estimates provided in official statistics (based on the NBP/CSO survey)?
- What proportion of bias can be explained by data source, city or time?
- Can the MNAR mechanism be observed not only for Poznań but also for other cities?

In order to answer these questions bias should first be estimated. In general, bias of θ can be expressed by equation (5.1) and MSE is defined by (5.2)

Definition 5.1. Bias of estimator $\hat{\theta}$ is the difference between its expectation $E(\hat{\theta})$ and the true value of characteristic y given by θ. This relation is expressed by the equation (5.1):

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta. \tag{5.1}$$

Definition 5.2. Mean square error (MSE) of $\hat{\theta}$ is the sum of $\hat{\theta}$ variance given by $V(\theta)$ and the square of $\hat{\theta}$ bias given by equation (5.1). MSE of $\hat{\theta}$ is defined by the following formula:

$$MSE(\hat{\theta}) = V(\hat{\theta}) + \Big(Bias(\hat{\theta})\Big)^2. \tag{5.2}$$

Definition 5.3. The coefficient of variation of $\hat{\theta}$ is the relation of the root square of $MSE(\hat{\theta})$ and estimator $\hat{\theta}$:

$$CV(\hat{\theta}) = \frac{\sqrt{MSE(\hat{\theta})}}{\hat{\theta}}. \tag{5.3}$$

In practice, θ is rarely known and should be estimated. In such cases θ is estimated based on sample surveys or register data. In chapter 3 we introduced the notation for three estimators of θ based on a sample survey given by $\hat{\theta}$, administrative sources $\tilde{\theta}$ and IDSs $\breve{\theta}$.

Now, let us assume that we are interested in estimating IDS-based $\breve{\theta}$ bias. Two cases should be considered. The first one is given by equation (5.4) and refers to the case when the true value θ is estimated based on a sample survey $\hat{\theta}$:

$$Bias(\breve{\theta}) = E(\breve{\theta}) - E(\hat{\theta}), \tag{5.4}$$

and for this case $MSE(\breve{\theta})$ is given by:

$$MSE(\breve{\theta}) = V(\breve{\theta}) + \Big(E(\breve{\theta}) - E(\hat{\theta})\Big)^2. \tag{5.5}$$

The second case is given by equation (5.6) and refers to the case when the true value θ is estimated using register data $\tilde{\theta}$ and this estimator is assumed to be unbiased or adjusted for bias

$$Bias(\breve{\theta}) = E(\breve{\theta}) - E(\tilde{\theta}), \tag{5.6}$$

and for this case $MSE(\breve{\theta})$ is given by:

$$MSE(\breve{\theta}) = V(\breve{\theta}) + \Big(E(\breve{\theta}) - E(\tilde{\theta})\Big)^2. \tag{5.7}$$

In the cases above we assume that estimates of θ based on register data ($\tilde{\theta}$) or a sample survey ($\hat{\theta}$) are unbiased, while IDS-based estimates ($\breve{\theta}$) are biased. In this setting we assume that the expectation of $\breve{\theta}$ will be given by:

$$E(\breve{\theta}) = \theta + b, \tag{5.8}$$

where θ is the true value and b is a constant that refers to bias and $\hat{\theta}$ will be given by:

$$E(\hat{\theta}) = \theta + e, e \sim N(0, \psi), \tag{5.9}$$

where ψ is known sampling variance. Moreover, for register-based estimates of θ we can assume two cases:

$$E(\tilde{\theta}) = \theta, \tag{5.10}$$

or

$$E(\tilde{\theta}) = \theta + \tilde{e}, \tilde{e} \sim N(0, \xi). \tag{5.11}$$

For the first case we assume that register-based estimates are unbiased and, owing to their character, do not contain an additional error component. In the second case it is assumed that initially $E(\tilde{\theta})$ was biased, but thanks to bias-adjustment methods it was reduced and, as a result, an extra error component $\tilde{e} \sim N(0, \xi)$ has been introduced, where ξ is known (sampling) variance resulting from the bias-adjustment procedure. Therefore, the estimator of $Bias(\breve{\theta})$ given by equation (5.4) has the form:

$$\widehat{Bias}(\breve{\theta}) = E(\breve{\theta}) - E(\hat{\theta}) = \theta + b - (\theta + e) = b + e, e \sim N(0, \psi) \tag{5.12}$$

and for (5.6) is given by:

$$\begin{aligned} &\widehat{Bias}(\breve{\theta}) = E(\breve{\theta}) - E(\tilde{\theta}) = \theta + b - \theta = b, \\ &\widehat{Bias}(\breve{\theta}) = E(\breve{\theta}) - E(\tilde{\theta}) = \theta + b - (\theta + e) = b + e, e \sim N(0, \xi). \end{aligned} \tag{5.13}$$

The approaches presented above refer to the case when the estimation of $Bias(\breve{\theta})$ is made at the level for which variance of $\hat{\theta}$ or bias-adjusted $\tilde{\theta}$ is low, that is estimates of θ are characterised by high precision. This precision can be estimated in terms of the coefficient of variation given by equation (5.3). Now let us assume that we are interested in estimating bias at domain level. Therefore, equation (5.4) will be given by the following equation:

$$Bias(\breve{\theta}_d) = E(\breve{\theta}_d) - \hat{\theta}_d, \tag{5.14}$$

where d refers to domain, $\hat{\theta}_d$ refers to unbiased estimate of θ_d. However, when estimation is made for domains that were not planned before conducting a survey, estimation of θ_d based on a sample survey is characterised by low precision (high CV). Therefore, direct estimation of $Bias(\breve{\theta}_d)$ is inappropriate and other approaches should be considered. One solution can be the application of a model-based approach known from small area estimation.

Fosen and Zhang (2011) and Zhang (2012a) proposed a model-based approach to estimate bias in a register-based survey at domain levels. This approach takes into account the fact that θ_d based on a sample survey is estimated with low precision (high CV), and direct estimation of bias can be unreliable. The approach proposed by Fosen and Zhang (2011) and Zhang (2012a) is based on the following assumptions:

- two data sources are available: a register which covers the target population and a separate sample survey of the same population,
- y is the target variable of interest (in Zhang (2012a) it was unemployment rate),
- θ denotes a target statistic of interest (for example the population mean of y),
- $\tilde{\theta}$ denotes a register-based estimator of θ, which is assumed to be biased,

- $\hat{\theta}$ denotes a survey-based estimator of θ, which is treated as a *gold standard* (reference point), hence it is assumed to be unbiased.

Fosen and Zhang (2011) and Zhang (2012a) applied small area estimation techniques to estimate the bias of $\tilde{\theta}$, which is defined by the following equation:

$$Bias(\tilde{\theta}_d) = \tilde{\theta}_d - \hat{\theta}_d, \tag{5.15}$$

where $\tilde{\theta}_d$ is an estimator of θ_d based on the register and $\hat{\theta}_d$ is an unbiased estimator of θ_d based on sample data in domain d. Assuming that $\tilde{\theta}_d = \theta_d + b_d$ is a biased register-based estimate of θ for d domain, $\hat{\theta}_d = \theta_d + e_i$ is an unbiased survey estimate of θ for d domain (b_d - the bias of $\tilde{\theta}_d$ and e_d - sampling error of $\hat{\theta}_d$) we get:

$$Bias(\tilde{\theta}_d) = \tilde{\theta}_d - \hat{\theta}_d = (\theta_d + b_d) - (\theta_d + e_i) = b_d + \epsilon_d, \tag{5.16}$$

where $\epsilon_d = -e_d$ and $e_d \sim N(0, \psi_d)$.

For model (5.16) Fosen and Zhang (2011) introduced a random-effect model of bias:

$$b_d = \beta + v_d, \tag{5.17}$$

where $E(v_d) = 0$ and $V(v_d) = \sigma_v^2$, which yields the following linear mixed model given by (5.16):

$$Bias(\tilde{\theta}_d) = \beta + v_d + \epsilon_d. \tag{5.18}$$

On fitting the model (5.18) we get:

$$\widehat{Bias}(\tilde{\theta}_d) = \hat{b}_d = \hat{\beta} + \hat{v}_d, \tag{5.19}$$

where $\hat{v}_d = \hat{\gamma}_d(Bias(\tilde{\theta}_d) - \hat{\beta})$, $\hat{\gamma}_d = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \tilde{\psi}_d)$ and $\tilde{\psi}_d$ is a smoothing function of ψ_d.

5.2 The proposed approach

5.2.1 Model specification

Initially, in their approach Fosen and Zhang (2011) and Zhang (2012a) considered only one data source (register) and multiple domains. Now, we would like to propose an extension of the model (5.18) to estimate bias in IDSs, taking into account multiple data sources and autocorrelation of bias in time. The main motivation to extend the model (5.18) is the character of IDSs data, collected by longitudinal observation and consisting of multiple overlapping data sources.

Taking into account the above setting, let $\breve{\theta}_{d,t}$ be the estimator of $\theta_{d,t}$ based on IDS and $\hat{\theta}_{d,t}$ be the unbiased estimator based on survey (in particular NBP/CSO survey) for domain $d \in \{1, ..., D\}$ and $t \in \{1, ..., T\}$. Then, direct estimator of $Bias(\breve{\theta}_{d,t})$ will be given by equation (5.20):

$$\widehat{Bias}(\breve{\theta}_{d,t}) = E(\breve{\theta}_{d,t}) - E(\hat{\theta}_{d,t}). \tag{5.20}$$

Due to $\Omega_{IDS,d,t} \cap \Omega_{IDS,d,t+1} = \emptyset$, which results in $cov(Bias(\breve{\theta}_{d,t}), Bias(\breve{\theta}_{d,t+1}) \neq 0$. The situation $\Omega_{IDS,d,t} \cup \Omega_{IDS,d,t+1} = \emptyset$ in the secondary real estate market is the result of the long process of selling properties. Some properties put up for sale in period t, will be also offered at time $t+1$. IDSs in the real estate market are organised in such a way that advertisements are displayed for periods of 15, 30 or 60 days. Moreover, outdated ads are often hard to identify and are not removed by the owner. As a result, it is necessary to take into account the autocorrelation of $Bias(\breve{\theta}_{d,t})$.

The second issue that should be considered are multiple overlapping data sources. Let $\breve{\theta}_{k,d,t}$ denote an estimator of $\theta_{d,t}$ based on IDS k. Because brokers and owners tend to use several advertising services we need to take into account that $cov(\breve{\theta}_{k,d,t}, \breve{\theta}_{k+1,d,t}) \neq 0$. Therefore, a direct estimator of $Bias(\breve{\theta}_{k,d,t})$ given by:

$$\widehat{Bias}(\breve{\theta}_{k,d,t}) = E(\breve{\theta}_{k,d,t}) - E(\hat{\theta}_{d,t}), \tag{5.21}$$

will be correlated between data sources $cov(Bias(\breve{\theta}_{k,d,t}), Bias(\breve{\theta}_{k+1,d,t})) \neq 0$.

The use of IDSs by brokers and owners will vary between cities and IDSs. This assumption means that the estimator of θ based on IDS k and IDS $k+1$ for the same domain d in the same period t can be different. This assumption leads to:

$$E(\breve{\theta}_{k,d,t}) \neq E(\breve{\theta}_{k+1,d,t}) \implies \widehat{Bias}(\breve{\theta}_{k,d,t}) \neq \widehat{Bias}(\breve{\theta}_{k+1,d,t}). \tag{5.22}$$

Finally, because each IDS is a non-probability sample and does not cover all units of the target population, random error should also be taken into account for IDSs. Therefore, the estimator of $\breve{\theta}_{k,d,t}$ can be expressed by the following equation:

$$\breve{\theta}_{k,d,t} = \theta_{dt} + b_{kdt} + \epsilon_{kdt}, \tag{5.23}$$

where θ_{dt} denotes the true value, b_{kdt} is source specific bias of θ_{dt} and $\epsilon_{kdt} \sim N(0, \phi_{kdt})$ denotes random error with known variance ϕ_{kdt}. Taking into account the characteristics of IDSs, the proposed approach is based on the following assumptions:

A1: an IDS provides information about apartments at the geographical domain level (here the city) at a given time or it can be aggregated,

A2: a *gold standard* exists for which direct estimates of some parameters and their sampling variances ψ_{dt} are available,

A3: an IDS is treated as a (big) sample of the target population, so there is a variance component ϕ_{kdt} which cannot be ignored a priori,

A4: owing to the massive character of Internet data sources it is assumed that the variance component is approximated under simple random sampling $V(\tilde{\theta}) = \phi = s^2/n$, where s is simple random sampling standard deviation of $\tilde{\theta}$,

A5: there are four random effects that effect bias – domain (city), AR(1) for domain, data source (web portal) and the interaction between domain and data source,

A6: the random effect for the data source is correlated, and this correlation matrix is known.

A7: we do not assume seasonality for y,

A8: we assume that random errors are correlated ϕ, ψ due to overlap between the survey sample and the IDS.

The estimation of variance is a crucial element of the proposed approach. If the survey is based on a complex sample, variance can be estimated in a straightforward manner, for instance using replication methods (in this case bootstrap). In this setting variance is calculated taking into account the sampling design applied in a given survey. Other approaches to the estimation of variance for sample surveys can be found in Wolter (2007). In addition, it should be noted that in the presence of small sample sizes estimation of variance may not be reliable, hence an smoothing function can be applied. For instance, variance is estimated under simple random sampling or using the generalized variance function.

On the other hand, in the case of IDSs, the sample mechanism is unknown and approaches known from the survey sampling theory may not hold. Given the massive character of Internet data, variance can be approximated under simple random sampling or the bootstrap approach. However, it should be noted that given the lack of knowledge about the selection mechanism, the above mentioned approaches may underestimate variance.

An overlap between the sample survey and IDSs leads to correlation between point estimates and their variances, in other words $\tilde{\theta}$ and $\hat{\theta}$ are not independent. In this case, the variance of the difference should be calculated, according to the equation (5.24) using the notation from the previous chapter:

$$\begin{aligned} V(\widehat{Bias}(\tilde{\theta})) =& V(\tilde{\theta} - \hat{\theta}) = V(\tilde{\theta}) + V(\hat{\theta}) - 2cov(\tilde{\theta}, \hat{\theta}) = \\ & V(\tilde{\theta}) + V(\hat{\theta}) - 2cor(\tilde{\theta}, \hat{\theta})\sqrt{V(\tilde{\theta})V(\hat{\theta})}, \end{aligned} \tag{5.24}$$

where cov denotes covariance and cor denotes the correlation coefficient. Assuming that $\tilde{\theta} = \theta + b + \epsilon, \epsilon \sim N(0, \psi)$ and $\hat{\theta} = \theta + e, e \sim N(0, \phi)$ equation (5.24) can be expressed as:

$$V(\widehat{Bias}(\tilde{\theta})) = \psi + \phi - 2cov(\tilde{\theta}, \hat{\theta}) = \psi + \phi - 2cor(\tilde{\theta}, \hat{\theta})\sqrt{\psi\phi}, \tag{5.25}$$

where ϕ is a known sampling error for the sample survey and ψ is a known random error for the IDS. The most problematic part is to estimate $cov(\tilde{\theta}, \hat{\theta})$ (see Alper and Berger, 2015; Berger and Priam, 2016; Qualité and Tillé, 2008; Smith et al., 2003; Tam, 1984; Wood, 2008). However, the approaches presented in the literature require access to unit-level data in order to compute $cov(\tilde{\theta}, \hat{\theta})$, which is not always possible in practice. Hence, to estimate $cov(\tilde{\theta}, \hat{\theta})$ the following approach is proposed:

1. calculate point estimates for each domain for each data source ($\tilde{\theta}_{k,d,t}$ and $\hat{\theta}_{d,t}$),

2. calculate differences for each domain and for each data source:

$$\delta_{k,d} = \tilde{\theta}_{k,d,t} - \tilde{\theta}_{k,d,t-1}, \tag{5.26}$$

and

$$\delta_{S,d} = \hat{\theta}_{d,t} - \hat{\theta}_{d,t-t}, \tag{5.27}$$

where S denotes sample-based estimates, k denotes k IDS.

3. assume that $cor(\tilde{\theta}_{k,d,t}, \hat{\theta}_{d,t}) = cor(\tilde{\theta}_{k,d}, \hat{\theta}_{d})$ and estimate $cor(\tilde{\theta}_{k,d}, \hat{\theta}_{d})$ using $\delta_{k,d}$ and $\delta_{S,d}$ according to the following formula:

$$cor(\tilde{\theta}_{k,d,t}, \hat{\theta}_{d,t}) = cor(\tilde{\theta}_{k,d}, \hat{\theta}_{d}) \approx cor(\delta_{k,d}, \delta_{S,d}). \tag{5.28}$$

5.2.2 Model description and estimation

The proposed approach is an extension of the model given by (5.16) and is given by the following formula:

$$Bias(\breve{\theta}_{k,d,t}) = \eta_{k,d,t} = \breve{\theta}_{k,d,t} - \hat{\theta}_{d,t} = b_{k,d,t} + \epsilon_{k,d,t}, \tag{5.29}$$

where $\breve{\theta}_{k,d,t}$ is an estimator of y based $k \in \{1, ..., K\}$ Internet data source for $d \in \{1, .., D\}$ domain in period $t \in \{1, ..., T\}$. Let M denote the number of observations given by $K \times D \times T$. Let $\hat{\theta}_{d,t}$ denote an estimator based on a sample survey for domain d in period t and let e_{kdt} denote known sampling variance for the survey and IDS, given by $\epsilon_{k,d,t} = e_{\phi,k,d,t} - e_{\psi,d,t}$, where $e_{\psi,d,t} \sim N(0, \psi_{d,t})$ is known sampling error and $e_{\psi,k,d,t} \sim N(0, \psi_{k,d,t})$ is known sampling error for IDS. Therefore:

$$\epsilon_{k,d,t} \sim N(0, \phi_{k,d,t} + \psi_{d,t} - 2cor(\delta_{k,d}, \delta_{S,d})\sqrt{\phi_{k,d,t}\psi_{d,t}}) = N(0, \omega_{k,d,t}), \tag{5.30}$$

where $cor(\delta_{k,d}, \delta_{S,d})$ was calculated according to (5.28). For the model given (5.29) we assume the linking model given by:

$$b_{k,d,t} = \boldsymbol{x}'_{k,d,t}\boldsymbol{\beta} + u_{1,d} + u_{2,d,t} + u_{3,k} + u_{4,d,k}, \tag{5.31}$$

for which we assume four random effects denoted by u with subscripts $\{1, 2, 3, 4\}$ and the fixed part denoted by $\boldsymbol{x}'_{k,d,t}\boldsymbol{\beta}$. By $\boldsymbol{x}'_{k,d,t}$ we denote vector of auxiliary variables for k, d, t observation and $\boldsymbol{\beta}$ is a vector of parameters. Random effects specified in the model (5.31) are defined as follows:

1. $u_{1,d} = (u_{1,1}, ..., u_{1,D})'$ – is random effect for domains (cities). We assume that this random effect has the following distribution $u_{1d} \sim N(0, \sigma_1^2)$;

2. $u_{2,d,t} = (u_{2,1,1}, ..., u_{2,D,T})'$ – random effect for domains (cities) for which we assume AR(1) structure given by $u_{2,d,t} = \tilde{\tilde{\rho}} u_{2,d,t-1} + \epsilon_{2,d,t}$ and $\epsilon_{2,d,t} \sim N(0, \sigma_2^2)$;
3. $u_{3,k} = (u_{3,1}, ..., u_{3,K})'$ – random effect for the data source, which is assumed to be correlated with known $\boldsymbol{R}_{K\times K}$ correlation matrix. Thus, $u_{3,k} \sim N(0, \sigma_3^2 \boldsymbol{R})$;
4. $u_{4,k,d} = (u_{4,1,1}, ..., u_{4,K,D})'$ – random effect for the interaction between the data source and the domain for which we assume $u_{4,k,d} \sim N(0, \sigma_4^2)$.

Consider the following vectors and matrices obtained by stacking the elements of the model in columns

$$\begin{aligned}
\boldsymbol{\eta} &= \underset{1\le k\le K}{col}(\underset{1\le d\le D}{col}(\underset{1\le t\le T}{col}(\eta_{k,d,t}))),\\
\boldsymbol{X} &= \underset{1\le k\le K}{col}(\underset{1\le d\le D}{col}(\underset{1\le t\le T}{col}(\boldsymbol{x}'_{k,d,t}))),\\
\boldsymbol{\epsilon} &= \underset{1\le k\le K}{col}(\underset{1\le d\le D}{col}(\underset{1\le t\le T}{col}(\epsilon_{k,d,t}))),\\
\boldsymbol{u}_1 &= \underset{1\le k\le K}{col}(\underset{1\le t\le T}{col}(u_{1,d})),\\
\boldsymbol{u}_2 &= \underset{1\le k\le K}{col}(\underset{1\le d\le D}{col}(\underset{1\le t\le T}{col}(u_{2,d,t}))),\\
\boldsymbol{u}_3 &= \underset{1\le d\le D}{col}(\underset{1\le t\le T}{col}(u_{3,k})),\\
\boldsymbol{u}_4 &= \underset{1\le t\le T}{col}(u_{4,k,d}).
\end{aligned} \tag{5.32}$$

To simplify, let $\boldsymbol{u} = (\boldsymbol{u_1}', \boldsymbol{u_2}', \boldsymbol{u_3}', \boldsymbol{u_4}')'$, and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3, \boldsymbol{Z}_4)$, where $\boldsymbol{Z}_1 = \mathbf{1}_K \otimes I_D \otimes \mathbf{1}_T$, $\boldsymbol{Z}_2 = \mathbf{1}_K \otimes \boldsymbol{I}_{DT}$, $\boldsymbol{Z}_3 = \mathbf{1}_{DT} \otimes \boldsymbol{I}_K$, $\boldsymbol{Z}_4 = \boldsymbol{I}_{KD} \otimes \mathbf{1}_T$, where $\mathbf{1}$ denotes a vector of a given size (for instance $\mathbf{1}_T$ is a vector of size T consisting only of 1) and $\boldsymbol{I}$ denotes an identity matrix of a given size (for instance $\boldsymbol{I}_K$ is an identity matrix of $K \times K$ dimensions). A small area model of $Bias(\breve{\boldsymbol{\theta}})$ can be expressed as a linear mixed model given by:

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}. \tag{5.33}$$

Let $\boldsymbol{\Delta} = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \tilde{\rho})'$ be a vector of unknown parameters involved in the covariance structure of the model. Let $\boldsymbol{\epsilon} \sim N(\mathbf{0}_M, \boldsymbol{V}_\epsilon)$, where $\mathbf{0}_M$ denotes a vector of zeros of size M, and $\boldsymbol{V}_\epsilon$ is a diagonal matrix:

$$\boldsymbol{V}_\epsilon = \underset{1\le k\le K}{diag}(\underset{1\le d\le D}{diag}(\underset{1\le t\le T}{diag}(\omega_{k,d,t}))), \tag{5.34}$$

where $\omega_{k,d,t}$ is given by (5.30), $\boldsymbol{u} \sim N(\mathbf{0}_m, \boldsymbol{V}_u(\boldsymbol{\Delta}))$ with a covariance matrix given by a block diagonal matrix:

$$\boldsymbol{V}_u(\boldsymbol{\Delta}) = diag\{\sigma_1^2 \boldsymbol{I}_{D\times D}, \sigma_2^2 \Omega(\tilde{\rho}), \sigma_3^2 \boldsymbol{R}_{K\times K}, \sigma_4^2 \boldsymbol{I}\}, \tag{5.35}$$

where

$$\Omega(\tilde{\rho}) = \underset{1\le k\le K}{diag}(\underset{1\le d\le D}{diag}(\Omega(\tilde{\rho})_d)), \tag{5.36}$$

or, alternatively, $\boldsymbol{I}_{KD} \otimes \Omega(\tilde{\rho})_d$ and $\Omega(\tilde{\rho})_d$ is given by:

$$\Omega(\tilde{\rho})_d = \frac{1}{1-\tilde{\rho}^2} \begin{pmatrix} 1 & \tilde{\rho} & \cdots & \tilde{\rho}^{T-1} \\ \tilde{\rho} & 1 & \cdots & \tilde{\rho}^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\rho}^{T-1} & \tilde{\rho}^{T-2} & \cdots & 1 \end{pmatrix}, \tag{5.37}$$

where $d \in \{1, .., D\}$. Thus, the covariance matrix of $\boldsymbol{\eta}$ is given by:

$$\boldsymbol{V}(\boldsymbol{\Delta}) = \boldsymbol{Z}\boldsymbol{V}_u(\boldsymbol{\Delta})\boldsymbol{Z}' + \boldsymbol{V}_\epsilon. \tag{5.38}$$

Then, we can derive $\boldsymbol{V}(\boldsymbol{\Delta})$ in the following way:

$$\boldsymbol{V}(\boldsymbol{\Delta}) = \boldsymbol{V}_\epsilon + \sigma_1^2 \boldsymbol{Z_1}\boldsymbol{Z_1}' + \sigma_2^2 \boldsymbol{\Omega}(\tilde{\rho}) + \sigma_3^2 \boldsymbol{Z_3}\boldsymbol{Z_3}' + \sigma_4^2 \boldsymbol{Z_4}\boldsymbol{Z_4}'. \tag{5.39}$$

Following Henderson (1975), the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is given by:

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\Delta}) = \{\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\Delta})\boldsymbol{X}\}^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{\Delta})\boldsymbol{\eta}. \tag{5.40}$$

The best linear unbiased predictor (BLUP) of $\boldsymbol{u}$ is given by:

$$\tilde{\boldsymbol{u}} = \boldsymbol{V}_u(\boldsymbol{\Delta})\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{\Delta})\{\boldsymbol{\eta} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\Delta})\}, \tag{5.41}$$

The second identity leads to BLUPs of $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3$ and $\boldsymbol{u}_4$, which are given respectively by:

$$\tilde{\boldsymbol{u}}_1 = \sigma_1^2 \boldsymbol{Z}_1'\boldsymbol{V}^{-1}(\boldsymbol{\Delta})\{\boldsymbol{\eta} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\Delta})\}, \tag{5.42}$$

$$\tilde{\boldsymbol{u}}_2 = \sigma_2^2 \Omega(\tilde{\rho})\boldsymbol{V}^{-1}(\boldsymbol{\Delta})\{\boldsymbol{\eta} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\Delta})\}, \tag{5.43}$$

$$\tilde{\boldsymbol{u}}_3 = \sigma_3^2 \boldsymbol{Z}_3'\boldsymbol{V}^{-1}(\boldsymbol{\Delta})\{\boldsymbol{\eta} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\Delta})\}, \tag{5.44}$$

$$\tilde{\boldsymbol{u}}_4 = \sigma_4^2 \boldsymbol{Z}_4'\boldsymbol{V}^{-1}(\boldsymbol{\Delta})\{\boldsymbol{\eta} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\Delta})\}. \tag{5.45}$$

Replacing $\boldsymbol{\Delta}$ in the previous equations with estimator $\hat{\boldsymbol{\Delta}}$ we obtain empirical BLUE (EBLUE) $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\Delta}})$ and empirical BLUPs (EBLUPs) of $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_4$ respectively:

$$\begin{aligned} \hat{\boldsymbol{u}}_1 &= \tilde{\boldsymbol{u}}_1(\hat{\boldsymbol{\Delta}}) = (\hat{u}_{1,1}, ..., \hat{u}_{1,D})', \\ \hat{\boldsymbol{u}}_2 &= \tilde{\boldsymbol{u}}_2(\hat{\boldsymbol{\Delta}}) = (\hat{u}_{1,1}, ..., \hat{u}_{1,K})', \\ \hat{\boldsymbol{u}}_3 &= \tilde{\boldsymbol{u}}_3(\hat{\boldsymbol{\Delta}}) = (\hat{u}_{3,1,1}, ..., \hat{u}_{3,K,1}, ..., \hat{u}_{3,K,D})', \\ \hat{\boldsymbol{u}}_4 &= \tilde{\boldsymbol{u}}_4(\hat{\boldsymbol{\Delta}}) = (\hat{u}_{4,d,t}, ..., \hat{u}_{4,d,T})'. \end{aligned} \tag{5.46}$$

The BLUP estimator of (5.31) is given by:

$$Bias_{BLUP}(\breve{\theta}_{k,d,t}) = \eta_{k,d,t} = \boldsymbol{x}'_{k,d,t}\boldsymbol{\beta} + u_{1,d} + u_{2,d,t} + u_{3,k} + u_{4,d,k}, \tag{5.47}$$

and the EBLUP estimator of (5.31) after estimating $\boldsymbol{\Delta}$ is:

$$\widetilde{Bias}(\breve{\theta}_{k,d,t}) = \widetilde{\eta}_{k,d,t} = \boldsymbol{x}'_{kdt}\hat{\boldsymbol{\beta}} + \hat{u}_{1,d} + \hat{u}_{2,d,t} + \hat{u}_{3,k} + \hat{u}_{4,k,d}. \tag{5.48}$$

To obtain estimates of $\boldsymbol{\Delta}$ for the model (5.31) the likelihood-based method will be used, in particular the restricted maximum likelihood (REML) approach. Equation 5.49 gives the form of REML estimation under log-likelihood functions:

$$\begin{aligned} \log l(\boldsymbol{\Delta}) = &-\frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\Delta})| - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\Delta})\boldsymbol{X}| \\ &-\frac{1}{2}\log \boldsymbol{r}'(\boldsymbol{\Delta})\boldsymbol{V}^{-1}(\boldsymbol{\Delta})\boldsymbol{r}(\boldsymbol{\Delta}) - \frac{M-p}{2}\log(2\pi), \end{aligned} \tag{5.49}$$

where

$$\boldsymbol{r}(\boldsymbol{\Delta}) = \boldsymbol{\eta} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\Delta})\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\Delta})^{-1}\boldsymbol{\eta}, \tag{5.50}$$

and M denotes the number of observations, p denotes the rank of $\boldsymbol{X}$. To obtain estimates of $\boldsymbol{\Delta}$ minimization of $-2\log l(\boldsymbol{\Delta})$ was applied.

The model was estimated with Bound Optimization BY Quadratic Approximation (BOBYQA) derivative-free optimizer proposed by Powell (2009), which is implemented in *bobyqa* function from *minqa* package (Bates et al., 2014) in **R** via *metafor* (Viechtbauer, 2010) package.

5.2.3 Limitations of the model

The proposed model is the first such approach to estimating bias in IDSs and is aimed at finding whether these sources significantly differ from official sources. The model assumes that there are two sources of variation that should be taken into accoun. First, an IDS is treated as a big non-probability sample; therefore variance of estimators based on these sources should also be calculated. The second source of variation are survey-based estimates. Owing to the character of IDSs data are commonly available at domains for which variance of sample-based estimators is high. Therefore, the approach takes into account these two sources to provide estimates of bias.

Moreover, in the proposed approach four random effects are assumed that answer specific questions and enable independent quantification of bias and its decomposition into four sources. This decomposition is crucial for determining whether bias is of the MNAR type. The model makes it possible to analyse the relation between the distribution of the domain random effect and distributions of the target variable. In particular, the proposed model is formulated to answer the following research questions:

- What is overall bias in IDSs taking into account uncertainty?
- What is the autocorrelation of bias in IDSs?
- What is the level of bias in domains (cities) and IDSs?

- Is bias similar across domains and IDSs?

The first question can be answered by analysing the intercept of the model. This parameter helps to identify overall bias in IDSs taking into account the variance of IDSs and sample-based estimators. The intercept of the proposed model has been estimated taking into account random effects and can be interpreted as overall IDS bias.

The second question addresses the specific problem concerning systematic bias observed in IDSs and its trend. A high value of the autocorrelation coefficient $\tilde{\rho}$ indicates that bias is highly correlated over time. Moreover, $\tilde{\rho}$ can also be considered as an proxy of the overlap between IDSs in period t and $t+1$. This interpretation is related to the character of IDSs, particularly those concerning the real estate market. Properties for sale are advertised and these ads are updated by algorithms on a daily or weekly basis.

The proposed model quantifies the influence of the data source and domain on bias. Therefore, it can be used to check whether IDSs significantly differ between one another. The interpretation of the IDS random effect should be connected with the character and popularity of IDS. For instance, advertisement services may specialize in a certain group of properties. The domain (city) random effect, as was previously stated, can be used to check whether the missing data mechanism is of the MNAR type. The relation between estimated random effects across domains can be used to quantify overall bias observed in domains.

Nonetheless, the model also has limitations, which will now be discussed. The model assumes that bias is only observed in IDSs while in practice, official statistics are not error-free. For instance, surveys tend to have high non-response bias, which can be connected with the target variable. Hence, informative missingness in the target variable is observed (MNAR) and reweighting procedures do not fully account for that error. To find out whether official statistics are biased, access to unit-level data is required, which is unlikely in practice.

The second objection that can be raised against the model is connected with the variance estimation process. In the proposed approach, owing to the limited access to data, simple random sampling was assumed for IDSs. However, under this assumption IDS variance can be underestimated because the probability of inclusion can vary between units. Moreover, to precisely estimate variance the self-selection mechanism should be studied. Nonetheless, bias in estimated IDS variance under simple random sampling can decrease with increasing sample size of the IDS. In addition, the approximation of $cor(\tilde{\theta}_{k,d,t}, \hat{\theta}_{d,t})$ can also introduce additional bias due to $cor(\delta_{k,d}, \delta_{S,d})$. A simulation study should be conducted to determine how much bias is introduced by $cor(\delta_{k,d}, \delta_{S,d})$.

Last but not least, it should be noted that IDSs contain objects that should first be transformed into statistical units. Identification of units can also introduce additional bias, particularly when probabilistic record linkage or statistical matching is applied. For instance, Samart (2011) studied mixed-models based on probability-linked data sources. In the case when only domain data are available it can be difficult to determine the true number of statistical units (ads with connections many-to-one).

Finally, the form of the model can be criticized. For instance, it assumes additive bias, while in the case of multiplicative base, the model may not hold. Some

examples are given by Lohr and Brick (2012) for dual survey estimation. Another assumption that may require modification is that bias is stationary (AR(1) process), while, in fact, it may be nonstationary, or a random intercept for the trend may be required. Possible solutions can be found in the literature on small area estimation. For instance Fay and Diallo (2012) proposes a time-area model under nonstationarity, Pfeffermann and Burck (1990) and Hobza and Morales (2012) discuss random coefficient area and time-area models. Another reservation may be connected with the way the fixed correlation matrix between data sources is specified. However, such an approach would lead to an over-parametrisation of the model, which could partially be resolved by the Bayesian approach. For instance, Manzi et al. (2011, sec. 4.1) modelled a correlation structure under the scaled inverse Wishart model.

5.3 Results of bias estimation

5.3.1 Point estimates

Figure 5.1 shows point estimates of the average price per m^2 in 12 cities between 2012Q1 and 2014Q4. The Y scale has been limited to match the range of results for each city. Point estimates based on the NBP/CSO survey are published in an MS Excel file and are available on the web page of NBP[1].

Point estimates of the average price per m^2 for domain d in period t for Nieruchomosci-Online.pl were calculated according to equation (5.51), which a ratio of y to x. In (5.51) y refers to the asking price of property i and x is the total floor area of property i observed in period t in domain d, $n_{d,t}$ denotes sample size in a given domain d in period t

$$\breve{\theta}_{d,t} = \frac{\sum_{i=1}^{n_{d,t}} y_{d,t,i}}{\sum_{i=1}^{n_{d,t}} x_{d,t,i}}. \tag{5.51}$$

Data for OtoDom.pl and Dom.Gratka.pl were made available in aggregated form. Therefore, $\breve{\theta}_{d,t}$ for these two sources was calculated according to (5.52), which is the weighted average price per m^2:

$$\breve{\theta}_{d,t} = \frac{\sum_{c=1}^{C_{d,t}} r_{d,t,c} \times n_{d,t,c}}{\sum_{c=1}^{C_{d,t}} n_{d,t,c}}, \tag{5.52}$$

where c denotes the aggregation level and $C_{d,t}$ the number of aggregation levels prepared by owners of OtoDom.pl and Dom.Gratka.pl. For instance, OtoDom.pl prepared a data set, which for each domain (city) contains information aggregated by month, the number of rooms and floor area. Therefore, the aggregation level was defined as a cross-classification by month, the number of rooms and floor area. The number of aggregation $C_{d,t}$ levels varied between domains and quarters. For OtoDom.pl $C_{d,t}$ ranged from 21 to 213 with a mean of 113.6, and for Dom.Gratka.pl $C_{d,t}$ it ranged from 56 to 96 with a mean of 80. In equation (5.52)

[1] See `http://www.nbp.pl/homen.aspx?f=/en/publikacje/inne/real_estate_market_q.html`.

$n_{d,t,c}$ refers to the number of cases observed for each aggregation level. The difference in the way of calculating the average price per m^2 for these two sources was motivated by the lack of access to raw data. Hence, the calculation of the average price per m^2 in this manner may introduce additional bias due to measurement error since (5.51) is not equal to (5.52). The underlying point estimates for Figure 5.1 are presented in the appendix in Table A.1.

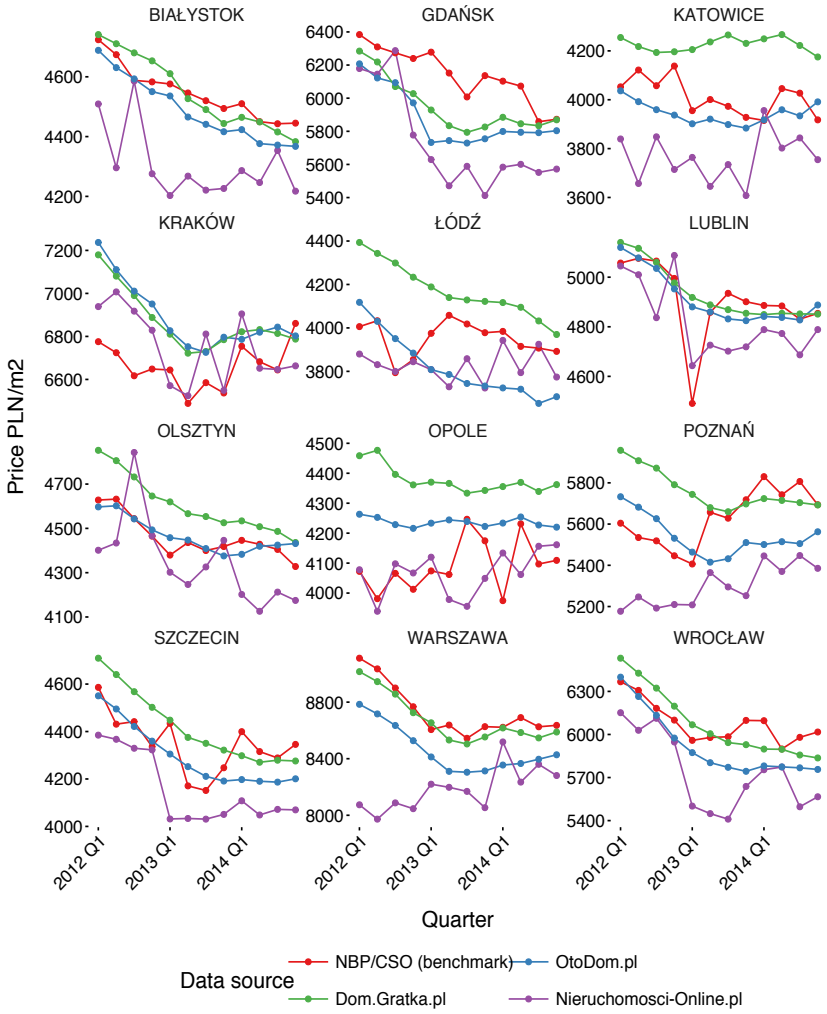


FIGURE 5.1: Comparison of average offer price of m^2 in 12 cities between 2012Q1 and 2014Q4 based on three IDS and NBP/CSO survey on the secondary market

Point estimates are presented in Figure 5.1. They reveal systematic differences between data sources and bias. However, the level of bias differs between the cities. For instance for Białystok a decrease in the average price per m^2 is the same in all the sources, but the NBP/CSO survey and Nieruchomosci-Online.pl systematically underestimate it. For Gdańsk and Warszawa all IDSs underestimate the average price per m^2, and for Nieruchomosci-Online.pl the mean bias is 500 PLN/m^2. For Warszawa the bias observed in Dom.Gratka.pl is the lowest. For Poznań after 2013Q1 there was an increase in the offer price per m^2 and a similar trend is visible in all IDS. Nonetheless, the change is not as high as in the

NBP/CSO survey. Data for all the cities from Dom.Gratka.pl and OtoDom.pl reveal similar trends, which are also smoothed due to aggregation.

An interesting pattern can be observed for Lublin. There is a substantial decrease in the NBP/CSO survey in 2013Q1; a similar fall can be seen in the data from Nieruchomosci-Online.pl. No such decline, however, is present in data from OtoDom.pl or Dom.Gratka.pl. This trend change may imply that an outlying observation was present both in the NBP/CSO survey and on Nieruchomosci-Online.pl. The relationship between these two sources indicates an overlap between the sample survey and the IDS, which influences the estimates.

Figure 5.2 presents direct estimates of $Bias(\breve{\theta}_{k,d,t})$, which is calculated according to equation (5.21). Mean absolute bias equals 183.07 with a standard deviation equal to 153.59, a minimum of 0.96, a maximum of 1062.6 and the median equal to 155.07. The analysis of bias indicates that the overall difference between the NBP/CSO survey and IDSs is small.

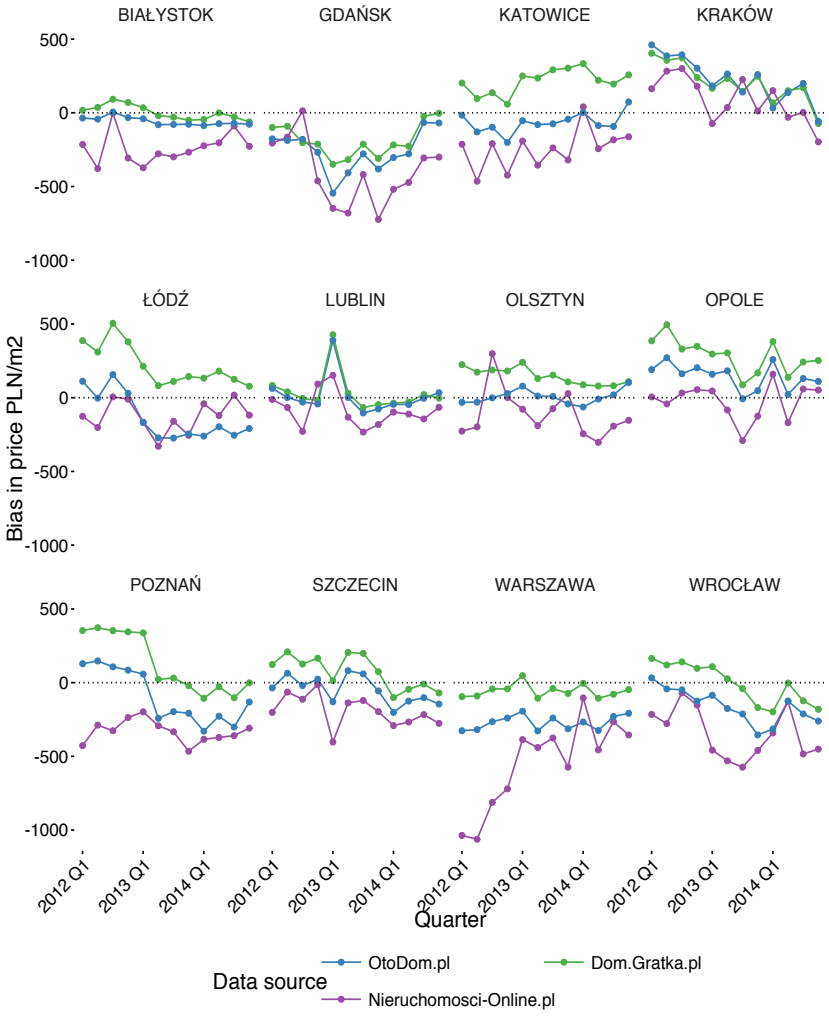


FIGURE 5.2: Comparison of bias in average offer price of m^2 in 12 cities between 2012Q1 and 2014Q4 based on three IDS and NBP/CSO survey on the secondary market

Figure 5.2 also shows a change in $\widehat{Bias}(\breve{\theta}_{k,d,t})$ over time. For instance, a similar level of bias can be seen for Warszawa and Gdańsk, particularly on OtoDom.pl. The difference between OtoDom.pl and the NBP/CSO survey is constant over time. However, in the case of Gdańsk from the begining of 2014 the bias decreases. The fall in the level of bias can be observed for Kraków. For Białystok, Olsztyn and Szczecin, all IDSs fluctuate about official statistics.

It should be noted that IDSs differ not only from official statistics but also between one another. For instance, for Łódź or Katowice Dom.Gratka.pl data overestimate, while data from OtoDom.pl underestimate the average offer price. However, for Białystok, Gdańsk, Kraków and Lublin differences between these two sources are not substantial. Pearson's correlation between $abs(\widehat{Bias}(\theta_{k,d,t}))$ and $\hat{\theta}_{d,t}$ is equal to 0.3, while Spearman's correlation coefficient is equal to 0.16. However, there is a correlation between IDSs. The highest correlation of $\widehat{Bias}(\theta_{k,d,t})$ is observed for OtoDom.pl and Dom.Gratka.pl. For instance, there is a constant difference between time series for these sources for Łódź, Opole, Poznań, Szczecin and Warszawa. In each of these cities, the average offer price on Dom.-Gratka.pl is higher than that on OtoDom.pl. Figure 5.3 shows a scatter plot of point estimates of bias for these data. The overall correlation of bias between the sources is high and exceeds 0.6. The highest correlation can be observed between Dom.Gratka.pl and OtoDom.pl (0.8) and the lowest between Dom.Gratka.pl and Nieruchomosci-Online.pl (0.6).

The explanation for the high correlation of $\widehat{Bias}(\theta_{k,d,t})$ between the data sources is straightforward. It is connected with brokers' and owners' activities online. Advertisements are placed on several online services either automatically (via special software) or manually. The level of $\widehat{Bias}(\theta_{k,d,t})$ correlation can be an proxy for the overlap between these sources. The high correlation between Dom.Gratka.pl and OtoDom.pl is due to their popularity as advertisement services for the real estate market. However, it should be noted that the correlation varies between the sources from 0.10 for Białystok according to Dom.Gratka.pl and Nieruchomosci-Online.pl to 0.99 for Kraków according to Dom.Gratka.pl and OtoDom.pl. The overall mean correlation between the sources for particular cities is equal to 0.69 and the median is equal to 0.69. Table A.4 shows correlation coefficients between the cities and sources.

The analysis of point estimates provides information about the relationship between IDSs and official statistics (the NBP/CSO survey). First of all, there is variation in the bias level between IDS and cities. Average bias for Dom.Gratka.pl was equal to 88.54 (standard deviation (SD) 174.36), for OtoDom.pl -58.62 (SD 229.32) and for Nieruchomosci-Online.pl -214.22 (SD 178.34). Median bias for Dom.Gratka.pl was equal to 82.48, for OtoDom.pl -54.84 and for Nieruchomosci-Online.pl -202.87. These statistics indicate that the highest systematic bias can be observed for Nieruchomosci-Online.pl, the less popular IDS (than the two other sources). On average OtoDom.pl underestimates the offer price in the secondary market, while Dom.Gratka.pl overestimates the level of the average offer price. The lower level of bias observed for these two data sources can be connected with the popularity of these sources as advertisement services.

However, it should be underlined that the absolute relative bias (ARB, absolute ratio of bias and the average offer price based on the NBP/CSO survey) given by

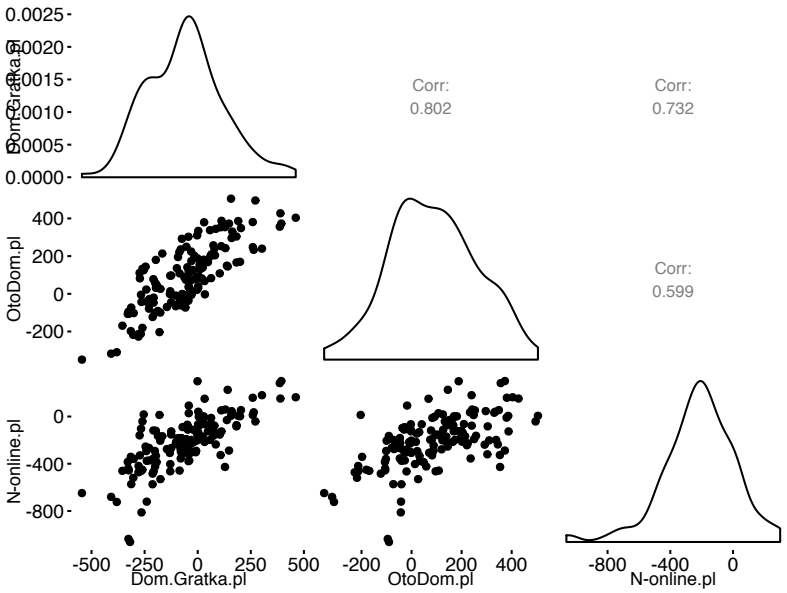


FIGURE 5.3: Scatterplot of $\widehat{Bias}(\tilde{\theta}_{k,d,t})$ 12 cities between 2012Q1 and 2014Q4 between three IDS on the secondary market

equation (4.2) for Dom.Gratka.pl was equal to 1.67 %, for OtoDom.pl 1.11 % and for Nieruchomosci-Online 4.04 %. A similar level of ARB can be observed between cities, where maximum ARB is equal to 4.66 % for Gdańsk and the minimum for Lublin is 0.28%. The correlation between bias varies from 0.2 to over 0.9, indicating a degree of overlap between the sources. The underlying statistics for the above analysis can be found in Table A.5 and Table A.6.

One crucial element of the model is the calculation of the variance of $Bias(\breve{\theta}_{d,t})$ for the NBP/CSO survey and the IDSs. Three components should be estimated, as indicated by equation (5.24) – variance for a given domain (city) and the period in the NBP/CSO survey, for each IDS and covariance between the sources. The last part is the most problematic and requires access to unit-level data and deterministic linkage between the sources. However, in practice this is unlikely due to (a) limited access to unit-level data from official sources or IDSs, (b) the lack of common identifiers between sources to enable deterministic linkage. In such cases, probabilistic linkage could be applied. In the present study access to unit-level data was available for Nieruchomosci-Online.pl. For the NBP/CSO survey, Dom.Gratka.pl and OtoDom.pl only aggregated data were available. Hence, to approximate the true variance that could be estimated using unit-level data, the following procedure was applied.

In the case of the NBP/CSO survey variance estimates of $\hat{\theta}_{d,t}$ are not published, so it should be estimated based on time series of direct estimates $\hat{\theta}_{d,t}$. Therefore, a simplified approach was adopted and is presented below. To estimate the variance of $\hat{\theta}_{d,t}$ based on the NBP/CSO survey a parametric bootstrap approach was applied, assuming AR(1) under a stationary model for each domain (city). The bootstrap steps are presented in Algorithm 2. All calculations were made in `boot` package and `tsboot` function (Canty and Ripley, 2015; Davison and Hinkley, 1997).

Data: NBP/CSO survey, D – the number of domains (cities, d), B - the number of bootstrap replications (b)
Result: Data frame with estimated variance for each time point for each domain
for $d \leftarrow 1$ **to** D **do**
  for $b \leftarrow 1$ **to** B **do**
    Calculate autoregressive order using function `ar`
    Return estimated order of $AR()$, mean of the time series and the time series for each doman;
  end
  Calculate variance for each time point
end

Algorithm 2: Pseudo-code for the algorithm for parametric bootstrap for $\hat{\theta}_{d,t}$

In the case of Nieruchomosci-Online.pl variance under simple random sampling was calculated. For Dom.Gratka.pl and OtoDom.pl variance under weighted simple random sampling was calculated. Weights were defined in the same way as in equation (5.52). Variance was calculated using `survey` package Lumley (2004, 2014). As was discussed ealier, this approach has several drawbacks. For instance, the assumption that simple random sampling was applied is not plausible, because owners decide independently whether or not to place an ad online. In addition, some owners, as was shown in Chapter 4, may choose not to advertise in IDSs (the cheapest and the most expensive ones). However, due to the lack of access to data and an unknown self-selection mechanism, simple random sampling is the most suitable option. Similar approaches can be found in the case of web surveys (Bethlehem, 2010).

Given the lack of access to unit-level data, direct estimation of covariance between estimates based on the NBP / CSO survey and a given IDS was not possible. Similarly, no information is available about the fraction of residential properties advertised online. Hence, an approximation was applied based on the correlation of first differences. The assumption underlying this approach is straightforward – if the overlap between the NBP/CSO survey and a given IDS is high, then first differences in a time series should be highly correlated; otherwise the correlation should be low. Moreover, correlation was only calculated at the level of IDSs, which resulted in only one correlation coefficient between the IDSs and the NBP/CSO survey for each domain. Figure 5.4 presents correlations between each IDS for each domain and period. The correlation between differences in the NBP/CSO survey and the IDSs are small: 0.26 for Dom.Gratka.pl, 0.2 for OtoDom.pl and 0.14 for Nieruchomosci-Online. However, it is noteworthy that first differences for OtoDom.pl and Dom.Gratka.pl are highly correlated (0.7).

Finally, $V(\widehat{Bias}(\tilde{\theta}_{k,d,t}))$ was calculated according to equation (5.30) assuming that covariance is constant in time and between domains. Table A.2 presents estimates of $V(\widehat{Bias}(\tilde{\theta}_{k,d,t}))$ for each data source and Table A.3 presents point

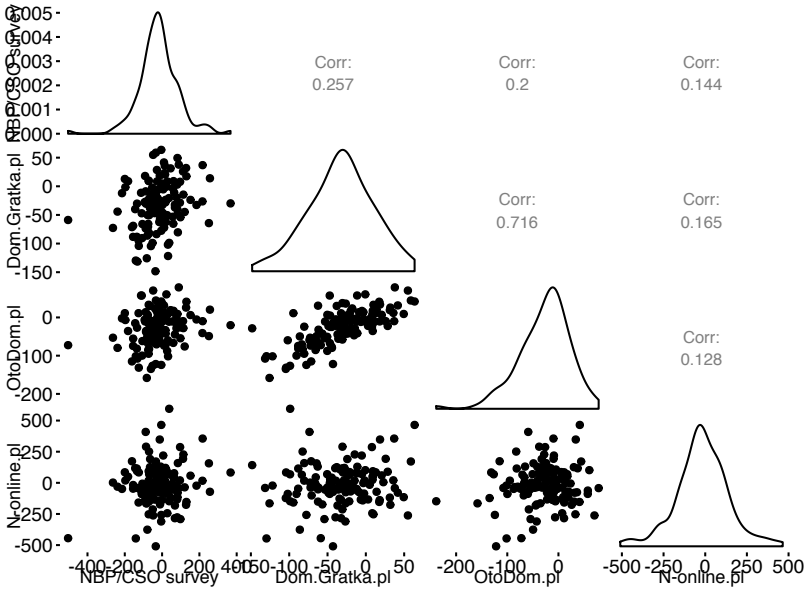


FIGURE 5.4: Correlation between $\delta_{k,d}$ and $\delta_{S,d}$ for all domains and all periods

estimates of $Bias(\breve{\theta}_{k,d,t})$ and its variance. However, the following possible reservations towards the proposed approach should be addressed. First of all, the assumption that covariance is constant in time and across periods is questionable. For instance, the coverage of the target population by an IDS is likely to depend on the overall use of the Internet by real estate brokers and agencies. The use of the Internet by companies involved in real estate market activities was presented in the first section of chapter 4. Another problem is that the use of first differences as a proxy for covariance might introduce additional bias. For example, residential properties observed in period t may also be observed in period $t+1$ in the NBP/CSO survey and it may be the same case for IDSs but for different apartments. Hence, if the price for these apartments changed in a similar way, or remained unchanged, the correlation as a proxy can be misleading.

Nonetheless, the approaches adopted in this study are highly dependent on the availability of data. For instance, approaches presented in Berger and Priam (2016), Qualité and Tillé (2008), Smith et al. (2003), Tam (1984), and Wood (2008) can only be applied if access to unit-level data is available.

5.3.2 Model results

Model specification

This section contains a description of the model and an interpretation of its results. The proposed model contains four random effects, which result in a complicated covariance structure. Figure 5.5 presents a visualisation of the covariance structure for each random effect. The data set used for analysis contained a total of 432 rows, so each matrix was of size 432×432. Black rectangles in Figure 5.5 represent the variance associated with a given random effect. For instance, in the top right graph the black rectangles refer to the random effect for the data source and

the dimensions correspond to 12 domains $\times$ 12 quarters, producing a 144×144 matrix (times 3 sources). The dimensions of the small black rectangles in the other graphs are associated with the number of quarters, namely 12×12.

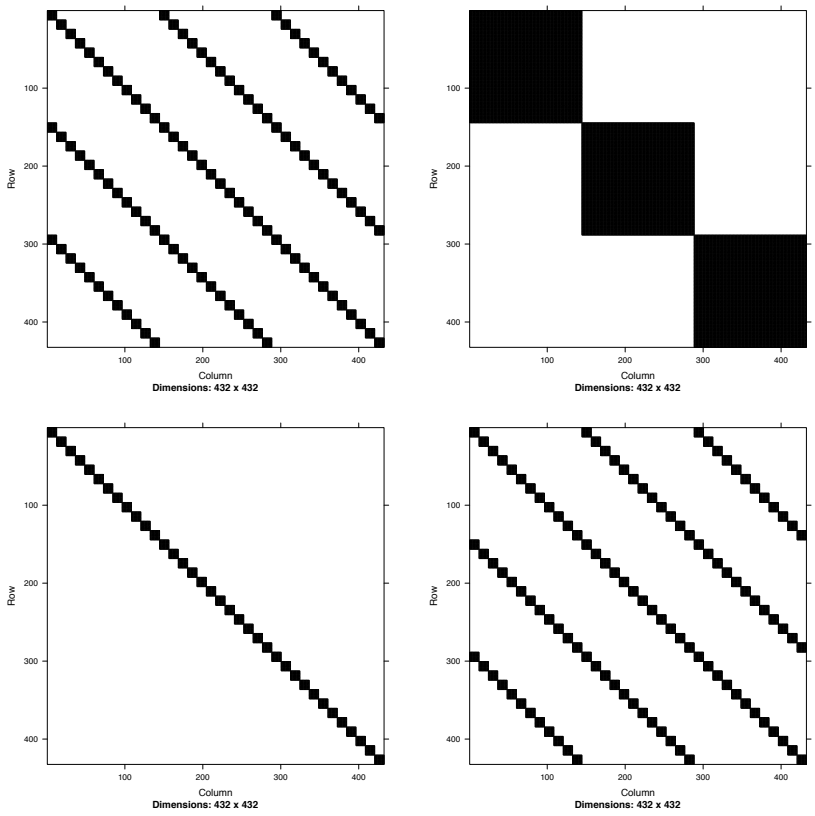


FIGURE 5.5: Visualisation of the covariance structure in the proposed model

Note: The left top graph represents the city random effect, right top graph represents the source random effect, left bottom graph represents the random effect of the interaction between city and source, the right bottom graph represents the AR(1) random effect for city.

The proposed model given by (5.47) includes a random effect for IDS and assumes that the levels of this random effect are correlated. Hence, the approximation of this correlation is presented in Figure 5.4. Pearson's correlation coefficient between estimates of $\theta_{d,t}$ based on OtoDom.pl and Dom.Gratka.pl is high (0.71), while for Nieruchomosci-Online.pl is assumed to be low (0.12, 0.16).

The following models were estimated to check whether the inclusion of random effects improves estimation.

- **Model 1**: only random effect for domain (city),
- **Model 2**: Model 1 and random effect for data source (IDS, correlated),
- **Model 3**: Model 2 and random effect for the interaction between domain and IDS,

- **Model 4**: Model 3 and AR(1) random effect for domain (city).

Table 5.1 contains a comparison of model statistics for the four estimated models. The following equations (5.53) define measures used for model selection. All models were estimated using REML.

$$Deviance = -2\times LogLik, \tag{5.53a}$$
$$AIC = -2\times LogLik + 2\times(p+1), \tag{5.53b}$$
$$BIC = -2\times LogLik + (p+1)\times \log(n), \tag{5.53c}$$
$$AICc = -2\times LogLik + 2\times(trace(H)+1). \tag{5.53d}$$

Column LRT shows the result of the Likelihood Ratio test (LRT) given by equation (5.54), where $-2\times LogLik(Null)$ is calculated for Model1, Model2 and Model3 (described above) separately, and $-2\times LogLik(\text{Model 4})$ refers to Model 4 (with all four random effects). Column pval shows the p-value and informs whether the hull hypothesis for LRT should be rejected. For each measure, the fit statistics for Model 4 are better than those for Model1, Model2 and Model3. LRT confirms that Model 4 is significantly better than any of the previous models. Hence, the proposed model is the most suitable to estimate bias in IDSs.

$$LRT = -2 \times LogLik(Null) - 2 \times LogLik(\text{Model 4}) \sim \chi^2(df-1). \tag{5.54}$$

TABLE 5.1: A comparison of the four estimated models

	LogLik	Dviance	AIC	BIC	AICc	LRT	pval
Model 1	-3121.5	6243.1	6247.1	6255.2	6247.1	826.8	$<.0001$
Model 2	-2852.0	5704.1	5710.1	5722.3	5710.1	287.8	$<.0001$
Model 3	-2816.7	5633.5	5641.5	5657.7	5641.6	217.2	$<.0001$
Model 4	-2708.2	5416.3	5428.3	5452.7	5428.5	-	-

Estimated parameters

Following the notation of the model given by (5.29) - (5.48), estimated parameters are listed below. Values in brackets for variance components refer to the fraction of total variance explained by all random effects. Analysis of the results of the estimated intercept and its standard error indicate that on average ($\hat{\beta}$) IDSs do not significantly differ from the officially published offer price per m^2 and the 95% confidence interval is equal to $(-371.69, 193.54)$. Hence, the hypothesis that IDSs do not differ significantly from the NBP/CSO was not rejected.

Moreover, the variance associated with the IDS effect is the highest and accounts for over 55% of all variance components. This means that the level of bias is mostly explained by differences between the IDSs of interest, that is Dom.Gratka.pl, OtoDom.pl and Nieruchomosci-Online.pl. Estimated variance for the data

source (IDS) effect $\hat{\sigma}_3^2$ can be due to the different popularity of the IDSs, which is manifested by owners' preferences to place ads on certain portals when offering properties in a certain price range; this results in differences in terms of the asking price and the average price per m^2. The random effect for Dom.Gratka.pl is equal to 173.69 PLN/m^2, for Nieruchomosci-Online.pl it is -123.22 PLN/m^2 and for OtoDom.pl 33.10 PLN/m^2. The results indicate that in terms of absolute bias, OtoDom.pl is characterised by the lowest bias, while Dom.Gratka.pl and Nieruchomosci-Online.p with the highest bias. It should be noted that this effect is independent of the other random effects.

- Intercept ($\hat{\beta}$): -89.07 (SE=144.19),
- Variance of city effect ($\hat{\sigma}_1^2$): 13074.18 (20.1%),
- Variance of AR(1) effect ($\hat{\sigma}_2^2$): 12107.87 (18.6%),
- Variance of IDS effect ($\hat{\sigma}_3^2$): 35975.03 (55.4%),
- Variance of city/IDS effect ($\hat{\sigma}_4^2$): 3804.09 (5.9%),
- Autocorrelation for AR(1) ($\tilde{\rho}$): 0.6919.

Next, other characteristics of the real estate market were investigated. Table 5.2 contains several characteristics for the 12 cities between 2012Q1 and 2014Q4. The following variables are included in Table 5.2 – the average offer price per m^2 from the NBP/CSO survey, the average transaction price per m^2 based on the Register of Transaction[2], the offer-to-transaction price index, the average number of transactions based on the Register of Transactions, the average number of companies classified into NACE rev. 2 Section L (Real estate market activities) and estimates of the city random effect denoted by $\hat{\boldsymbol{u}}_2$.

The offer price per m^2 is systematically overestimated in IDSs for 7 cities (Kraków, Opole, Olsztyn, Łódź, Lublin, Katowice and Szczecin), and it is underestimated for the remaining cities (Białystok, Poznań, Wrocław, Gdańsk and Warszawa). Kraków, Gdańsk and Warszawa are cities with the highest level of bias in the offer price per m^2 and are also characterised by the highest offer price according to the NBP/CSO survey. The only exception is Opole, but the average sample size in the NBP/CSO survey for this city was only 34 units. On average, the cities where the average price per m^2 is lower are characterised by smaller bias. Pearson's correlation coefficient between the absolute value of the random effect for city ($abs(\hat{\boldsymbol{u}}_2)$) and the average offer price per m^2 at domain level ($\bar{\hat{\theta}}_d$) equals 0.64 (p-value = 0.02), and between $\hat{\boldsymbol{u}}_2$ and $\bar{\hat{\theta}}_d$ equals to -0.51 (p-value = 0.09). Spearman's correlation coefficient is equal to 0.58 (p-value = 0.05) and -0.47 (p-value = 0.12) respectively. Th results imply a moderate correlation between $abs(\hat{\boldsymbol{u}}_2)$ and the average offer price per m^2.

Pearson's correlation coefficient for the absolute city random effect for bias ($abs(\hat{\boldsymbol{u}}_2)$) and the transaction price per m^2 is equal to 0.64 (p-value = 0.02),

[2]Information about transactions was taken from CSO publications connected *Real estate transactions*, see http://stat.gov.pl/obszary-tematyczne/infrastruktura-komunalna-nieruchomosci/nieruchomosci-budynki-infrastruktura-komunalna/obrot-nieruchomosciami-w-2014-r,4,12.html.

TABLE 5.2: A comparison of the random effect of bias

City	Average price per m^2					City random effects	
	Offer (O)	Transaction (T)	Index (O/T)	# trans	Section L	$\hat{\boldsymbol{u}}_2$	$\lvert\hat{\boldsymbol{u}}_2\rvert$
Warszawa	8733.6	7627.0	87.3	10405.7	20321.8	-179.8	179.8
Kraków	6663.4	6016.0	90.3	8893.7	7656.8	188.9	188.9
Wrocław	6080.7	5479.7	92.7	5050.7	9322.1	-84.7	84.7
Gdańsk	6140.6	5183.5	90.1	5907.0	6616.3	-163.9	163.9
Poznań	5632.2	4852.2	86.2	4178.3	4994.1	-29.8	29.8
Lublin	4902.8	4462.4	91.0	1613.7	1766.1	38.0	38.0
Białystok	4546.1	4234.5	93.1	1653.7	1216.3	-29.6	29.6
Olsztyn	4459.1	4064.1	91.1	1752.0	1821.9	53.9	53.9
Szczecin	4345.7	4088.9	94.0	2043.7	4824.5	2.5	2.5
Opole	4091.8	3717.0	90.8	846.3	1834.3	168.5	168.5
Łódź	3951.2	3483.5	88.2	5190.0	4692.2	45.2	45.2
Katowice	4011.0	3211.3	80.1	639.7	2921.2	21.1	21.1

Note: *Offer (O)* denotes the average offer price per m^2 based on the NBP/CSO survey between 2012Q4 and 2014Q4; *Transaction (T)* denotes the average transaction price per m^2 based on the Register of Transactions between 2012Q4 and 2014Q4; *Index (O/T)* denotes the ratio of the average offer to transaction price per m^2 between 2012Q4 and 2014Q4; *# trans* denotes the average number of transactions between 2012Q4 and 2014Q4, *Section L* is the average number of companies classified into section L (Real estate market activities, NACE rev. 2) based on the REGON register, $\hat{\boldsymbol{u}}_2$ is the estimated city random effect.

and for bias ($\hat{\boldsymbol{u}}_2$) -0.46 (p-value = 0.12), Spearman's correlation coefficient is equal to 0.56 (p-value = 0.06) and -0.48 (p-value = 0.12) respectively. Pearson's correlation of $abs(\hat{\boldsymbol{u}}_2)$ with the transaction price per m^2 lower than that for the offer price, however still significant (p-value = 0.0259). Next, let us analyse the correlation between the offer-to-transaction price index (O/T index). If the correlation between this measure proves significant, it will imply that bias is related to market liquidity and competition. Pearson's and Spearman's correlation coefficients between the O/T index and $abs(\hat{\boldsymbol{u}}_2)$ are equal to 0.06 (p-value = 0.85) and -0.18 (p-value = 0.59) respectively. For $\hat{\boldsymbol{u}}_2$ the correlation coefficients are lower and equal to 0.06 (p-value = 0.85) and 0.11 (p-value = 0.71). The estimated correlation coefficients indicate that there is no relation between the O/T index and bias (correlation coefficients are not significant). Results indicate that bias is not associated with liquidity in the secondary real estate market for the cities of interest.

The final step of the analysis concerns the relation between the market size defined as the average number of transactions and the number of companies classified into Section L. If the relation between the market size and random city effect $\hat{\boldsymbol{u}}_2$ is significant, it may imply that the bigger the market, the more difference there is between properties offered online and offline due to competition in the market. However, it should be noted that the market size depends on the situation in the market, particularly on the price levels. Pearson's and Spearman's correlation coefficients between the average transaction price per m^2 in the 12 cities are equal to 0.87 and 0.73 respectively. Therefore, Pearson's and Spearman's correlation coefficients between $abs(\hat{\boldsymbol{u}}_2)$ and the number of transaction are equal to 0.66 (p-value = 0.02) and 0.58 (p-value = 0.05), but after including information about the

transaction price per m^2, Pearson's and Spearman's partial correlation is equal to 0.27 (p-value = 0.42) and 0.30 (p-value = 0.36). The correlation with the number of companies classified into Section L was equal to 0.54 p-value = 0.07) and 0.51 (p-value = 0.09). After including information about the transaction price per m^2, Pearson's and Spearman's partial correlation equals to -0.06 (p-value = 0.86) and 0.21 (p-value = 0.53) respectively.

To sum up, based on the results and correlations between characteristics presented in Table 5.2 one can identify a moderate relation between the estimated city random effect, offer and transaction price per m^2. This relation indicates that bias is higher for those cities with high property prices. Indirectly, bias is associated with the market size.

The autocorrelation coefficient estimated for the AR(1) random effect indicates that there is systematic bias in the IDSs. The autocorrelation coefficient equals $\tilde{\rho} = 0.70$ and indicates strong autocorrelation. This level of autocorrelation can be explained by the long time to sell and the fraction of residential properties offered in quarter t and are also observed in quarter $t + 1$. Moreover, the autocorrelation coefficient can be also interpreted as an overlap between properties offered in period t and $t + 1$ and $\tilde{\rho} = 0.70$ indicates that the bias decreases over time ($\tilde{\rho} < 1$). For instance, bias of $\breve{\theta}$ for Gdańsk in 2013 was on average equal to -439.5492, while at the end of 2014 it fell to -124.2698. Another visible decline in bias can be seen for Warszawa according to Nieruchomosci-Online.pl. At the beginning, the average offer price per m^2 is underestimated by nearly 1000 PLN/m^2, while at the end the difference is down to almost 400 PLN/m^2.

Figure 5.6 presents the relationship between the direct estimator ($\widehat{Bias}(\breve{\theta}_{k,d,t})$) and the EBLUP estimator ($\widetilde{Bias}(\breve{\theta}_{kdt})$). The figure shows that estimates based on the proposed model are consistent and highly correlated with the direct estimator (Pearson's correlation coefficient is equal to 0.95). Most of the points lie on the red line denoting $y = x$ and both distributions are asymmetric. There are outliers in the left tail, which are related to properties in Warszawa and Gdańsk.

To complete the analysis of the estimated bias let us compare its values between data sources and cities. Figure 5.7 presents $\widetilde{Bias}(\breve{\theta}_{kdt})$ broken down by IDS and city. Red denotes the direct estimator and blue the EBLUP based on the proposed model. In all cases EBLUP is smoothed in comparison to the direct estimator. For instance, a peak observed for Lublin (which is an outlier in direct estimates based on the NBP/CSO survey) is smoothed by EBLUP for each IDS, which makes it more reliable. Another example is Warszawa and OtoDom.pl, where EBLUP indicates constant systematic bias.

According to Figure 5.7 data from OtoDom.pl for Białystok, Katowice, Lublin, Olsztyn and Szczecin provide almost unbiased estimates of the average offer price per m^2. However, for Gdańsk, Łódź, Poznań, Warszawa and Wrocław systematic bias can be observed: the average offer price per m^2 is underestimated. On the other hand, Dom.Gratka.pl provides unbiased estimates for Warszawa and Poznań. This may imply that characteristics of residential properties offered in these markets are similar to those found in the target population. Unlike OtoDom.pl, Dom.Gratka.pl overestimates prices for Katowice and Łódź, but the lack of access to unit-level data prevents a thorough investigation of the source of bias. The

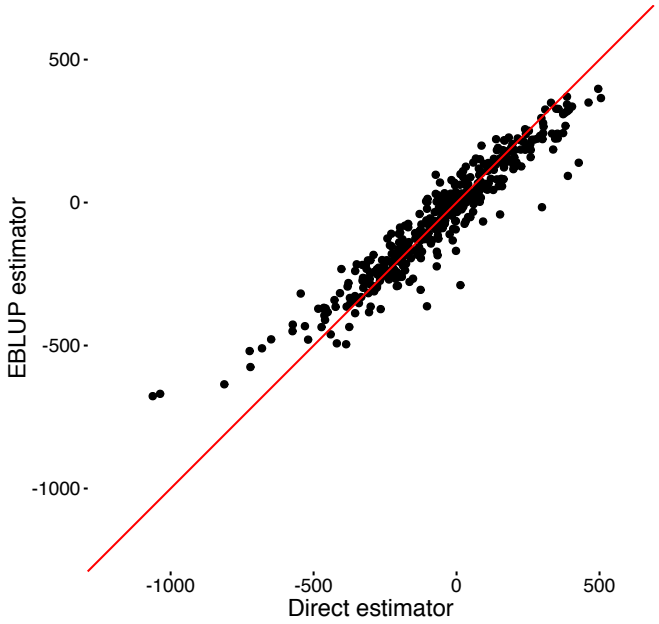


FIGURE 5.6: Comparison of direct estimator and EBLUP for $Bias(\breve{\theta}_{kdt})$

most unstable level of bias is observed for Nieruchomosci-Online.pl, where unit-level data were available. For instance, ads placed on Nieruchomosci-Online.pl about properties in the secondary market in Warszawa contained properties with lower prices than the population average.

TABLE 5.3: Descriptive statistics of model-based estimates of $Bias(\theta_{k,d,t})$ broken down by IDS

Statistic	Overall	OtoDom.pl	Dom.Gratka.pl	Nieruchomosci-Online.pl
Minimum	-677	-318.20	-217.80	-677.00
Q_1	-205.4	-195.70	-2.02	-331.20
Median	-56.13	-47.31	97.79	-215.30
Mean	-60.9	-58.30	88.60	-213.00
Q_3	75.91	21.44	189.70	-102.70
Maximum	397.7	349.80	397.70	223.40
RB [%]	-0.94	-0.97	2.10	-3.94
ARB [%]	3.11	2.31	2.80	4.21

Table 5.3 contains descriptive statistics of model-based estimated bias broken down by IDS. In addition, information about relative bias and absolute relative bias is presented. On average (for all domains and quarters) OtoDom.pl underestimated true average offer prices by -58.3 PLN/m^2, Dom.Gratka.pl overestimated by 88.6 PLN/m^2 and Nieruchomosci-Online.pl was characterised the highest bias of -213.0 PLN/m^2. The maximum bias observed in each data source was 349.8 PLN/m^2 for OtoDom.pl, Dom.Gratka.pl 397.7 PLN/m^2 and 223.4 PLN/m^2 for Nieruchomosci-Online.pl. However, it should be noted that relative absolute bias was low and equal to 2.31% for OtoDom.pl, 2.80% for Dom.Gratka.pl and 4.23

% for Nieruchomosci-Online.pl. Table A.7 presents similar statistics but including a breakdown by domain (city). For instance, the results indicate that relative bias was between -6.68 % to 6.96 % and on average -0.94%. On the other hand relative absolute bias was between 0.46% and 6.96% and on average 3.11%, but the median was lower and equal to 2.96%.

The results of the analysis can be summarized as follows:

- OtoDom.pl and Dom.Gratka.pl provide less biased point estimates of the average offer price per m^2, while data from Nieruchomosci-Online.pl are characterised by higher bias.
- The self-selection process in the residential real estate market is strongly associated with the target variable, hence it is informative.
- Bias is small for cities where the average offer price per m^2 is low, which means that the market is homogeneous.

Model diagnostics

In order to assess the model the following steps were taken. First, residuals based on the proposed model were analysed; second, the distribution of random effects was compared. The literature on small area models describes several methods to assess model performance (ch. 5.4 Rao and Molina, 2015). The basic diagnostics that should be calculated are conditional residuals given by equation (5.55), standardized conditional residuals given by equation (5.56) and leverage measures (hat values) given by equation (5.57):

$$r_{cond} = \widehat{Bias}(\breve{\theta}_{kdt}) - \widetilde{Bias}(\breve{\theta}_{kdt}), \tag{5.55}$$

where $\widehat{Bias}(\breve{\theta}_{kdt})$ is the direct estimator of bias, $\widetilde{Bias}(\breve{\theta}_{kdt})$ is the model-based estimate given by equation (5.48).

$$r_{cond}^{stand} = r_{cond}/\sqrt{var(r_{cond})}, \tag{5.56}$$

where r_{cond} is defined as in equation (5.55):

$$\boldsymbol{H}_i = diag(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}(\hat{\boldsymbol{\Delta}})\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}(\hat{\boldsymbol{\Delta}})^{-1}). \tag{5.57}$$

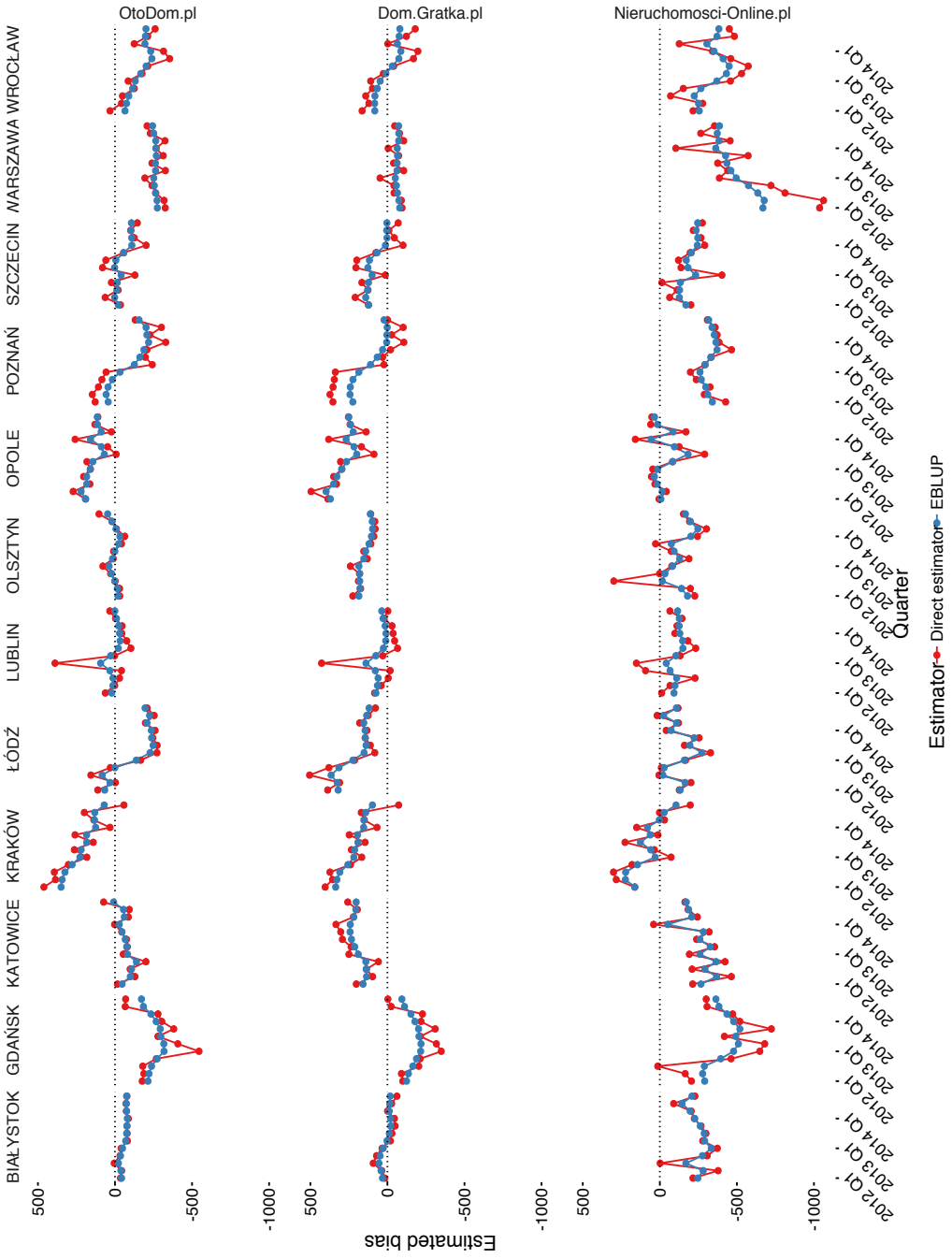


FIGURE 5.7: Comparison of direct and EBLUP estimates for 3 IDS and 12 cities between 2012Q1 and 2014Q4

Table 5.4 contains basic statistics on conditional residuals (r_{cond}). The overall mean is equal to -0.53 and is close to 0, but the median is -1.99. The minimum conditional residual is equal to -385.60 and the maximum is 315.00. These values indicate the presence of outliers in residuals. Hence, to check whether residuals are normally distributed standardized residuals are presented in a quantile-quantile plot (eq:standarizedcondresid) in Figure 5.8. The top part of Figure 5.8 indicates that the right tail is approximately normally distributed. The left bottom part has several outliers. Shapiro's normality test indicates that the distribution of standardized residuals is not normal, but the p-value equals 0.033. The results of the test are available in the Appendix, in R sample code A.4.6.

TABLE 5.4: Basic statistics for conditional residuals

Minimum	Q1	Median	Mean	Q3	Maximum
-385.60	-37.96	-1.99	-0.53	37.42	315.00

Figure 5.9 presents a scatterplot of leverage and standardized conditional residuals. Analysis of the relation between the leverage measure and r_{cond}^{stand} indicates the presence of several influential observations. Observations with the highest level of leverage refer mainly to estimates based on Nieruchomosci-Online.pl, particularly for Katowice (2012Q1, 2014Q4), Łódź (2012Q1, 2014Q4), Białystok (2014Q4), Opole (2014Q4) and Kraków (2012Q1, 2014Q4).

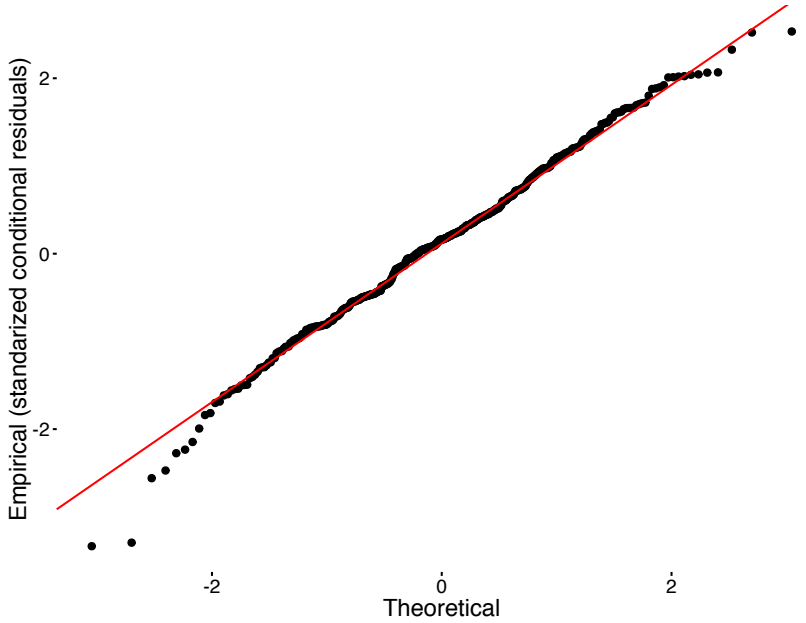


FIGURE 5.8: Quantile-Quantile plot of standardized conditional residuals

The last set of model diagnostics refers to random effects, which are assumed to be normally distributed. Due to the low number of levels (three sources), the data source effect was not analysed, nor was the AR(1) effect. Figure 5.10 presents a quantile-quantile plot for the domain random effect and the random effect of the

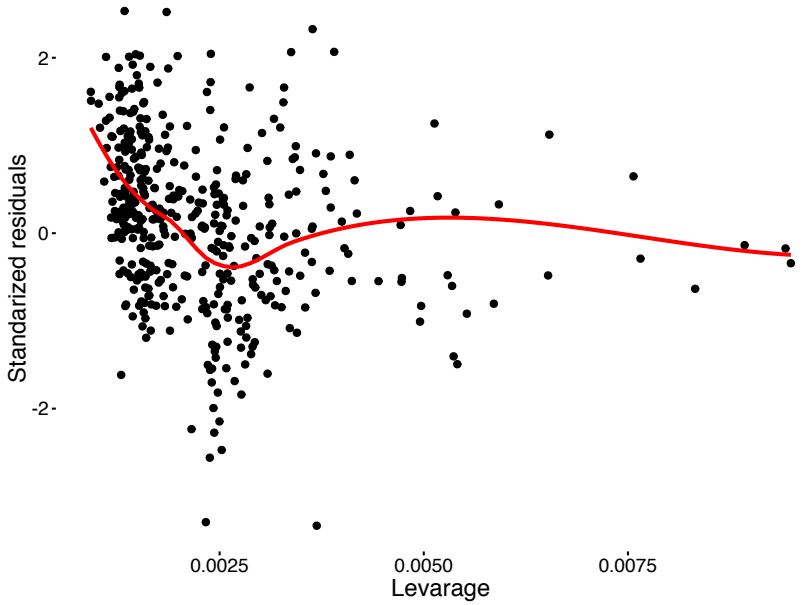


FIGURE 5.9: Scatterplot of leverage vs standardized conditional residuals

interaction between domain and IDS. In both cases the analysis was conducted using standardized residuals. In both cases points are close to the red line, which indicates a normal distribution. However, in Figure 5.10 outliers are present on both tails, but values lie between -2 and 2. The outliers in Figure 5.10 are Kraków, Warszawa, Opole and Gdańsk. For domain-IDS effect deviations from the red line are observed, but are not as high as for the domain random effect. Both effects were tested for normality (Shapiro and Kolmogorov-Smirnof test) and turned out positive. The results for these tests are available in R sample code A.4.7.

5.4 Conclusions

The above chapter has presented the results of estimating bias observed in IDSs, in this case bias in the average offer price per m^2. Because the NBP/CSO survey is a sample survey and IDSs can be treated as non-probability samples, a small area estimator was proposed. The proposed model is a linear mixed-model, which takes into account four random effects – domain (city) effect, correlated IDS (data source) effect, interaction between IDS and the domain effect and autocorrelation in bias. The model also takes into accout the variance of estimates based on the NBP/CSO survey and IDSs. The proposed approach makes it possible to decompose bias into these four random effect components and determine whether the MNAR process can be observed for the cities of interest. The results reported above show that the IDS (data source) effect accounts for for 55.4% of the bias and the domain (city) effect accounts for over 20.1%. Autocorrelation in the bias is equal to 0.69, which implies the presence of systematic bias in IDSs. Moreover, the relationship between the domain effect and the average offer and transaction price per m^2 for the cities indicates that bias is positively correlated with both types of prices. Hence, the higher average price per m^2, the bigger the bias

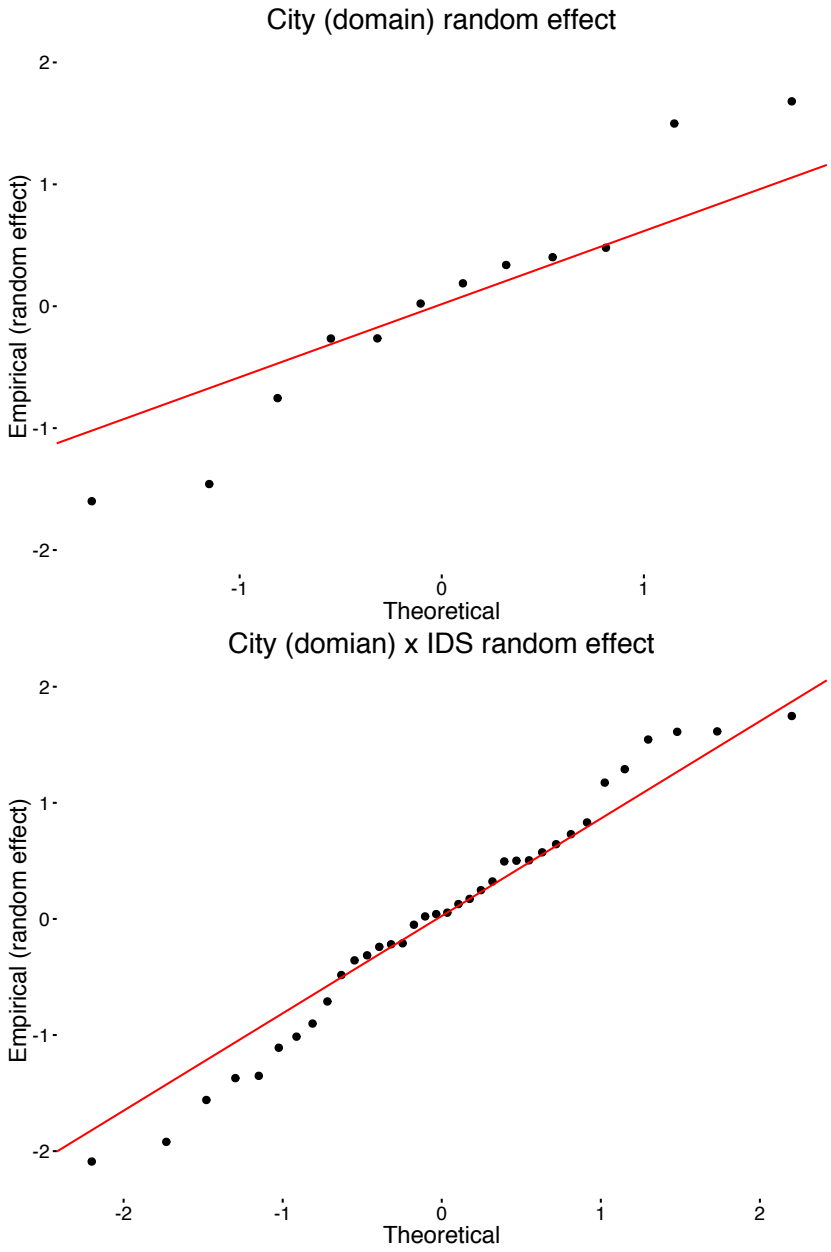


FIGURE 5.10: Quantile-quantile plot of city and city/IDS random effect

in IDSs. This relationship shows that the selection mechanism underlying undercoverage is connected with the target variable; in other words, the MNAR process is not only observed for Poznań (which was demonstrated in Chapter 4) but also for the other cities. Therefore, estimates based only on IDS provide systematically biased estimates of the offer price m^2 owing to the MNAR process. In order to provide accurate IDS-based estimates, further studies should involve methods that take into account the MNAR process.

Conclusions

Summary

The present dissertation is the first attempt of this kind at a comprehensive assessment of the usefulness of Internet data sources for official statistics about the real estate market. In particular, it is an attempt to:

- apply basic statistical concepts, such as population, statistical unit and target variable to IDSs,
- assess the representativeness of IDSs,
- assess bias of estimates for basic variables based on IDSs,
- assess the possibility of inference based on IDSs.

The main goal of the dissertation is *the evaluation of the Internet as a data source for real estate market statistics*. To achieve the main goal, the following hypothesis is formulated: *Internet data sources enable acceptable estimation of real estate market characteristics*.

The dissertation focuses on Internet data sources, which are not yet recognized by statistics. The way of applying basic statistical concepts to IDSs, the approach to measuring the representativeness of IDSs and assessing bias in estimates of the target variables are the first comprehensive attempts to assess the suitability of IDSs for statistical purposes. The issues raised in the dissertation have not been widely discussed in scientific literature before.

The results indicate that IDSs are suitable for providing statistics about the secondary real estate market. IDSs can be used to produce estimates of the offer price per m^2 with an acceptable level of absolute relative bias. It has been demonstrated that the overall difference between IDS-based estimates of the target variable (offer price per m^2) and those based on the NBP/CSO survey is not significant. The main reason for that is the relatively high variance of the estimates of the offer price per m^2 based on the NBP/CSO survey. Hence, the main goal of the dissertation has been achieved.

Scientific contribution

Problems connected with the use of IDSs for statistics covered in the dissertation have not been previously discussed in such detail in statistical literature. Chapter 1, Chapter 3 and partially Chapter 2 address the classification of IDSs with reference to statistical data sources, the definitional vagueness of populations in the real estate market and the problem of estimation that takes into account character

of IDSs. IDSs have been classified as non-probability and self-selected samples, a characteristic that has not been previously underlined in the statistical literature. New data sources, in particular the Internet and big data, are mainly discussed in the context of possible information they can provide. However, from the point of view of survey methodology self-selection inherent in these data sources is the most important issue.

The author adopts a two-phase life cycle of integrated statistical micro data proposed by Zhang (2012b) to IDSs. The main sources of non-sampling errors are identified and exemplified in the context of the secondary real estate market in Poland. Errors identified in advertisement services are also relevant in the case of other new data sources, particularly big data.

The representativeness of IDS is assessed using a two-step measurement procedure proposed by the author. Chapter 3 provides a detailed description of the procedure. It takes into account different data sources (e.g. surveys, register data) and aggregation levels (unit and domain). The procedure is not limited to IDSs but can easily be applied to other types of new data sources. Because the approach was practically tested, it could easily be applied to other fields. The proposed approach was applied to data sources about the secondary real estate market, as described in Chapter 4. This is the first comprehensive approach to measuring the representativeness of advertisement services.

Another theoretical contribution is related to the extension of a small area model for the estimation of bias. The model used in the study can be applied not only to IDSs but also to other data sources. It is not limited to data from the real estate market but can also be applied to other fields in order to assess and decompose bias. An important element of the model is the decomposition of bias into four components, making it possible to determine whether the response mechanism follows the pattern of missing not at random.

More importantly, the author has managed to verify the four hypotheses put forward in the dissertation, which are described in detail below:

H1 *The approach proposed by the author to measure representativeness can be effectively used to assess the representativeness of Internet data sources about the secondary real estate market* – Chapter 4 contains the results of applying the procedure to measure representativeness of IDSs using data for Poznań. As pointed out earlier, the Internet is widely used by estate market brokers operating in Poland. However, the representativeness of IDSs cannot easily be evaluated using existing official statistics. The author has identified differences between the Register of Real Estate Prices and Values and the NBP/CSO survey in terms of the distribution of prices. The first comparison indicates that there are two specific groups of residential properties that are underrepresented online: (1) low-priced properties with small floor area, and (2) high-priced properties with large floor area. There can be several reasons for this situation: (1) some properties are sold by court bailiffs at auctions for a price lower than the market price; (2) properties sold between family members or relatives are not presented online; (3) some transactions take place without the use of advertisement services; (4) brokers do not advertise all their properties for sale online; (5) the cheapest and the most expensive residential properties are presented to a specific

group of customers. The distributions of the number of rooms and floor area in the NBP/CSO survey and between IDSs are significantly different, particularly in the group of small (one room, less than 40 m^2) and big (four and more rooms, over 80m^2) properties.

H2 *Internet data sources are biased and this bias varies between sources and domains* – Chapter 5 has provided the theoretical basis for the estimation of bias in the average offer price per m^2. The results indicate that the bias of the IDSs-based average offer price per m^2 is higher for less popular data sources (Nieruchomosci-Online.pl) and lower for more popular data sources. Moreover, it has been demonstrated that the inclusion of the random effect for domain, data source and the interaction between the domain and the source significantly improves the model. It turns out that bias is highly autocorrelated in time.

H3 *Self-selection in Internet data sources about the real estate market is informative (depends on the target variable)* – Chapter 5 contains results based on the small area model used to estimate the bias of the average price per m^2 using IDS data. The model accounts for the random effect for domain, which was used to determine whether self-selection is informative (in other words whether the selection mechanism is MNAR). The results indicate the presence of informative self-selection for the 12 cities of interest. The bias of the average offer price per m^2 is positively correlated with average offer and transaction price per m^2. Results indicate that bias is higher for cities where average price of properties is high, and lower for cities with cheaper properties.

H4 *Internet data sources can be used to estimate the offer price per m^2 in the secondary real estate market with acceptable error measured by absolute relative bias* – Chapter 5 reports bias estimates based on the small area model, specifically a linear mixed model with four random effects. The overall average bias (represented by the intercept in the model) was found not to significant, which implies that the differences between IDSs and statistics based on the NBP/CSO survey are not substantial. The results indicate that the relative absolute bias in IDSs ranges from 0.2% to just over 4%. NSIs publish survey based estimates alongside with measure of precision defined by coefficient of variation (CV). Threshold necessary for publication of direct survey estimates differ between countries. For instance, according to Office for National Statistics (2006), estimates that do not meet the 20% threshold are not published, ISTAT has set the threshold at 15% for domains and at 18% for small domains and Statistics Canada applies the 16.5% threshold for estimates released with no restrictions and 33.3% for those published with a warning (Eurostat, 2013, ch. 2.4). However, NSI publications do not include information about bias. Taking into account the above criteria for evaluating the precision and the threshold of 15%–20%, the results can be treated as acceptable.

Possible directions of future research

The author has addressed important problems concerning representativeness and the estimation of bias in IDSs. However, the proposals described in the dissertations do not resolve other major problems connected with the use of IDSs, which can be investigated in the future. The possible areas of research include:

1. *Unit-error theory* – one of the main problems associated with IDSs is the difficulty of transforming objects (actions, non-statistical units) into statistical units. Zhang (2011) proposes an approach to assess uncertainty for the household unit based on register data. However, in the context of IDSs the identification of statistical units is not straightforward. Uncertainty is not only connected with the identification of statistical units but has to be taken into account when detecting non-statistical units (Zhang (2011) refers to them as base units). Some work is being done in this field, especially in the context of profiling users of social media or mobile phones (Daas et al., 2015, sec. 3.4).

2. *Detection of the self-selection mechanism in Internet data sources* – the self-selection mechanism in IDSs is unknown, and should be carefully studied. Moreover, self-selection bias can depend on the data source, the economic sector or the geographic location. This means that separate models should be constructed for specific purposes. Explanation of this mechanism is one of the most important challenges that need to be tackled, because only correct model specification can ensure unbiased estimates of the target population.

3. *Inference under MNAR / informative self-selection* – some possible approaches to the modelling process under informative sampling are proposed in Gelman (2007), Pfeffermann (2011), Pfeffermann and Sverchkov (2003), and Pfeffermann and Sverchkov (2007). Another approach under informative sampling and non-response is proposed by Feder and Pfeffermann (2015). Riddles (2013) describes a propensity score that takes into account the assumptions of MNAR . This problem requires models of the selection and response mechanism that resemble the population model. Empirical likelihood might be an interesting approach to estimate population characteristics under MNAR (see Feder and Pfeffermann, 2015; Owen, 2001; Owen, 2013).

4. *Inference in the absence of population information* – Internet data sources could provide information at the level of aggregation which is not available in official statistics. However, there is a lack of reference official information for the target population, either for the country or domain level. For instance, Reilly et al. (2001) analyse post-stratification of political polling data in the absence of auxiliary information at the population level.

5. *Weighting IDSs/big data* – in survey methodology a vector of weights is calculated to provide unbiased estimates of totals, means or quantiles. This approach is also applied to administrative data. There are two possible

avenues of research: (1) weighting based on linkage with surveys or/and administrative sources; (2) propensity weighting based on models of the selection mechanism. One possible solution can be based on the Generalized Weight Share Method under probabilistic record linkage (Lavallée and Caron, 2001).

6. *Linkage with official and non-official sources* – linkage with official (surveys, census, reporting) and non-official (in particular registers) sources is not straightforward. On the one hand, IDSs often require the extraction of variables that can be used for linkage; on the other hand, probabilistic methods may have limited applications for massive sources. Possible solutions for linkage (entity resolution) can be found in Steorts et al. (2013, 2015) and Winkler (2006).

7. *Measurement of estimation and representativeness under probabilistically-linked data* – probabilistic record linkage is discussed in the dissertation, but is not the subject of further study. Integration of register or survey data with IDSs may require approaches that account for the probabilistic character of the linkage. For instance, R-indicators may not provide reliable estimates under this setting. Foundations for modelling under probability-linked data are proposed by Chambers (2009) and further studied by Kim and Chambers (2012) and Samart and Chambers (2014). Samart (2011) provides the theory for linear mixed models based on probabilistically linked data.

8. *Adoption of current survey methods for IDSs/big data* – due to the massive and *organic* rather than *designed* character of such data, existing survey methods may not be applicable. For one thing, there are computational problems; another difficulty is related to natural language processing. Consequently, what is required is an interplay between information systems, technology and statistics (often referred to broadly as *data science*). Approaches including the use of high performance computing (i.e. parallel processing) in the context of official statistics should be studied. For instance, Daas et al. (2015, sec. 3) discusses problems related to the use of new data sources for official statistics.

9. *Estimation of MSE for the proposed model* – the dissertation focuses on the estimation of bias in IDSs using a small area model but MSE for the proposed estimator was not calculated. Parametric bootstrap or block bootstrap can be used to calculate MSE for the model.

It can be argued that the dissertation fills a research gap in that it provides an evaluation of the suitability of IDSs for statistical purposes. The author's major contribution consists in assessing the quality and representativeness of IDS and developing a model-based approach to bias estimation. The topic is worthy of further investigation, particularly in the areas outlined above, and is relevant for broadening the scope of knowledge about IDSs.

Bibliography

Abramowicz, Witold (2013). *Knowledge-based information retrieval and filtering from the Web*. Vol. 746. Springer Science & Business Media (see pp. 1, 3).

Abramowicz, Witold, Pawel J Kalczynski, and Krzysztof Wecel (2002). *Filtering the Web to feed data warehouses*. Springer Science & Business Media (see p. 3).

Ahas, Rein, Anto Aasa, Ülar Mark, Taavi Pae, and Ain Kull (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management* 28 (3), pp. 898–910 (see p. 24).

Alper, Melike Oguz and Yves G Berger (2015). Variance Estimation of Change in Poverty Rates : an Application to the Turkish EU-SILC Survey. *Journal of Official Statistics* 31 (2), pp. 155–175 (see p. 160).

Ann Keller, Sallie, Steven E Koonin, and Stephanie Shipp (2012). Big data and city living–what can it do for us? *Significance* 9 (4), pp. 4–7 (see p. 20).

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D Shapiro (2014). *Using social media to measure labor market flows*. Tech. rep. National Bureau of Economic Research. URL: http://www-personal.umich.edu/~shapiro/papers/LaborFlowsSocialMedia.pdf (see p. 20).

Bacchini, F, M D'Alò, S Falorsi, A Fasulo, and C Pappalardo (2014). "Does Google index improve the forecast of Italian labour market?" URL: www.sis2014.it/proceedings/allpapers/3019.pdf (see pp. 24, 25).

Baffour, Bernard, Thomas King, and Paolo Valente (2013). The Modern Census: Evolution, Examples and Evaluation. *International Statistical Review* 81 (3), pp. 407–425. ISSN: 03067734. DOI: 10.1111/insr.12036 (see pp. 10, 87).

Bailar, BA (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association* 70 (349), pp. 23–30. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1975.10480255 (see p. 72).

Baker, Reg, J Michael Brick, Nancy A Bates, Mike Battaglia, Mick P Couper, Jill A Dever, Krista J Gile, and Roger Tourangeau (2013). Summary Report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* 1, pp. 90–143 (see p. 4).

Bakker, Bart F M (2010). "Micro-Integration: State of the art". URL: http://www.istat.it/it/files/2013/12/FinalReport_WP1.pdf (see p. 72).

Bapna, Ravi, Paulo Goes, Ram Gopal, and James R Marsden (2006). Moving from data-constrained to data-enabled research: experiences and challenges in collecting, validating and analyzing large-scale e-commerce data. *Statistical Science*, pp. 116–130 (see pp. 3, 17).

Barcaroli, Giulio, Monica Scannapieco, Alessandra Nurra, Marco Scarn, and Sergio Salamone (2015). Internet as Data Source in the Istat Survey on ICT in Enterprises. 44 (April), pp. 31–43. DOI: 10.17713/ajs.v44i2.53 (see pp. 22, 34).

Bates, Douglas, Katharine M. Mullen, John C. Nash, and Ravi Varadhan (2014). minqa: Derivative-free optimization algorithms by quadratic approximation. R package version 1.2.4. URL: http://CRAN.R-project.org/package=minqa (see p. 164).

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 (see p. 12).

Bayer, M. and Douglas Laney (2012). *The Importance of 'Big Data': A Definition*. URL: https://www.gartner.com/doc/2057415/importance-big-data-definition (see p. 12).

Berger, Yves G. and Rodolphe Priam (2016). A simple variance estimator of change for rotating repeated surveys: an application to the EU-SILC household surveys. *Journal of Royal Statistical Society: Series A* XXX (XXX), p. XXX. ISSN: 1467985X. DOI: 10.1111/rssa.12116. URL: http://eprints.soton.ac.uk/347142/ (see pp. 160, 172).

Beręsewicz, Maciej and Marcin Szymkowiak (2015). Big data w statystyce publicznej - nadzieje, osiągnięcia, wyzwania i zagrożenia. *Ekonometria* 2 (47) (see pp. 36, 104).

Bethlehem, Jelke (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics* 4 (3), pp. 251–260 (see p. 82).

Bethlehem, Jelke (2009). *Applied survey methods: A statistical perspective*. John Wiley & Sons (see pp. 6, 9, 10, 84, 86, 132, 145).

Bethlehem, Jelke (2010). Selection bias in web surveys. *International Statistical Review* 78 (2), pp. 161–188 (see pp. 79, 82, 83, 171).

Bethlehem, Jelke and Silvia Biffignandi (2011). *Handbook of web surveys*. John Wiley & Sons (see pp. 26, 78, 79, 82, 83).

Biemer, Paul P and Lars E Lyberg (2003). *Introduction to Survey Quality*. Wiley Series in Survey Methodology. John Wiley & Sons, Inc. ISBN: 3175723993 (see p. 73).

Boonstra, Harm Jan and Bart Buelens (2011). *Model-based estimation*. URL: http://www.cbs.nl/NR/rdonlyres/DA57C4D8-A631-4C04-B4EA-9165D264D0D6/0/2011x3706.pdf (see p. 11).

Borg, Andreas and Murat Sariyar (2015). RecordLinkage: Record Linkage in R. R package version 0.4-7. URL: http://CRAN.R-project.org/package=RecordLinkage (see p. 99).

Bosch, Olav ten and Dick Windmeijer (2014). "On the use of Internet robots for Official Statistics". Dublin, Manila. URL: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2014/Topic_3_NL.pdf (see p. 21).

Brackstone, GJ (1987). Issues in the use of administrative records for statistical purposes. *Survey methodology* 13 (1), pp. 29–43 (see p. 30).

Brakel, Jan A Van den and Sabine Krieg (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology* 35, pp. 177–190 (see p. 72).

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen (1984). *Classification and Regression Trees*. CRC Press (see p. 78).

Brick, J Michael (2013). Unit Nonresponse and Weighting Adjustments : A Critical Review. *Journal of Sfficial statistics* 29 (3), pp. 329–353. DOI: 10.2478/jos-2013-0026 (see pp. 4, 81, 92).

Buelens, Bart, Piet J. H. Daas, Joep Burger, Marco Puts, and Jan van den Brakel (2014). *Selectivity of Big Data*. URL: http://www.cbs.nl/NR/rdonlyres/457A097A-DA43-4006-AFE0-A8E8316CFEF0/0/201411x10pub.pdf (see pp. 3, 25, 90).

Bureau of Labour Statistics (2015). *Measurement of Unemployment in States and Local Areas*. URL: http://www.bls.gov/opub/hom/pdf/homch4.pdf (see p. 11).

Canty, Angelo and Brian Ripley (2015). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-17. URL: http://CRAN.R-project.org/package=boot (see p. 170).

Cavallo, A., B. Neiman, and R. Rigobon (2014a). Currency Unions, Product Introductions, and the Real Exchange Rate. en. *The Quarterly Journal of Economics* 129 (2), pp. 529–595. ISSN: 0033-5533. DOI: 10.1093/qje/qju008. URL: http://qje.oxfordjournals.org/content/129/2/529.full (see p. 20).

Cavallo, Alberto (2012). Scraped data and sticky prices. *MIT Sloan Research Paper*. URL: http://www.mit.edu/%7Eafc/papers/Cavallo-Scraped.pdf (see pp. 4, 20).

Cavallo, Alberto (2013). Online and Official Price Indexes: Measuring Argentina's Inflation. *Journal of Monetary Economics* 60 (2), pp. 152–165 (see pp. 4, 20).

Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia (2014b). *Inflation Expectations, Learning and Supermarket Prices: Evidence from Field Experiments*. Tech. rep. National Bureau of Economic Research (see p. 20).

Cavallo, Alberto, Brent Neiman, and Roberto Rigobon (2014c). *The price impact of joining a currency union: evidence from Latvia*. Tech. rep. National Bureau of Economic Research (see p. 20).

Census Bureau (2015). *Model-based Small Area Income & Poverty Estimates (SAIPE) for School Districts, Counties, and States*. URL: http://www.census.gov/did/www/saipe/ (see p. 11).

Chambers, Ray (2009). "Regression Analysis of Probability-Linked Data". Wellington. URL: http://www.statisphere.govt.nz/official-statistics-research/series/default.htm (see pp. 99, 107, 189).

Chandrasekaran, Deepak, David Costello, and Paul Stubbs (2012). *Social media profiling*. US Patent App. 13/465,335 (see p. 97).

Chen, Sixia, Jae-Kwang Kim, et al. (2014). Two-phase sampling experiment for propensity score estimation in self-selected samples. *The Annals of Applied Statistics* 8 (3), pp. 1492–1515 (see p. 4).

Choi, Hyunyoung and Hal Varian (2012). Predicting the present with Google Trends. *Economic Record* 88 (1), pp. 2–9 (see p. 19).

Christen, Peter (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media (see p. 99).

Citro, Constance F (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey methodology* 40 (2), pp. 137–161 (see pp. 2, 11, 15, 16).

Cleveland, Robert B, William S Cleveland, Jean E McRae, and Irma Terpenning (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics* 6 (1), pp. 3–73 (see p. 109).

Cleveland, William S (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association* 74 (368), pp. 829–836 (see p. 109).

Cobben, Fannie (2009). "Nonresponse in sample surveys: methods for analysis and adjustment". PhD thesis. URL: http://www.cbs.nl/nr/rdonlyres/2c300d9d-c65d-4b44-b7f3-377bb6cea066/0/2009x11cobben.pdf (see p. 79).

Coleman, David (2013). The Twilight of the Census. *Population and Development Review* 38 (s1), pp. 334–351. ISSN: 00987921. DOI: 10.1111/j.1728-4457.2013.00568.x (see pp. 10, 87).

Consiglio, Loredana Di and Tiziana Tuoto (2015). Coverage Evaluation on Probabilistically Linked Data. 31 (3), pp. 415–429 (see p. 99).

Couper, M. P. (2011). The Future of Modes of Data Collection. *Public Opinion Quarterly* 75 (5), pp. 889–908. ISSN: 0033-362X. DOI: 10.1093/poq/nfr046. URL: http://poq.oxfordjournals.org/cgi/doi/10.1093/poq/nfr046 (see pp. 19, 73).

Couper, Mick P (2013). Is the Sky Falling ? New Technology , Changing Media , and the Future of Surveys. *Survey Research Methods* 7 (3), pp. 145–156 (see p. 19).

CSO (2014a). *Information society in Poland in 2010-2014*. URL: http://stat.gov.pl/obszary-tematyczne/nauka-i-technika-spoleczenstwo-informacyjne/spoleczenstwo-informacyjne/spoleczenstwo-informacyjne-w-polsce-wyniki-badan-statystycznych-z-lat-2010-2014,1,8.html (see p. 102).

CSO (2014b). *The Household Budget Survey*. URL: http://stat.gov.pl/obszary-tematyczne/warunki-zycia/dochody-wydatki-i-warunki-zycia-ludnosci/budzety-gospodarstw-domowych-w-2014-r-,9,9.html (see p. 102).

CSO (2015a). *National Official Business Register, REGON*. Accessed: 2015-11-15. URL: http://bip.stat.gov.pl/en/regon/ (see p. 49).

CSO (2015b). *Use of ICT technology in enterprises and households in 2015*. URL: http://stat.gov.pl/obszary-tematyczne/nauka-i-technika-spoleczenstwo-informacyjne/spoleczenstwo-informacyjne/wykorzystanie-technologii-informacyjno-telekomunikacyjnych-w-przedsiebiorstwach-i-gospodarstwach-domowych-w-2015-r-,3,13.html (see p. 102).

Daas, P and MJH Puts (2014). *Social Media Sentiment and Consumer Confidence*. URL: http://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.pdf (see p. 22).

Daas, P., Marco Roos, Mark van de Ven, and Joyce Neroni (2012). *Twitter as a potential data source for statistics*. URL: http://pietdaas.nl/beta/pubs/pubs/DiscPaper_Twitter.pdf (see pp. 3, 22).

Daas, Piet and Joep Burger (2015). *Profiling Big Data sources to assess their selectivity*. NTTS Conference 2015. URL: http://www.cros-portal.eu/sites/default/files//Daas-etal_NTTS%202015%20abstract%20unblinded-v3z_903.pdf (see pp. 22, 97).

Daas, Piet, Marko Roos, Chris de Blois, Rutger Hoekstra, Olav ten Bosch, and Yinyi Ma (2011). *New data sources for statistics: Experiences at Statistics Netherlands*. Paper for the 2011 European New Technique and Technologies for Statistics conference, February (see p. 3).

Daas, PJH, MJ Puts, B Buelens, and PAM Van den Hurk (2013). *Big Data and official statistics*. Paper for the 2013 New Techniques and Technologies for Statistics conference, Brussels, Belgium (see p. 23).

Daas, PJH, MJ Puts, B Buelens, and PAM Van den Hurk (2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics* 2 (31) (see pp. 2, 23, 24, 32, 90, 97, 104, 188, 189).

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and Their Applications*. ISBN 0-521-57391-2. Cambridge University Press (see p. 170).

Dehnel, Grażyna (2009). Rejestr podatkowy w estymacji pośredniej dla małych firm na podstawie badania SP3. *Zeszyty Naukowe/Uniwersytet Ekonomiczny w Poznaniu* (116), pp. 27–37 (see p. 87).

Dehnel, Grażyna (2015). Rejestr podatkowy oraz rejestr ZUS jako źródło informacji dodatkowej dla statystyki gospodarczej–możliwości i ograniczenia. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu* (384), pp. 51–59 (see p. 87).

Dehnel, Grażyna and Elżbieta Gołata (2012). Wykorzystanie rejestrów administracyjnych w statystyce przedsiębiorstw. *Zeszyty Naukowe/Uniwersytet Ekonomiczny w Poznaniu* (227), pp. 63–83 (see p. 87).

Demunter, Christophe and Fernando Reis (2015). "Using mobile positioning data for official statistics: daydream nation or promised land?" NTTS. URL: http://www.cros-portal.eu/sites/default/files//Presentation%20S8AP2.pdf (see p. 24).

Deville, J and Pierre Lavallée (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology* 32 (2), p. 165 (see p. 64).

Deville, Jean-Claude and Carl-Erik Särndal (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association* 87 (418), pp. 376–382 (see p. 80).

Dittmann, Iwona (2013). Primary and secondary residential real estate markets in Poland–analogies in offer and transaction price development. *Real Estate Management and Valuation* 21 (1), pp. 39–48 (see p. 145).

D'Orazio, Marcello (2015). *StatMatch: Statistical Matching*. R package version 1.2.3. URL: http://CRAN.R-project.org/package=StatMatch (see pp. 101, 108, 109).

D'Orazio, Marcello, Marco Di Zio, and Mauro Scanu (2006). *Statistical matching: Theory and practice*. John Wiley & Sons (see pp. 100, 101, 107–109).

Dz.U. [Journal of Laws] No. 38, item 454, as amended (2001). *Rozporządzenie Ministra Rozwoju Regionalnego i Budownictwa z dnia 29 marca 2001 r. w sprawie ewidencji gruntów i budynków*. URL: http://isap.sejm.gov.pl/DetailsServlet?id=WDU20010380454 (see pp. 50, 51).

Dz.U. [Journal of Laws] No. 85, item 388, as amended (1994). *Ustawa z dnia 24 czerwca 1994 r. o własności lokali*. URL: http://isap.sejm.gov.pl/DetailsServlet?id=WDU19940850388 (see p. 52).

Dz.U. [Journal of Laws] No. 985 (2015). *Ustawa z dnia 15 maja 2015 r. o zmianie ustawy o gospodarce nieruchomościami*. URL: http://isap.sejm.gov.pl/DetailsServlet?id=WDU20150000985 (see p. 51).

Edelman, Benjamin (2012). Using Internet Data for Economic Research. *The Journal of Economic Perspectives* 26 (2), pp. 189–206 (see p. 20).

Einav, Liran and Jonathan Levin (2014). Economics in the age of big data. *Science* 346 (6210), p. 1243089 (see pp. 17, 20).

Eurostat (2013). *Handbook on precision requirements and variance estimation for ESS households surveys*. URL: http://ec.europa.eu/eurostat/documents/3859598/5927001/KS-RA-13-029-EN.PDF (see p. 187).

Eurostat (2015). *Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics*. Tech. rep. New York, NY; Geneva: Eurostat. URL: http://mobfs.positium.ee (see p. 24).

Eurostat (2015a). *Information society*. Accessed: 2015-11-15. URL: http://stat.gov.pl/cps/rde/xbcr/gus/lud_raport_z_wynikow_NSP2011.pdf (see pp. 92, 93).

Eurostat (2015b). *Methodological Manual for statistics on the Information Society*. Survey year 2015, version 1.0, Accessed: 2015-11-15. URL: http://ec.europa.eu/eurostat/web/information-society/methodology (see p. 94).

Fay, Robert E (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association* 91 (434), pp. 490–498 (see p. 79).

Fay, Robert E and Mamadou S Diallo (2012). *Small Area Estimation Alternatives for the National Crime Victimization Survey* (see p. 166).

Fay, Robert E and Roger A Herriot (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74 (366a), pp. 269–277 (see p. 11).

Feder, Mosche and Danny Pfeffermann (2015). *Statistical inference under non-ignorable sampling and non-response*. URL: http://eprints.soton.ac.uk/378245/ (see p. 188).

Feenstra, Robert C and Matthew D Shapiro (2007). *Scanner data and price indexes*. Vol. 64. University of Chicago Press (see p. 21).

Fellegi, Ivan P and Alan B Sunter (1969). A theory for record linkage. *Journal of the American Statistical Association* 64 (328), pp. 1183–1210 (see p. 99).

Flekova, Lucie and Iryna Gurevych (2013). "Can we hide in the web? large scale simultaneous age and gender author profiling in social media". *CLEF 2012 Labs and Workshop, Notebook Papers*. Citeseer (see p. 97).

Fondeur, Y. and F. Karamé (2013). Can Google data help predict French youth unemployment? *Economic Modelling* 30, pp. 117–125. ISSN: 02649993. DOI:

10.1016/j.econmod.2012.07.017. URL: http://linkinghub.elsevier.com/retrieve/pii/S0264999312002490 (see p. 19).

Fosen, Johan and Li-Chun Zhang (2011). "The approach to quality evaluation of the micro-integrated employment statistics". ESSnet Data Integration (see pp. 6, 76, 155, 157, 158).

Gelman, Andrew (2007). Struggles with Survey Weighting and Regression Modeling. *Statistical Science* 22 (2), pp. 153–164. ISSN: 0883-4237. DOI: 10.1214/088342306000000691. URL: http://projecteuclid.org/euclid.ss/1190905511 (see pp. 4, 81, 188).

Ghosh, Malay and JNK Rao (1994). Small area estimation: an appraisal. *Statistical science*, pp. 55–76 (see pp. 4, 11).

Ginsberg, Jeremy, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant (2008). Detecting influenza epidemics using search engine query data. *Nature* 457 (7232), pp. 1012–1014 (see pp. 17, 18).

Goel, Sharad, Jake M Hofman, Sébastien Lahaie, David M Pennock, and Duncan J Watts (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America* 107 (41), pp. 17486–17490. ISSN: 0027-8424. DOI: 10.1073/pnas.1005962107 (see p. 19).

Gołata, Elzbieta (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 (see pp. 1, 11).

Gołata, Elżbieta and Grażyna Dehnel (2013). Rozbieżności szacunków NSP 2011 i BAEL. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu* 20 (278 Klasyfikacja i analiza danych-teoria i zastosowania), pp. 120–130 (see p. 87).

Gower, John C (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pp. 857–871 (see pp. 101, 107).

Gołata, Elżbieta (2012). Spis ludności i prawda/Population Census and Truth. *Studia Demograficzne* 1 (161). URL: http://www.degruyter.com/view/j/studdem.2012.1.issue-161/v10274-012-0002-y/v10274-012-0002-y.xml (see p. 87).

Grient, HA Van der and Jan de Haan (2010). "The use of supermarket scanner data in the Dutch CPI". *Joint ECE/ILO Workshop on Scanner Data*. Vol. 10 (see p. 21).

Griffioen, Robert, Jan de Haan, and Leon Willenborg (2014). *Collecting clothing data from the Internet*. URL: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/UNECE-ILO_2014_Griffioen_deHaan_Willenborg.pdf (see p. 22).

Groves, Robert M (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly* 70 (5), pp. 646–675 (see p. 92).

Groves, Robert M (2011a). *"Designed Data" and "Organic Data"*. Accessed: 2014-02-15. URL: http://directorsblog.blogs.census.gov/2011/05/31/designed-data-and-organic-data/ (see pp. 14, 16).

Groves, Robert M (2011b). Three Eras of Survey Research. *Public Opinion Quarterly* 75 (5), pp. 861–871. ISSN: 0033-362X. DOI: 10.1093/poq/nfr057. URL: http://poq.oxfordjournals.org/cgi/doi/10.1093/poq/nfr057 (see pp. 10, 11, 14, 16).

Groves, Robert M, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau (2011). *Survey methodology*. Vol. 561. John Wiley & Sons (see p. 10).

Haan, Jan de and Heymerik A Van der Grient (2011). Eliminating chain drift in price indexes based on scanner data. *Journal of Econometrics* 161 (1), pp. 36–46 (see p. 21).

Haan, Jan De (2002). Generalised Fisher Price Indexes and the Use of Scanner Data in the Consumer Price Index (CPI). 18 (1), pp. 61–85 (see p. 21).

Hansen, Morris H and William N Hurwitz (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics* 14 (4), pp. 333–362 (see p. 10).

Henderson, Charles R (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pp. 423–447 (see p. 163).

Hobza, Tomáš and Domingo Morales (2012). Small area estimation under random regression coefficient models. *Journal of Statistical Computation and Simulation* 9655 (August 2013), pp. 1–18. ISSN: 0094-9655. DOI: 10.1080/00949655.2012.684094. URL: http://www.tandfonline.com/doi/abs/10.1080/00949655.2012.684094 (see p. 166).

Hoekstra, Rutger, Olav ten Bosch, and Frank Harteveld (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), pp. 99–111 (see pp. 2, 3, 21).

Hollander, Myles, Douglas A Wolfe, and Eric Chicken (2013). *Nonparametric statistical methods*. John Wiley & Sons (see p. 110).

Holt, D. Tim (2007). The Official Statistics Olympic Challenge. *The American Statistician* 61 (1), pp. 1–8. ISSN: 0003-1305. DOI: 10.1198/000313007X168173. URL: http://www.tandfonline.com/doi/abs/10.1198/000313007X168173 (see pp. 3, 11).

Horrigan, Michael W. (2013). *Big Data: A Perspective from the BLS*. Accessed: 2014-02-15. URL: http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/ (see pp. 14, 16, 20).

Horvitz, Daniel G and Donovan J Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47 (260), pp. 663–685 (see p. 10).

Houbiers, Marianne (2004). Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. *Journal of official statistics* 20 (1), pp. 55–75 (see p. 79).

Ivancic, Lorraine, W Erwin Diewert, and Kevin J Fox (2011). Scanner data, time aggregation and the construction of price indexes. *Journal of Econometrics* 161 (1), pp. 24–35 (see p. 21).

Jank, Wolfgang and Galit Shmueli (2010). *Modeling online auctions*. Vol. 91. John Wiley & Sons (see p. 17).

Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, and Abe Usher (2015). Big Data in Survey Research AAPOR Task Force Report. *Public Opinion Quarterly* 79 (4), pp. 839–880 (see pp. 4, 17, 20).

Jensen, A. (1924). The report on the representative method in statistics. *Bulletin of the Inter- national Statistical Institute* XXII (1), 359–380 (see p. 9).

Józefowski, Tomasz and Beata Rynarzewska-Pietrzak (2010). Ocena możliwości wykorzystania rejestru PESEL w spisie ludności,[w:] E. *Gołata (red.), Pomiar i informacja w gospodarce, Zeszyt Naukowy WIGE, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań* (see p. 87).

Keller, Sallie Ann, Steven E Koonin, and Stephanie Shipp (2012). Big data and city living–what can it do for us? *Significance* 9 (4), pp. 4–7 (see p. 14).

Kiaer, AN (1897). The representative method of statistical surveys. *Norwegian Academy of Science and Letters, II The Historical, philosophical Section* (4) (see p. 9).

Kim, Gunky and Raymond Chambers (2012). Regression analysis under incomplete linkage. *Statistica Neerlandica* 56 (9), pp. 2756–2770. DOI: http://dx.doi.org/10.1016/j.csda.2012.02.026. URL: http://www.sciencedirect.com/science/article/pii/S0167947312001089 (see pp. 99, 107, 189).

Kim, Jae Kwang and Minsun Kim Riddles (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Survey Methodology* 38 (2), pp. 157–165. URL: http://www.statcan.gc.ca/pub/12-001-x/2012002/article/11754-eng.pdf (see p. 4).

Kim, Jae Kwang and Chris J Skinner (2013). Weighting in survey analysis under informative sampling. *Biometrika* 100 (2), pp. 385–398. DOI: 10.1093/biomet/ass085. eprint: http://biomet.oxfordjournals.org/content/100/2/385.full.pdf+html. URL: http://biomet.oxfordjournals.org/content/100/2/385.abstract (see p. 4).

Kitchin, Rob (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS* 31 (3), pp. 471–481 (see p. 30).

Knottnerus, Paul and Coen van Duin (2006). Variances in repeated weighting with an application to the Dutch Labour Force Survey. *Journal of Official Statistics* 22 (3), p. 565 (see p. 99).

Kreuter, Frauke (2013). *Improving surveys with paradata: Analytic uses of process information*. Vol. 581. John Wiley & Sons (see p. 31).

Krueger, Alan, Alexandre Mas, and Xiaotong Niu (2014). *The Evolution of Rotation Group Bias: Will the Real Unemployment Rate Please Stand Up?* Tech. rep. National Bureau of Economic Research (see p. 72).

Kruskal, W and F Mosteller (1979a). Representative sampling I: Non-scientific literature. *International Statistical Review* 47 (1), pp. 13–24. URL: http://www.jstor.org/stable/1402564 (see pp. 6, 83–87, 102).

Kruskal, W and F Mosteller (1979b). Representative sampling II: Scientific literature excluding statistics. *International Statistical Review* 47 (2), pp. 111–123. URL: http://www.jstor.org/stable/1402564 (see pp. 6, 83–87, 102).

Kruskal, W and F Mosteller (1979c). Representative sampling III: The current statistical literature. *International Statistical Review* 47 (3), pp. 245–265. URL: http://www.jstor.org/stable/1402647 (see pp. 6, 83–87, 102).

Lahiri, Partha (2015). "Can big data help in the production of reliable local area statistics?" SDAL, Virginia Tech. URL: http://www.slideshare.net/kimlyman/lahiri-talksdal3 (see p. 25).

Lahti, Leo, Przemyslaw Biecek, Janne Huovari, and Markus Kainu (2014). *Eurostat R package*. URL: https://github.com/rOpenGov/eurostat (see pp. 108, 112).

Laney, Doug (2001). *3-D data management: Controlling data volume, velocity and variety*. Accessed: 2014-02-27. URL: http://goo.gl/Bo3GS (see p. 12).

Lang, Duncan Temple (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 (see p. 119).

Lang, Duncan Temple (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 (see p. 119).

Lavallée, Pierre (2009). *Indirect sampling*. Vol. 7397. Springer Science & Business Media (see pp. 64, 65).

Lavallée, Pierre and Pierre Caron (2001). Estimation Using the Generalised Weight Share Method : The Case of Record Linkage. *Survey Methodology* 27 (2), pp. 155–169 (see pp. 64, 99, 189).

Lazer, David M, Ryan Kennedy, Gary King, and Alessandro Vespignani (2014). The parable of Google Flu: traps in big data analysis. American Association for the Advancement of Science (AAAS) (see pp. 1, 4, 17, 18).

Lee, Kwan Ok and Masaki Mori (2015). Do Conspicuous Consumers Pay Higher Housing Premiums? Spatial and Temporal Variation in the United States. *Real Estate Economics*. DOI: 10.1111/1540-6229.12115 (see p. 2).

Lee, Sunghee (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics* 22 (2), p. 329 (see pp. 4, 78).

Lehtonen, R and A Vejanen (2012). Small area poverty estimation by model calibration. *Journal of the Indian Society of Agricultural Statistics* 66 (1), pp. 125–133 (see p. 4).

Lehtonen, Risto and Ari Veijanen (1998). Logistic generalized regression estimators. *Survey Methodology* 24, pp. 51–56 (see p. 4).

Lehtonen, Risto and Ari Veijanen (2009). Design-based methods of estimation for domains and small areas. *Handbook of statistics* 29, pp. 219–249 (see p. 4).

Liao, Dan (2014). *Using "Small Data" to Improve the Use of "Big Data"*. Accessed: 2015-05-03. URL: https://blogs.rti.org/surveypost/2014/02/03/using-small-data-to-improve-the-use-of-big-data/ (see p. 25).

Lohr, Sharon (2009). *Sampling: design and analysis*. Cengage Learning (see pp. 10, 79).

Lohr, Sharon and JN K Rao (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association* 101 (475), pp. 1019–1030 (see p. 41).

Lohr, Sharon L and JNK Rao (2000). Inference from dual frame surveys. *Journal of the American Statistical Association* 95 (449), pp. 271–280 (see p. 41).

Lohr, SL and JM Brick (2012). Blending domain estimates from two victimization surveys with possible bias. *Canadian Journal of Statistics* 40 (4), pp. 679–696. URL: http://onlinelibrary.wiley.com/doi/10.1002/cjs.11153/full (see pp. 72, 166).

Lumley, Thomas (2004). Analysis of complex survey samples. *Journal of Statistical Software* 9 (1), pp. 1–19 (see p. 171).

Lumley, Thomas (2014). survey: analysis of complex survey samples. R package version 3.30. URL: http://CRAN.R-project.org/package=survey (see p. 171).

Lundström, Sixten and Carl-Erik Särndal (1999). Calibration as a standard method for treatment of nonresponse. *Journal Of Official Statistics* 15 (2), pp. 305–328. URL: http://www.jos.nu/Articles/abstract.asp?article=152305 (see p. 80).

Lynch, Clifford (2008). Big data: How do your data grow? *Nature* 455 (7209), pp. 28–29 (see p. 1).

Manzi, Giancarlo, David J. Spiegelhalter, Rebecca M. Turner, Julian Flowers, and Simon G. Thompson (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 174 (1), pp. 31–50 (see p. 166).

Marchetti, Stefano, Caterina Giusti, Monica Pratesi, Nicola Salvati, Fosca Giannotti, Dino Padreschi, Salvatore Rinzivillo, Luca Pappalardo, and Lorenzo Gabrielli (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics* 2 (31) (see pp. 24, 25).

Marpsat, Maryse and Nicolas Razafindratsima (2010). Survey methods for hard-to-reach populations : introduction to the special issue. *Methodological Innovations Online* 5 (2), pp. 3–16. DOI: 10.4256/mio.2010.0014 (see p. 61).

Miller, Greg (2011). Social scientists wade into the tweet stream. *Science* 333 (6051), pp. 1814–1815 (see pp. 1, 3, 4).

Mohorko, Anja, Edith de Leeuw, and Joop Hox (2013). Internet coverage and coverage bias in Europe: developments across countries and over time. *Journal of Official Statistics* 29 (4), pp. 609–622 (see p. 16).

Moriarity, Chris and Fritz Scheuren (2001). Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics* 17 (3), pp. 407–422 (see p. 100).

Munoz-Lopez, Juan (2015). "Mobility analysis from Twitter data". NTTS. URL: http://www1.unece.org/stat/platform/download/attachments/109252755/Big%20Data%20Satellite%20Workshop.pptx?version=1&modificationDate=1425992394314&api=v2 (see p. 22).

Nagler, Jonathan and Joshua A Tucker (2015). Drawing Inferences and Testing Theories with Big Data. *PS: Political Science & Politics* 48 (01), pp. 84–88 (see p. 16).

NBP (2014a). Report on the situation on the markets of residential and commercial property in Poland in 2013. Finance stability department. Warsaw, Poland. URL: http://www.nbp.pl/en/publikacje/inne/annual_report_2013.pdf (see pp. 49, 103).

NBP (2014b). The real estate market - Information Quarterly. Finance stability department. Warsaw, Poland. URL: http://nbp.pl/home.aspx?f=/publikacje/rynek_nieruchomosci/index2.html (see pp. 49, 103).

NBP (2015a). Report on the situation on the markets of residential and commercial property in Poland in 2014. Finance stability department. Warsaw, Poland. URL: http://www.nbp.pl/en/publikacje/inne/annual_report_2014.pdf (see p. 49).

NBP (2015b). The real estate market - Information Quarterly. Finance stability department. Warsaw, Poland. URL: http://www.nbp.pl/homen.aspx?f=/en/publikacje/inne/real_estate_market_q.html (see p. 49).

Neyman, Jerzy (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, pp. 558–625 (see p. 9).

Office for National Statistics (2006). *Model-based estimates of ILO unemployment for LAD/UAs in Great Britain: guide for users*. URL: Availablefromhttp://www.ons.gov.uk/ons/guide-method/method-quality/specific/labour-market/subnational-labour/model-based-estimates-of-ilo-unemployment-for-lad-uas-in-great-britain---guide-for-users.pdf (see p. 187).

Ouwehand, Pim and Barry Schouten (2014). Measuring Representativeness of Short-Term Business Statistics. *Journal of Official Statistics* 30 (4), pp. 623–649. DOI: http://dx.doi.org/10.2478/JOS-2014-0041 (see pp. 4, 87, 105).

Owen, Art B (2001). *Empirical likelihood*. CRC press (see p. 188).

Owen, Art B. (2013). Self-concordance for empirical likelihood. *Canadian Journal of Statistics* 41 (3), pp. 387–397. ISSN: 03195724. DOI: 10.1002/cjs.11183 (see p. 188).

Pavlopoulos, Dimitris and Jeroen K Vermunt (2015). Measuring temporary employment. Do survey or register data tell the truth ? *Survey Methodology* 41 (1), pp. 197–214 (see p. 87).

Pfeffermann, Danny (2002). Small Area Estimation-New Developments and Directions. *International Statistical Review* 70 (1), pp. 125–143 (see pp. 4, 11).

Pfeffermann, Danny (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it. *Survey Methodology* 37 (2), pp. 115–136 (see pp. 76, 77, 85, 86, 89, 102, 188).

Pfeffermann, Danny (2013). New important developments in small area estimation. *Statistical Science* 28 (1), pp. 40–68 (see pp. 4, 11).

Pfeffermann, Danny (2015). Methodological issues and challenges in the production of official statistics. *Journal of Survey Statistics and Methodology* 3 (4), pp. 425–483 (see p. 2).

Pfeffermann, Danny and L Burck (1990). Robust Small Area Estimation Combining Time Series and Cross-Sectional Data. *Survey Methodology* 16, pp. 217–237 (see p. 166).

Pfeffermann, Danny and M Sverchkov (2003). Fitting generalized linear models under informative sampling. *Analysis of survey Data*, pp. 175–195 (see p. 188).

Pfeffermann, Danny and Michail Sverchkov (2007). Small-Area Estimation Under Informative Probability Sampling of Areas and Within the Selected Areas. *Journal of the American Statistical Association* 102 (480), pp. 1427–1439. ISSN: 0162-1459. DOI: 10.1198/016214507000001094. URL: http://www.tandfonline.com/doi/abs/10.1198/016214507000001094 (see pp. 4, 188).

Porter, Aaron T, Scott H Holan, Christopher K Wikle, and Noel Cressie (2014). Spatial Fay–Herriot models for small area estimation with functional covariates. *Spatial Statistics* 10, pp. 27–42 (see p. 24).

Powell, Michael JD (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. Report No. DAMTP 2009/NA06, Centre for Mathematical Sciences, University of Cambridge, UK. URL: http://www.damtp.cam.ac.uk/user/na/NA_papers/NA2009_06.pdf (see p. 164).

Pratesi, Monica, Dino Pedreschi, Fosca Giannotti, Stefano Marchetti, Nicola Salvati, and Filomena Maggino (2013). "Small area model-based estimators using big data sources". NTTS. URL: http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_208.pdf (see pp. 24, 25).

Pratesi, Monica, Fosca Giannotti, Caterina Giusti, Stefano Marchetti, Dino Pedreschi, and Nicola Salvati (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 (see p. 24).

Puts, Marco, Piet Daas, and Martijn Tennekes (2015). "High frequency road sensor data for official statisitics". NTTS. URL: http://www.cros-portal.eu/content/high-frequency-road-sensor-data-official-statisitics-marco-puts-piet-daas-martijn-tennekes (see p. 23).

Qualité, Lionel and Yves Tillé (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology* 34 (2), pp. 173–181 (see pp. 160, 172).

Rao, JNK (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association* 91 (434), pp. 499–506 (see p. 79).

Rao, JNK and Isabel Molina (2015). *Small area estimation*. Second Edition. John Wiley & Sons (see pp. 4, 179).

Rao, JNK et al. (2011). Impact of frequentist and Bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science* 26 (2), pp. 240–256 (see p. 11).

Rao, Jonathan NK (2003). *Small area estimation*. 1st. Vol. 331. John Wiley & Sons (see pp. 4, 11, 108).

Rässler, Susanne (2012). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. Vol. 168. Springer Science & Business Media (see pp. 100, 101, 106).

Reilly, Cavan, Andrew Gelman, and Jonathan Katz (2001). Poststratification Without Population Level Information on the Poststratifying Variable With Application to Political Polling. *Journal of the American Statistical Association* 96 (453), pp. 1–11. DOI: 10.1198/016214501750332640 (see pp. 81, 89, 188).

Renssen, Robbert H, AH Kroese, and AJ Willeboordse (2001). Aligning estimates by repeated weighting. *Report, Central Bureau of Statistics, The Netherlands* (see pp. 81, 99).

Riddles, Minsun Kim (2013). "Propensity score adjusted method for missing data". PhD Thesis. Iowa State University. URL: http://lib.dr.iastate.edu/etd/13287 (see p. 188).

Rosenbaum, Paul R and Donald B Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1), pp. 41–55 (see pp. 4, 78, 106).

Roszka, Wojciech (2013). "Statystyczna integracja danych w badaniach społeczno-ekonomicznych". PhD Thesis. Poznań University of Economics. URL: http://www.wbc.poznan.pl/dlibra/docmetadata?id=265243\&from=publication (see p. 99).

Rubin, Donald B (1976). Inference and missing data. *Biometrika* 63 (3), pp. 581–592 (see pp. 4, 70, 77).

Rubin, Donald B (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics* 4 (1), pp. 87–94 (see p. 100).

Rubin, Donald B (1987). *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons (see p. 79).

Rubin, Donald B (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association* 91 (434), pp. 473–489 (see p. 79).

Samart, Klairung (2011). "Analysis of probabilistically linked data". PhD thesis. Doctor of Philosophy thesis, School of Mathematics and Applied Statistics, University of Wollongong. URL: http://ro.uow.edu.au/theses/3513/ (see pp. 99, 107, 165, 189).

Samart, Klairung and Ray Chambers (2014). Linear Regression With Nested Errors Using Probability-Linked Data. *Australian & New Zealand Journal of Statistics* 56 (1), pp. 27–46. ISSN: 13691473. DOI: 10.1111/anzs.12052. URL: http://doi.wiley.com/10.1111/anzs.12052 (see pp. 99, 107, 189).

Sariyar, Murat and Andreas Borg (2010). The RecordLinkage package: Detecting errors in data. *The R Journal* 2 (2), pp. 61–67 (see p. 99).

Sariyar, Murat, Andreas Borg, and Klaus Pommerening (2012). Missing values in deduplication of electronic patient data. *Journal of the American Medical Informatics Association* 19 (1), pp. 76–82 (see p. 99).

Särndal, Carl-Erik (2007). The calibration approach in survey theory and practice. *Survey Methodology* 33 (2), pp. 99–119 (see p. 80).

Särndal, Carl-Erik and Sixten Lundström (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons (see pp. 4, 79, 80).

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman (2003). *Model assisted survey sampling*. Springer Science & Business Media (see p. 11).

Scholtus, Sander, Bart F M Bakker, and van Del (2015). "Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables". URL: http://www.cbs.nl/NR/rdonlyres/4E94A4AA-585E-4B50-AF9F-AFAA62EFBE99/0/modelling{_}measurement{_}error.pdf (see p. 87).

Schonlau, Matthias, Arthur Van Soest, Arie Kapteyn, and Mick Couper (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research* 37 (3), pp. 291–318 (see p. 79).

Schouten, Barry, Fannie Cobben, and Jelke Bethlehem (2009). Indicators for the representativeness of survey response. *Survey Methodology* 35 (1), pp. 101–113 (see pp. 4, 84, 85, 102).

Schouten, Barry, Natalie Shlomo, and Chris Skinner (2011). Indicators for Monitoring and Improving Representativeness of Response. *Journal of Official Statistics* 27 (2), pp. 1–24 (see p. 105).

Schouten, Barry, Jelke Bethlehem, Koen Beullens, Øyvin Kleven, Geert Loosveldt, Annemieke Luiten, Katja Rutar, Natalie Shlomo, and Chris Skinner (2012). Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators. *International Statistical Review* 80 (3), pp. 382–399 (see p. 105).

Scott, Steven L and Hal R Varian (2014). Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation* 5 (1), pp. 4–23 (see p. 19).

Shao, Jun et al. (2003). Impact of the bootstrap on sample surveys. *Statistical Science* 18 (2), pp. 191–198 (see p. 79).

Shirley, Kenneth E (2015). Hierarchical models for estimating state and demographic trends in US death penalty public opinion, pp. 1–28 (see p. 81).

Shlomo, Natalie and Barry Schouten (2013). *Theoretical Properties of Partial Indicators for Representative Theoretical Properties of Partial Indicators for Representative*. Tech. rep. January. University of Southampton. URL: http://www.risq-project.eu/papers/shlomo-schouten-2013.pdf (see p. 105).

Shlomo, Natalie, Chris Skinner, and Barry Schouten (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference* 142 (1), pp. 201–211. ISSN: 03783758. DOI: 10.1016/j.jspi.2011.07.008. URL: http://linkinghub.elsevier.com/retrieve/pii/S0378375811002783 (see p. 105).

Shmueli, Galit, Wolfgang Jank, and Ravi Bapna (2005). "Sampling eCommerce data from the web: Methodological and practical issues". *ASA Proc. Joint Statistical Meetings*. Vol. 941, p. 948. URL: https://archive.nyu.edu/bitstream/2451/14953/2/USEDBOOK11.pdf (see pp. 3, 17).

Smith, Paul, Mark Pont, and Tim Jones (2003). Developments in business survey methodology in the Office for National Statistics, 1994–2000. *Statistician* 52 (3), pp. 257–286 (see pp. 160, 172).

Statistics Finland (2004). Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland. *Handbook Series* (45) (see pp. 10, 99).

Statistics Netherlands (2015). *Labour force, monthly figures*. URL: http://www.cbs.nl/en-GB/menu/themas/arbeid-sociale-zekerheid/methoden/dataverzameling/korte-onderzoeksbeschrijvingen/labour-force-monthly-figures.htm (see p. 11).

Steorts, Rebecca C, Rob Hall, and Stephen E Fienberg (2013). A Bayesian approach to graphical record linkage and de-duplication. *Journal of American Statistical Association*. arXiv: 1312.4645. URL: http://arxiv.org/abs/1312.4645 (see p. 189).

Steorts, Rebecca C et al. (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis* 10 (4), pp. 849–875 (see p. 189).

Strączkowski, Łukasz (2011). Buyers' Preferences on the Local Housing Market. *Zeszyty Naukowe/Uniwersytet Ekonomiczny w Poznaniu* (188), pp. 103–116 (see pp. 1, 2).

Struijs, Peter and Piet Daas (2014). "Quality Approaches to Big Data in Official Statistics". the European Conference on Quality in Official Statistics, Vienna, Austria. URL: http://www.pietdaas.nl/beta/pubs/pubs/Q2014_session_33_paper.pdf (see p. 25).

Swier, Nigel (2015). "Using Web Scraped Data to Construct Consumer Price Indices". NTTS. URL: http://www.cros-portal.eu/sites/default/files//Presentation%20S6AP3.pdf (see p. 22).

Szreder, Mirosław (2007). O roli informacji spoza próby w badaniach sondażowych. *Przegląd Socjologiczny* (1), pp. 97–108 (see p. 4).

Szreder, Mirosław (2010). Losowe i nielosowe próby w badaniach statystycznych. *Przegląd Statystyczny* 57 (4), pp. 168–174 (see pp. 4, 77).

Szreder, Mirosław (2011). O niektórych źródłach i konsekwencjach braków odpowiedzi w badaniach ankietowych. *Marketing i Rynek* (5), pp. 2–5 (see p. 79).

Szreder, Mirosław (2015). Big data wyzwaniem dla człowieka i statystyki. *Wiadomości Statystyczne* 651 (8), pp. 1–11 (see p. 2).

Szymkowiak, Marcin (2007). Przyczynek do kalibracji w badaniach statystycznych z brakami odpowiedzi. *Zeszyty Naukowe / Akademia Ekonomiczna w Poznaniu* (nr 96), pp. 194–204. URL: http://bazekon.icm.edu.pl/bazekon/element/bwmeta1.element.ekon-element-000152383884?q=632fd6f5-24b8-4d2b-9f77-7e8b8578d188\$5\&qt=IN_PAGE (see p. 4).

Szymkowiak, Marcin (2009). "Estymatory kalibracyjne w badaniu budżetów gospodarstw domowych". PhD Thesis. Poznań University of Economics. URL: http://www.wbc.poznan.pl/dlibra/docmetadata?id=115816 (see p. 80).

Tam, SM (1984). On covariances from overlapping samples. *The American Statistician* 38 (4), pp. 288–289 (see pp. 160, 172).

Tennekes, Martijn and Marco Puts (2015). "Projection of road sensors to the Dutch road network". NTTS. URL: http://www.cros-portal.eu/content/projection-road-sensors-dutch-road-network-martijn-tennekes-marco-puts-statistics (see pp. 23, 24).

The Geodetic and Cartographic Act (1989). *Ustawa z dnia 17 maja 1989 r. Prawo geodezyjne i kartograficzne, Dz.U. 1989 nr 30 poz. 163 (in Polish)*. URL: http://isap.sejm.gov.pl/DetailsServlet?id=WDU19890300163 (see pp. 100, 115).

The Real Estate Management Act (1997). *Ustawa z dnia 21 sierpnia 1997 r. o gospodarce nieruchomościami, Dz.U. 1997 nr 115 poz. 741 (in Polish)*. URL: http://isap.sejm.gov.pl/DetailsServlet?id=WDU19971150741 (see p. 100).

Trojanek, Radosław (2008). *Wahania cen na rynku mieszkaniowym*. Wydawnictwo Akademii Ekonomicznej (see p. 145).

UNECE (2007). *Register-based statistics in the Nordic countries*. Tech. rep. New York, NY; Geneva: UNECE (see p. 10).

UNECE (2011). *Using Administrative and Secondary Sources for Official Statistics. A handbook of Principles and Practices*. New York and Geneva. URL:

http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf (see p. 13).

UNECE (2014). *How big is Big Data? Exploring the role of Big Data in Official Statistics*. URL: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=99484307 (see p. 13).

Varian, Hal R (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, pp. 3–27. DOI: 10.1257/jep.28.2.3 (see pp. 1, 20).

Vicente, María Rosalía, Ana J López-menéndez, and Rigoberto Pérez (2015). Technological Forecasting & Social Change Forecasting unemployment with internet search data : Does it help to improve predictions when job destruction is skyrocketing ? *Technological Forecasting & Social Change* 92, pp. 132–139. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2014.12.005. URL: http://dx.doi.org/10.1016/j.techfore.2014.12.005 (see p. 19).

Viechtbauer, Wolfgang (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 36 (3), pp. 1–48. URL: http://www.jstatsoft.org/v36/i03/ (see p. 164).

Vosen, Simeon and Torsten Schmidt (2011). Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting* 30 (6), pp. 565–578 (see p. 19).

Wallgren, Anders and Britt Wallgren (2014). *Register-based Statistics*. Second. Wiley Series in Survey Methodology. John Wiley & Sons, Inc. (see pp. 1, 3, 10, 11, 30, 31, 80, 87, 97).

Wang, Wei, David Rothschildb, Sharad Goelb, and Andrew Gelman (2015). Forecasting Elections with Non-Representative Polls. *International Journal of Forecasting* 21 (3), 980–991 (see pp. 4, 81).

Wickham, Hadley (2015). httr: Tools for Working with URLs and HTTP. R package version 0.6.1. URL: http://CRAN.R-project.org/package=httr (see p. 119).

Wikipedia (2015). *Web scraping — Wikipedia, The Free Encyclopedia*. [Online; accessed 03.05.2015]. URL: http://en.wikipedia.org/wiki/Web_scraping (see p. 21).

Winkler, William E (2006). *Overview of record linkage and current research directions*. Tech. rep. Bureau of the Census. URL: https://www.census.gov/srd/papers/pdf/rrs2006-02.pdf (see pp. 99, 189).

Wolter, Kirk (2007). *Introduction to variance estimation*. Springer Science & Business Media (see pp. 108, 160).

Wood, John (2008). On the Covriance Between Related Horvitz-Thompson Estimators. *Journal of Official Statistics* 24 (1), p. 53 (see pp. 160, 172).

Wu, Jing and Yongheng Deng (2015). Intercity information diffusion and price discovery in housing markets: Evidence from Google Searches. *The Journal of Real Estate Finance and Economics* 50 (3), pp. 289–306 (see pp. 1, 2).

Wu, Lynn and Erik Brynjolfsson (2014). "The future of prediction: How Google searches foreshadow housing prices and sales". *Economics of Digitization*. University of Chicago Press (see pp. 1, 2, 19).

Xu, Wei, Ziang Li, Cheng Cheng, and Tingting Zheng (2012). Data mining for unemployment rate prediction using search engine query data. *Service Oriented Computing and Applications* 7 (1), pp. 33–42. ISSN: 1863-2386. DOI:

10.1007/s11761-012-0122-2. URL: http://link.springer.com/10.1007/s11761-012-0122-2 (see p. 19).

Yildiz, Dilek and Peter W F Smith (2015). Models for Combining Aggregate-Level Administrative Data in the Absence of a Traditional Census. *Journal of Official Statistics* 31 (3), pp. 431–451 (see p. 99).

Zhang, Li-chun (2005). On the Bias in Gross Labour Flow Estimates Due to Non-response and Misclassification. *Journal of Official Statistics* 21 (4), pp. 591–604 (see p. 73).

Zhang, Li-Chun (2011). A Unit-Error Theory for Register-Based Household Statistics. *Journal of Official Statistics* 27 (3), pp. 415–432 (see pp. 1, 3, 9–11, 97, 116, 188).

Zhang, Li-Chun (2012a). *On the accuracy of register-based census employment statistics*. European Conference on Quality in Official Statistics. URL: http://www.q2012.gr/articlefiles/sessions/23.4_Zhang_AaccuracyRegisterStatistics.pdf (see pp. 6, 76, 157, 158).

Zhang, Li-Chun (2012b). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica* 66 (1), pp. 41–63. ISSN: 00390402. DOI: 10.1111/j.1467-9574.2011.00508.x (see pp. 1, 3, 6, 10–12, 64, 65, 69–72, 87, 99, 155, 186).

Zhang, Li-Chun (2015). On modelling register coverage errors. *Journal of Official Statistics* 31 (3), 381–396. DOI: 10.1515/jos-2015-0023 (see p. 1).

Zhang, Li-Chun, Ib Thomsen, and Øyvin Kleven (2013). On the Use of Auxiliary and Paradata for Dealing With Non-sampling Errors in Household Surveys. *International Statistical Review* 81 (2), pp. 270–288. DOI: 10.1111/insr.12009 (see p. 31).

Appendix A

Appendix

A.1 Description of data, point and variance estimates

A.1.1 Point and variance estimates

TABLE A.1: Point estimates of average price m2 in 12 cities between 2012 Q1 and 2014 Q4 based on NBP/CSO survey Nieruchomosci-online.pl, Dom.Gratka.pl and OtoDom.pl

City	Quarter	NBP CSO	Dom.Gratka.pl	NieOnline	OtoDom.pl
BIAŁYSTOK	2012 Q1	4724.00	4741.20	4509.12	4688.55
BIAŁYSTOK	2012 Q2	4674.00	4710.56	4295.76	4630.69
BIAŁYSTOK	2012 Q3	4588.00	4680.06	4585.96	4593.04
BIAŁYSTOK	2012 Q4	4583.00	4653.33	4275.69	4550.39
BIAŁYSTOK	2013 Q1	4576.00	4610.43	4202.89	4535.87
BIAŁYSTOK	2013 Q2	4546.00	4526.87	4267.74	4465.44
BIAŁYSTOK	2013 Q3	4520.00	4490.49	4220.90	4441.06
BIAŁYSTOK	2013 Q4	4494.00	4443.83	4226.49	4416.29
BIAŁYSTOK	2014 Q1	4510.00	4464.44	4286.26	4423.29
BIAŁYSTOK	2014 Q2	4450.00	4449.04	4246.37	4376.48
BIAŁYSTOK	2014 Q3	4443.00	4415.49	4353.40	4371.69
BIAŁYSTOK	2014 Q4	4445.00	4383.09	4217.52	4367.36
GDAŃSK	2012 Q1	6384.00	6284.25	6178.93	6207.31
GDAŃSK	2012 Q2	6309.52	6218.63	6144.92	6122.09
GDAŃSK	2012 Q3	6273.74	6070.28	6287.15	6093.62
GDAŃSK	2012 Q4	6239.75	6027.39	5777.39	5971.73
GDAŃSK	2013 Q1	6277.90	5928.61	5629.39	5732.79
GDAŃSK	2013 Q2	6151.73	5834.02	5471.21	5743.99
GDAŃSK	2013 Q3	6007.49	5794.07	5588.71	5728.97
GDAŃSK	2013 Q4	6136.03	5826.00	5412.28	5754.79
GDAŃSK	2014 Q1	6102.82	5884.72	5583.62	5799.16
GDAŃSK	2014 Q2	6073.01	5845.66	5600.64	5794.08
GDAŃSK	2014 Q3	5858.06	5833.61	5551.87	5791.70
GDAŃSK	2014 Q4	5872.59	5869.45	5571.97	5803.54
KATOWICE	2012 Q1	4052.69	4254.64	3839.43	4036.53
KATOWICE	2012 Q2	4121.55	4217.67	3657.29	3992.06
KATOWICE	2012 Q3	4057.38	4193.46	3848.21	3959.27

Continued on next page

Point estimates of average price m2 in 12 cities between 2012 Q1 and 2014 Q4 based on NBP/CSO survey Nieruchomosci-online.pl, Dom.Gratka.pl and OtoDom.pl

City	Quarter	NBP CSO	Dom.Gratka.pl	NieOnline	OtoDom.pl
KATOWICE	2012 Q4	4137.88	4196.34	3714.70	3937.12
KATOWICE	2013 Q1	3955.35	4205.27	3763.97	3901.61
KATOWICE	2013 Q2	4000.76	4236.57	3645.46	3920.73
KATOWICE	2013 Q3	3973.02	4265.08	3734.61	3898.41
KATOWICE	2013 Q4	3927.87	4230.71	3607.95	3883.95
KATOWICE	2014 Q1	3915.26	4249.19	3955.71	3917.78
KATOWICE	2014 Q2	4045.57	4266.67	3802.22	3958.91
KATOWICE	2014 Q3	4026.95	4222.14	3843.69	3934.01
KATOWICE	2014 Q4	3917.61	4175.00	3754.72	3991.26
KRAKÓW	2012 Q1	6775.56	7179.34	6938.68	7237.00
KRAKÓW	2012 Q2	6724.24	7080.42	7007.05	7110.48
KRAKÓW	2012 Q3	6617.04	6989.93	6917.57	7010.61
KRAKÓW	2012 Q4	6648.38	6888.01	6828.89	6950.85
KRAKÓW	2013 Q1	6644.00	6810.40	6571.17	6826.60
KRAKÓW	2013 Q2	6488.58	6722.09	6524.21	6752.08
KRAKÓW	2013 Q3	6585.00	6729.99	6811.22	6726.00
KRAKÓW	2013 Q4	6537.00	6784.85	6549.25	6796.38
KRAKÓW	2014 Q1	6754.22	6821.80	6904.76	6786.58
KRAKÓW	2014 Q2	6682.36	6831.71	6652.02	6819.72
KRAKÓW	2014 Q3	6644.17	6813.92	6646.45	6843.35
KRAKÓW	2014 Q4	6860.48	6787.50	6663.17	6802.45
ŁÓDŹ	2012 Q1	4006.00	4393.48	3879.73	4117.82
ŁÓDŹ	2012 Q2	4033.00	4343.12	3830.90	4030.29
ŁÓDŹ	2012 Q3	3794.00	4298.91	3799.88	3950.70
ŁÓDŹ	2012 Q4	3854.00	4233.50	3844.37	3884.08
ŁÓDŹ	2013 Q1	3975.00	4188.66	3808.11	3808.12
ŁÓDŹ	2013 Q2	4058.00	4139.86	3728.99	3785.71
ŁÓDŹ	2013 Q3	4018.00	4128.73	3858.20	3744.30
ŁÓDŹ	2013 Q4	3978.00	4122.14	3723.43	3732.54
ŁÓDŹ	2014 Q1	3984.00	4116.76	3942.60	3723.69
ŁÓDŹ	2014 Q2	3915.00	4095.05	3794.16	3717.80
ŁÓDŹ	2014 Q3	3907.00	4032.16	3924.65	3652.51
ŁÓDŹ	2014 Q4	3892.00	3969.77	3773.81	3683.29
LUBLIN	2012 Q1	5057.00	5140.10	5045.18	5119.96
LUBLIN	2012 Q2	5076.00	5116.44	5009.95	5077.79
LUBLIN	2012 Q3	5065.00	5059.64	4836.67	5035.42
LUBLIN	2012 Q4	4995.00	4977.10	5087.69	4952.79
LUBLIN	2013 Q1	4491.00	4918.27	4643.13	4880.27
LUBLIN	2013 Q2	4858.00	4888.40	4725.98	4860.00
LUBLIN	2013 Q3	4935.00	4869.08	4701.83	4831.68
LUBLIN	2013 Q4	4901.00	4854.32	4719.06	4824.67

Continued on next page

Point estimates of average price m2 in 12 cities between 2012 Q1 and 2014 Q4 based on NBP/CSO survey Nieruchomosci-online.pl, Dom.Gratka.pl and OtoDom.pl

City	Quarter	NBP CSO	Dom.Gratka.pl	NieOnline	OtoDom.pl
LUBLIN	2014 Q1	4886.00	4849.26	4788.54	4841.72
LUBLIN	2014 Q2	4884.00	4854.22	4772.88	4837.63
LUBLIN	2014 Q3	4831.00	4851.89	4686.78	4827.48
LUBLIN	2014 Q4	4854.00	4851.15	4788.58	4887.83
OLSZTYN	2012 Q1	4628.00	4852.41	4401.32	4597.01
OLSZTYN	2012 Q2	4632.00	4805.77	4433.90	4601.94
OLSZTYN	2012 Q3	4544.00	4732.18	4842.69	4542.86
OLSZTYN	2012 Q4	4465.00	4646.05	4466.53	4493.23
OLSZTYN	2013 Q1	4379.38	4619.61	4301.98	4458.00
OLSZTYN	2013 Q2	4436.00	4566.40	4246.85	4447.29
OLSZTYN	2013 Q3	4400.00	4553.43	4325.89	4408.90
OLSZTYN	2013 Q4	4418.00	4525.45	4445.92	4375.67
OLSZTYN	2014 Q1	4446.00	4533.87	4201.73	4382.95
OLSZTYN	2014 Q2	4428.31	4507.59	4125.83	4419.16
OLSZTYN	2014 Q3	4405.00	4486.44	4212.35	4424.62
OLSZTYN	2014 Q4	4327.85	4436.20	4174.73	4431.50
OPOLE	2012 Q1	4072.46	4458.54	4077.90	4263.20
OPOLE	2012 Q2	3981.23	4476.57	3939.28	4252.70
OPOLE	2012 Q3	4065.77	4396.18	4097.69	4228.53
OPOLE	2012 Q4	4012.61	4361.22	4067.10	4216.06
OPOLE	2013 Q1	4074.00	4370.27	4119.79	4233.42
OPOLE	2013 Q2	4062.00	4366.15	3978.56	4244.25
OPOLE	2013 Q3	4246.39	4333.52	3955.76	4238.92
OPOLE	2013 Q4	4174.43	4342.68	4049.05	4222.51
OPOLE	2014 Q1	3974.74	4355.51	4133.77	4233.70
OPOLE	2014 Q2	4231.23	4369.50	4061.92	4254.06
OPOLE	2014 Q3	4097.00	4339.25	4156.32	4227.23
OPOLE	2014 Q4	4109.18	4362.08	4161.30	4220.04
POZNAŃ	2012 Q1	5604.16	5957.76	5177.03	5732.57
POZNAŃ	2012 Q2	5534.71	5907.30	5246.48	5682.00
POZNAŃ	2012 Q3	5518.54	5871.39	5192.33	5626.17
POZNAŃ	2012 Q4	5446.27	5790.95	5209.84	5531.20
POZNAŃ	2013 Q1	5406.26	5744.02	5207.99	5464.06
POZNAŃ	2013 Q2	5657.19	5679.72	5364.54	5415.22
POZNAŃ	2013 Q3	5628.48	5659.84	5295.03	5431.74
POZNAŃ	2013 Q4	5717.87	5697.64	5252.58	5510.35
POZNAŃ	2014 Q1	5830.06	5723.52	5445.52	5500.71
POZNAŃ	2014 Q2	5742.14	5714.72	5370.18	5514.54
POZNAŃ	2014 Q3	5806.61	5704.09	5447.30	5505.19
POZNAŃ	2014 Q4	5693.84	5692.73	5385.31	5562.61
SZCZECIN	2012 Q1	4586.21	4709.15	4384.76	4550.42
SZCZECIN	2012 Q2	4431.00	4640.13	4367.27	4495.04

Continued on next page

Point estimates of average price m2 in 12 cities between 2012 Q1 and 2014 Q4 based on NBP/CSO survey Nieruchomosci-online.pl, Dom.Gratka.pl and OtoDom.pl

City	Quarter	NBP CSO	Dom.Gratka.pl	NieOnline	OtoDom.pl
SZCZECIN	2012 Q3	4441.71	4567.94	4329.48	4421.35
SZCZECIN	2012 Q4	4336.21	4502.05	4322.69	4359.52
SZCZECIN	2013 Q1	4434.36	4447.95	4031.48	4305.00
SZCZECIN	2013 Q2	4170.89	4375.24	4033.50	4251.99
SZCZECIN	2013 Q3	4151.63	4350.28	4030.72	4211.06
SZCZECIN	2013 Q4	4247.23	4321.69	4050.63	4191.29
SZCZECIN	2014 Q1	4399.45	4297.99	4108.08	4197.48
SZCZECIN	2014 Q2	4315.46	4270.52	4048.25	4190.35
SZCZECIN	2014 Q3	4288.72	4279.10	4071.82	4186.73
SZCZECIN	2014 Q4	4345.76	4275.83	4069.72	4200.71
WARSZAWA	2012 Q1	9110.86	9015.72	8074.08	8784.65
WARSZAWA	2012 Q2	9035.28	8944.95	7972.67	8716.70
WARSZAWA	2012 Q3	8899.90	8856.76	8087.60	8635.10
WARSZAWA	2012 Q4	8767.55	8725.31	8046.52	8527.45
WARSZAWA	2013 Q1	8606.02	8653.97	8219.76	8412.44
WARSZAWA	2013 Q2	8637.97	8532.20	8197.70	8310.69
WARSZAWA	2013 Q3	8544.23	8504.29	8168.93	8304.36
WARSZAWA	2013 Q4	8626.63	8553.71	8053.71	8313.45
WARSZAWA	2014 Q1	8622.18	8617.32	8518.66	8355.20
WARSZAWA	2014 Q2	8690.68	8585.04	8234.54	8366.11
WARSZAWA	2014 Q3	8625.78	8547.04	8360.37	8396.22
WARSZAWA	2014 Q4	8635.85	8588.73	8281.04	8427.70
WROCŁAW	2012 Q1	6367.00	6531.36	6151.53	6399.31
WROCŁAW	2012 Q2	6307.00	6426.91	6029.14	6264.66
WROCŁAW	2012 Q3	6182.00	6322.86	6112.63	6133.25
WROCŁAW	2012 Q4	6100.00	6197.15	5947.65	5975.01
WROCŁAW	2013 Q1	5959.00	6067.49	5501.10	5873.57
WROCŁAW	2013 Q2	5979.00	6005.16	5448.65	5803.41
WROCŁAW	2013 Q3	5984.00	5943.35	5410.23	5771.32
WROCŁAW	2013 Q4	6098.00	5928.83	5637.99	5743.37
WROCŁAW	2014 Q1	6096.00	5898.24	5754.48	5780.93
WROCŁAW	2014 Q2	5899.00	5896.63	5773.18	5774.17
WROCŁAW	2014 Q3	5980.00	5857.01	5496.25	5767.62
WROCŁAW	2014 Q4	6017.00	5836.41	5566.04	5756.44

TABLE A.2: Variance of point estimates of average price m2 in 12 cities between 2012 Q1 and 2014 Q4 based on NBP/CSO survey Nieruchomosci-online.pl, Dom.Gratka.pl and OtoDom.pl

City	Quarter	NBP CSO	Dom.Gratka.pl	NieOnline	OtoDom.pl
BIAŁYSTOK	2012 Q1	7608.45	9120.12	6167.77	1893.54
BIAŁYSTOK	2012 Q2	6940.71	10505.23	3782.02	1867.12
BIAŁYSTOK	2012 Q3	6709.70	10106.33	4873.25	2118.08
BIAŁYSTOK	2012 Q4	6749.42	9654.07	3345.82	1971.87
BIAŁYSTOK	2013 Q1	7086.34	8556.36	1071.08	1773.19
BIAŁYSTOK	2013 Q2	7141.79	8647.46	1227.10	1821.62
BIAŁYSTOK	2013 Q3	7071.29	9123.04	1289.28	1655.08
BIAŁYSTOK	2013 Q4	7162.65	9833.96	1327.80	1932.08
BIAŁYSTOK	2014 Q1	7578.77	10215.65	486.56	2210.91
BIAŁYSTOK	2014 Q2	7500.02	10794.15	893.39	2263.69
BIAŁYSTOK	2014 Q3	7287.39	10241.06	1381.48	2192.21
BIAŁYSTOK	2014 Q4	7494.91	10401.43	928.83	2732.78
GDAŃSK	2012 Q1	26114.51	5765.81	7928.34	3141.36
GDAŃSK	2012 Q2	26416.73	6227.44	4874.60	3323.31
GDAŃSK	2012 Q3	26498.54	5063.70	4062.40	3166.59
GDAŃSK	2012 Q4	26720.98	6003.82	3233.34	8433.89
GDAŃSK	2013 Q1	26748.01	6686.57	1968.28	10870.68
GDAŃSK	2013 Q2	27023.81	6838.44	1385.54	2387.65
GDAŃSK	2013 Q3	25736.95	6323.69	2285.80	2274.05
GDAŃSK	2013 Q4	24707.89	7340.77	2008.87	2481.08
GDAŃSK	2014 Q1	25271.41	9638.15	628.79	2579.60
GDAŃSK	2014 Q2	25620.50	9017.15	783.42	2366.71
GDAŃSK	2014 Q3	25172.08	8589.63	1774.87	2403.94
GDAŃSK	2014 Q4	24770.66	9520.83	1075.02	2186.29
KATOWICE	2012 Q1	5359.47	11192.34	5717.10	1949.44
KATOWICE	2012 Q2	5132.50	10670.38	4264.42	2108.71
KATOWICE	2012 Q3	5545.65	10503.31	4926.45	1689.25
KATOWICE	2012 Q4	5321.33	12445.34	2334.35	2194.22
KATOWICE	2013 Q1	5569.73	13906.19	3145.89	1717.20
KATOWICE	2013 Q2	4777.60	17601.44	1905.82	1897.96
KATOWICE	2013 Q3	5296.91	19503.56	1971.21	1791.44
KATOWICE	2013 Q4	5033.26	20517.23	1214.88	1879.06
KATOWICE	2014 Q1	5452.20	21578.66	1528.81	1833.72
KATOWICE	2014 Q2	5286.82	19411.86	1593.14	2674.31
KATOWICE	2014 Q3	5364.17	18942.46	2179.89	2457.35
KATOWICE	2014 Q4	5337.47	18656.73	1924.79	3579.35
KRAKÓW	2012 Q1	9722.53	5450.42	1420.90	4244.00
KRAKÓW	2012 Q2	9316.02	5069.89	1616.15	4038.43
KRAKÓW	2012 Q3	10781.00	5194.84	1166.16	4036.17
KRAKÓW	2012 Q4	9790.26	4795.14	955.29	3833.74
KRAKÓW	2013 Q1	9957.43	5013.07	1632.70	3740.95

Continued on next page

Variance of point estimates of average price m2 in 12 cities between 2012 Q1 and 2014 Q4 based on NBP/CSO survey Nieruchomosci-online.pl, Dom.Gratka.pl and OtoDom.pl

City	Quarter	NBP CSO	Dom.Gratka.pl	NieOnline	OtoDom.pl
KRAKÓW	2013 Q2	9724.54	8408.56	1477.08	3815.29
KRAKÓW	2013 Q3	9698.42	6021.69	1495.36	3313.46
KRAKÓW	2013 Q4	9883.29	6409.68	1083.16	3488.95
KRAKÓW	2014 Q1	10517.50	7200.14	717.70	3353.89
KRAKÓW	2014 Q2	10332.66	7581.30	1084.53	2845.57
KRAKÓW	2014 Q3	10246.06	6873.25	1583.56	3276.73
KRAKÓW	2014 Q4	9823.54	6632.69	1430.49	3240.05
ŁÓDŹ	2012 Q1	5674.99	6839.62	689.29	989.51
ŁÓDŹ	2012 Q2	5806.84	7111.14	1078.35	750.44
ŁÓDŹ	2012 Q3	5802.47	7352.53	542.80	706.56
ŁÓDŹ	2012 Q4	5637.95	6767.80	476.04	621.15
ŁÓDŹ	2013 Q1	5911.63	6343.89	2447.52	523.27
ŁÓDŹ	2013 Q2	5935.34	6601.47	1459.04	582.16
ŁÓDŹ	2013 Q3	5770.34	7651.08	2396.98	641.23
ŁÓDŹ	2013 Q4	5671.08	7278.24	846.69	734.51
ŁÓDŹ	2014 Q1	5688.78	6680.69	731.08	750.78
ŁÓDŹ	2014 Q2	5324.40	6662.79	591.58	712.93
ŁÓDŹ	2014 Q3	5347.38	7212.57	1651.65	755.09
ŁÓDŹ	2014 Q4	5534.23	6571.01	881.70	586.66
LUBLIN	2012 Q1	20672.60	2403.36	9763.29	932.41
LUBLIN	2012 Q2	24918.78	2372.00	7929.02	722.56
LUBLIN	2012 Q3	22809.07	2153.02	4186.58	666.69
LUBLIN	2012 Q4	20747.70	1705.18	35442.22	724.98
LUBLIN	2013 Q1	20848.12	1420.08	1536.99	706.61
LUBLIN	2013 Q2	24277.24	1936.14	1228.92	708.75
LUBLIN	2013 Q3	19823.40	2217.03	1087.85	734.57
LUBLIN	2013 Q4	21838.81	2252.85	893.48	548.89
LUBLIN	2014 Q1	23290.85	2235.36	358.48	483.41
LUBLIN	2014 Q2	21651.34	2530.25	1008.76	426.48
LUBLIN	2014 Q3	21013.85	2805.78	1355.81	417.65
LUBLIN	2014 Q4	24363.75	2295.08	902.04	537.41
OLSZTYN	2012 Q1	8005.52	7942.03	5108.27	2547.64
OLSZTYN	2012 Q2	8262.44	7384.91	3065.10	2528.61
OLSZTYN	2012 Q3	8304.31	6778.46	14708.83	2193.96
OLSZTYN	2012 Q4	8340.70	5506.96	3149.57	1863.09
OLSZTYN	2013 Q1	8734.99	5608.94	3311.63	1865.20
OLSZTYN	2013 Q2	8923.50	4951.82	3052.41	1526.58
OLSZTYN	2013 Q3	8750.03	4688.47	2833.62	1263.12
OLSZTYN	2013 Q4	8589.27	4783.92	2353.30	1280.57
OLSZTYN	2014 Q1	8595.99	4557.02	727.21	1301.66
OLSZTYN	2014 Q2	7886.92	4143.24	1858.75	1555.34

Continued on next page

Variance of point estimates of average price m2 in 12 cities between 2012 Q1 and 2014 Q4 based on NBP/CSO survey Nieruchomosci-online.pl, Dom.Gratka.pl and OtoDom.pl

City	Quarter	NBP CSO	Dom.Gratka.pl	NieOnline	OtoDom.pl
OLSZTYN	2014 Q3	7638.04	4472.43	3157.52	1628.93
OLSZTYN	2014 Q4	8100.91	4493.62	1473.88	4552.63
OPOLE	2012 Q1	6876.36	4126.57	4137.83	1141.15
OPOLE	2012 Q2	7064.02	4820.12	4089.83	1067.61
OPOLE	2012 Q3	7383.37	5034.94	4180.34	969.97
OPOLE	2012 Q4	6447.30	5650.17	2151.92	1408.30
OPOLE	2013 Q1	6816.38	4938.62	1658.90	1428.74
OPOLE	2013 Q2	6793.61	5319.30	2152.20	1181.04
OPOLE	2013 Q3	7445.31	5736.15	2183.27	1534.49
OPOLE	2013 Q4	7227.08	5815.26	1628.44	1463.50
OPOLE	2014 Q1	7215.57	5621.15	769.70	1266.09
OPOLE	2014 Q2	7019.94	5946.22	925.30	1214.80
OPOLE	2014 Q3	7342.63	5573.34	2175.38	1112.27
OPOLE	2014 Q4	6808.80	3936.80	1461.28	1180.24
POZNAŃ	2012 Q1	17046.50	12112.54	1695.77	2631.74
POZNAŃ	2012 Q2	17568.79	12923.02	1864.53	2594.36
POZNAŃ	2012 Q3	16835.65	13247.10	1501.46	2438.15
POZNAŃ	2012 Q4	17126.50	12748.20	1326.54	2444.33
POZNAŃ	2013 Q1	15988.10	13385.65	1408.69	2173.18
POZNAŃ	2013 Q2	17231.07	13105.71	1822.34	1694.98
POZNAŃ	2013 Q3	16731.71	13310.66	1613.15	1498.51
POZNAŃ	2013 Q4	17633.68	14681.04	1390.65	1916.62
POZNAŃ	2014 Q1	17138.95	13058.85	645.92	2040.17
POZNAŃ	2014 Q2	16698.32	13191.82	640.16	1850.81
POZNAŃ	2014 Q3	16599.67	13433.74	862.75	1668.32
POZNAŃ	2014 Q4	17170.64	14480.52	785.60	1255.92
SZCZECIN	2012 Q1	13633.07	4628.23	2125.92	815.58
SZCZECIN	2012 Q2	13625.71	4628.13	1769.27	791.39
SZCZECIN	2012 Q3	14716.65	4399.62	1308.10	703.30
SZCZECIN	2012 Q4	14058.77	4562.75	1189.02	651.00
SZCZECIN	2013 Q1	13811.19	4693.58	1246.58	711.24
SZCZECIN	2013 Q2	13676.04	4991.28	1168.24	723.19
SZCZECIN	2013 Q3	13337.48	5848.85	1175.17	791.86
SZCZECIN	2013 Q4	13953.07	6304.74	1107.10	974.81
SZCZECIN	2014 Q1	14952.56	6863.05	301.66	951.91
SZCZECIN	2014 Q2	14312.12	6722.82	675.72	985.17
SZCZECIN	2014 Q3	14445.41	6847.04	1223.33	941.92
SZCZECIN	2014 Q4	14116.61	6706.69	741.08	857.14
WARSZAWA	2012 Q1	29657.48	11401.39	11893.63	6718.12
WARSZAWA	2012 Q2	30523.66	11962.83	17603.50	7478.19
WARSZAWA	2012 Q3	31336.81	12827.58	9302.86	7819.59
WARSZAWA	2012 Q4	30538.87	12083.45	10268.01	7450.80

Continued on next page

Variance of point estimates of average price m2 in 12 cities between 2012 Q1 and 2014 Q4 based on NBP/CSO survey Nieruchomosci-online.pl, Dom.Gratka.pl and OtoDom.pl

City	Quarter	NBP CSO	Dom.Gratka.pl	NieOnline	OtoDom.pl
WARSZAWA	2013 Q1	30265.76	11867.36	850.46	7367.25
WARSZAWA	2013 Q2	33423.05	17917.77	926.83	7510.67
WARSZAWA	2013 Q3	33081.38	12902.16	713.40	7936.53
WARSZAWA	2013 Q4	29146.23	13490.02	809.06	8221.37
WARSZAWA	2014 Q1	29108.19	14450.65	299.73	8085.45
WARSZAWA	2014 Q2	31203.93	14564.29	464.40	7834.65
WARSZAWA	2014 Q3	31691.70	12796.47	799.79	7380.39
WARSZAWA	2014 Q4	29773.53	13656.12	499.07	7676.53
WROCŁAW	2012 Q1	18979.17	4470.62	2999.01	2385.49
WROCŁAW	2012 Q2	20152.89	4803.92	1781.66	2610.24
WROCŁAW	2012 Q3	19379.97	5272.62	2384.04	2550.57
WROCŁAW	2012 Q4	19080.63	5565.76	1192.58	2282.30
WROCŁAW	2013 Q1	19130.25	5565.34	1940.08	2444.62
WROCŁAW	2013 Q2	18249.43	5716.40	1507.05	2130.39
WROCŁAW	2013 Q3	18067.37	5981.74	1851.70	2183.79
WROCŁAW	2013 Q4	17809.63	6313.25	1039.59	2601.70
WROCŁAW	2014 Q1	17405.66	6280.99	403.19	3106.28
WROCŁAW	2014 Q2	18383.27	6434.38	938.37	2693.61
WROCŁAW	2014 Q3	20758.57	6044.55	772.38	2469.81
WROCŁAW	2014 Q4	18959.74	6226.60	529.61	2545.28

TABLE A.3: Point estimates and variance of $Bias(\breve{\theta}_{kdt})$

Source	City	Quarter	Bias	SE Bias	Var Bias
Dom.Gratka.pl	BIAŁYSTOK	2012 Q1	17.20	119.96	14391.04
Dom.Gratka.pl	BIAŁYSTOK	2012 Q2	36.56	122.92	15108.41
Dom.Gratka.pl	BIAŁYSTOK	2012 Q3	92.06	120.33	14478.50
Dom.Gratka.pl	BIAŁYSTOK	2012 Q4	70.33	118.60	14065.97
Dom.Gratka.pl	BIAŁYSTOK	2013 Q1	34.43	115.35	13305.17
Dom.Gratka.pl	BIAŁYSTOK	2013 Q2	-19.13	115.98	13451.73
Dom.Gratka.pl	BIAŁYSTOK	2013 Q3	-29.51	117.71	13856.81
Dom.Gratka.pl	BIAŁYSTOK	2013 Q4	-50.17	121.07	14659.09
Dom.Gratka.pl	BIAŁYSTOK	2014 Q1	-45.56	124.33	15456.89
Dom.Gratka.pl	BIAŁYSTOK	2014 Q2	-0.96	126.32	15956.64
Dom.Gratka.pl	BIAŁYSTOK	2014 Q3	-27.51	123.25	15190.92
Dom.Gratka.pl	BIAŁYSTOK	2014 Q4	-61.91	124.73	15558.81
Dom.Gratka.pl	GDAŃSK	2012 Q1	-99.75	171.88	29542.80
Dom.Gratka.pl	GDAŃSK	2012 Q2	-90.89	174.09	30306.65
Dom.Gratka.pl	GDAŃSK	2012 Q3	-203.46	170.95	29224.72
Dom.Gratka.pl	GDAŃSK	2012 Q4	-212.36	174.32	30387.28
Dom.Gratka.pl	GDAŃSK	2013 Q1	-349.29	176.34	31097.06
Dom.Gratka.pl	GDAŃSK	2013 Q2	-317.71	177.55	31524.73
Dom.Gratka.pl	GDAŃSK	2013 Q3	-213.42	172.40	29723.11
Dom.Gratka.pl	GDAŃSK	2013 Q4	-310.02	172.37	29711.13
Dom.Gratka.pl	GDAŃSK	2014 Q1	-218.11	180.48	32572.03
Dom.Gratka.pl	GDAŃSK	2014 Q2	-227.35	179.72	32300.12
Dom.Gratka.pl	GDAŃSK	2014 Q3	-24.45	177.27	31424.18
Dom.Gratka.pl	GDAŃSK	2014 Q4	-3.14	178.76	31953.97
Dom.Gratka.pl	KATOWICE	2012 Q1	201.95	119.22	14214.28
Dom.Gratka.pl	KATOWICE	2012 Q2	96.12	116.04	13465.36
Dom.Gratka.pl	KATOWICE	2012 Q3	136.07	117.10	13711.44
Dom.Gratka.pl	KATOWICE	2012 Q4	58.46	124.21	15429.14
Dom.Gratka.pl	KATOWICE	2013 Q1	249.92	130.91	17138.39
Dom.Gratka.pl	KATOWICE	2013 Q2	235.82	141.57	20041.52
Dom.Gratka.pl	KATOWICE	2013 Q3	292.07	149.88	22462.94
Dom.Gratka.pl	KATOWICE	2013 Q4	302.83	152.36	23212.97
Dom.Gratka.pl	KATOWICE	2014 Q1	333.92	157.14	24693.33
Dom.Gratka.pl	KATOWICE	2014 Q2	221.10	149.54	22361.16
Dom.Gratka.pl	KATOWICE	2014 Q3	195.19	148.22	21969.11
Dom.Gratka.pl	KATOWICE	2014 Q4	257.39	147.16	21656.67
Dom.Gratka.pl	KRAKÓW	2012 Q1	403.78	113.29	12835.42
Dom.Gratka.pl	KRAKÓW	2012 Q2	356.18	109.77	12048.38
Dom.Gratka.pl	KRAKÓW	2012 Q3	372.88	116.78	13638.31
Dom.Gratka.pl	KRAKÓW	2012 Q4	239.63	110.67	12247.88
Dom.Gratka.pl	KRAKÓW	2013 Q1	166.40	112.40	12632.97
Dom.Gratka.pl	KRAKÓW	2013 Q2	233.51	125.68	15795.58
Dom.Gratka.pl	KRAKÓW	2013 Q3	144.99	115.68	13382.58

Continued on next page

Point estimates and variance of $Bias(\breve{\theta}_{kdt})$

Source	City	Quarter	Bias	SE Bias	Var Bias
Dom.Gratka.pl	KRAKÓW	2013 Q4	247.85	118.13	13955.44
Dom.Gratka.pl	KRAKÓW	2014 Q1	67.58	124.02	15380.11
Dom.Gratka.pl	KRAKÓW	2014 Q2	149.35	124.81	15576.44
Dom.Gratka.pl	KRAKÓW	2014 Q3	169.75	121.58	14781.79
Dom.Gratka.pl	KRAKÓW	2014 Q4	-72.98	118.82	14118.70
Dom.Gratka.pl	ŁÓDŹ	2012 Q1	387.48	100.88	10177.08
Dom.Gratka.pl	ŁÓDŹ	2012 Q2	310.12	102.86	10580.45
Dom.Gratka.pl	ŁÓDŹ	2012 Q3	504.91	104.01	10817.48
Dom.Gratka.pl	ŁÓDŹ	2012 Q4	379.50	100.34	10068.23
Dom.Gratka.pl	ŁÓDŹ	2013 Q1	213.66	99.59	9918.00
Dom.Gratka.pl	ŁÓDŹ	2013 Q2	81.86	100.99	10199.28
Dom.Gratka.pl	ŁÓDŹ	2013 Q3	110.73	105.28	11083.89
Dom.Gratka.pl	ŁÓDŹ	2013 Q4	144.14	103.01	10611.79
Dom.Gratka.pl	ŁÓDŹ	2014 Q1	132.76	100.16	10031.94
Dom.Gratka.pl	ŁÓDŹ	2014 Q2	180.05	98.23	9649.67
Dom.Gratka.pl	ŁÓDŹ	2014 Q3	125.16	101.11	10222.43
Dom.Gratka.pl	ŁÓDŹ	2014 Q4	77.77	98.83	9767.72
Dom.Gratka.pl	LUBLIN	2012 Q1	83.10	144.01	20738.43
Dom.Gratka.pl	LUBLIN	2012 Q2	40.44	157.97	24953.25
Dom.Gratka.pl	LUBLIN	2012 Q3	-5.36	150.41	22624.56
Dom.Gratka.pl	LUBLIN	2012 Q4	-17.90	141.83	20115.36
Dom.Gratka.pl	LUBLIN	2013 Q1	427.27	141.18	19930.67
Dom.Gratka.pl	LUBLIN	2013 Q2	30.40	154.52	23875.86
Dom.Gratka.pl	LUBLIN	2013 Q3	-65.92	140.37	19702.90
Dom.Gratka.pl	LUBLIN	2013 Q4	-46.68	147.49	21754.13
Dom.Gratka.pl	LUBLIN	2014 Q1	-36.74	152.28	23188.69
Dom.Gratka.pl	LUBLIN	2014 Q2	-29.78	147.80	21844.06
Dom.Gratka.pl	LUBLIN	2014 Q3	20.89	146.57	21482.11
Dom.Gratka.pl	LUBLIN	2014 Q4	-2.85	155.95	24321.31
Dom.Gratka.pl	OLSZTYN	2012 Q1	224.41	116.66	13610.02
Dom.Gratka.pl	OLSZTYN	2012 Q2	173.77	115.37	13309.82
Dom.Gratka.pl	OLSZTYN	2012 Q3	188.18	112.89	12745.24
Dom.Gratka.pl	OLSZTYN	2012 Q4	181.05	107.29	11510.13
Dom.Gratka.pl	OLSZTYN	2013 Q1	240.23	109.57	12006.40
Dom.Gratka.pl	OLSZTYN	2013 Q2	130.40	107.41	11537.80
Dom.Gratka.pl	OLSZTYN	2013 Q3	153.43	105.36	11100.98
Dom.Gratka.pl	OLSZTYN	2013 Q4	107.45	105.05	11035.67
Dom.Gratka.pl	OLSZTYN	2014 Q1	87.87	104.00	10815.48
Dom.Gratka.pl	OLSZTYN	2014 Q2	79.28	98.45	9692.64
Dom.Gratka.pl	OLSZTYN	2014 Q3	81.44	98.86	9772.95
Dom.Gratka.pl	OLSZTYN	2014 Q4	108.35	101.28	10257.00
Dom.Gratka.pl	OPOLE	2012 Q1	386.08	93.09	8665.40
Dom.Gratka.pl	OPOLE	2012 Q2	495.34	97.71	9546.61

Continued on next page

Point estimates and variance of $Bias(\breve{\theta}_{kdt})$

Source	City	Quarter	Bias	SE Bias	Var Bias
Dom.Gratka.pl	OPOLE	2012 Q3	330.40	100.40	10080.79
Dom.Gratka.pl	OPOLE	2012 Q4	348.61	98.79	9759.94
Dom.Gratka.pl	OPOLE	2013 Q1	296.27	97.04	9417.48
Dom.Gratka.pl	OPOLE	2013 Q2	304.15	98.87	9775.38
Dom.Gratka.pl	OPOLE	2013 Q3	87.14	104.13	10843.93
Dom.Gratka.pl	OPOLE	2013 Q4	168.25	103.46	10704.82
Dom.Gratka.pl	OPOLE	2014 Q1	380.77	102.47	10499.19
Dom.Gratka.pl	OPOLE	2014 Q2	138.27	103.10	10628.64
Dom.Gratka.pl	OPOLE	2014 Q3	242.25	102.85	10578.44
Dom.Gratka.pl	OPOLE	2014 Q4	252.90	91.70	8408.07
Dom.Gratka.pl	POZNAŃ	2012 Q1	353.59	163.77	26821.51
Dom.Gratka.pl	POZNAŃ	2012 Q2	372.59	167.79	28154.28
Dom.Gratka.pl	POZNAŃ	2012 Q3	352.84	166.57	27745.23
Dom.Gratka.pl	POZNAŃ	2012 Q4	344.68	165.94	27537.17
Dom.Gratka.pl	POZNAŃ	2013 Q1	337.76	164.43	27036.23
Dom.Gratka.pl	POZNAŃ	2013 Q2	22.53	167.33	27999.26
Dom.Gratka.pl	POZNAŃ	2013 Q3	31.36	166.45	27704.85
Dom.Gratka.pl	POZNAŃ	2013 Q4	-20.23	173.14	29977.20
Dom.Gratka.pl	POZNAŃ	2014 Q1	-106.54	166.91	27860.28
Dom.Gratka.pl	POZNAŃ	2014 Q2	-27.42	165.99	27552.62
Dom.Gratka.pl	POZNAŃ	2014 Q3	-102.52	166.42	27695.89
Dom.Gratka.pl	POZNAŃ	2014 Q4	-1.11	171.21	29313.63
Dom.Gratka.pl	SZCZECIN	2012 Q1	122.94	126.19	15923.77
Dom.Gratka.pl	SZCZECIN	2012 Q2	209.13	126.16	15916.32
Dom.Gratka.pl	SZCZECIN	2012 Q3	126.23	129.53	16778.75
Dom.Gratka.pl	SZCZECIN	2012 Q4	165.84	127.61	16283.99
Dom.Gratka.pl	SZCZECIN	2013 Q1	13.59	127.15	16167.25
Dom.Gratka.pl	SZCZECIN	2013 Q2	204.35	127.79	16329.80
Dom.Gratka.pl	SZCZECIN	2013 Q3	198.65	129.80	16848.80
Dom.Gratka.pl	SZCZECIN	2013 Q4	74.46	133.87	17920.28
Dom.Gratka.pl	SZCZECIN	2014 Q1	-101.47	139.56	19478.08
Dom.Gratka.pl	SZCZECIN	2014 Q2	-44.94	136.74	18697.42
Dom.Gratka.pl	SZCZECIN	2014 Q3	-9.62	137.68	18954.93
Dom.Gratka.pl	SZCZECIN	2014 Q4	-69.93	135.96	18485.77
Dom.Gratka.pl	WARSZAWA	2012 Q1	-95.14	196.78	38721.35
Dom.Gratka.pl	WARSZAWA	2012 Q2	-90.33	200.37	40148.96
Dom.Gratka.pl	WARSZAWA	2012 Q3	-43.14	204.52	41826.87
Dom.Gratka.pl	WARSZAWA	2012 Q4	-42.24	200.71	40284.79
Dom.Gratka.pl	WARSZAWA	2013 Q1	47.95	199.49	39795.60
Dom.Gratka.pl	WARSZAWA	2013 Q2	-105.77	221.37	49003.29
Dom.Gratka.pl	WARSZAWA	2013 Q3	-39.94	208.92	43646.01
Dom.Gratka.pl	WARSZAWA	2013 Q4	-72.92	200.75	40298.73
Dom.Gratka.pl	WARSZAWA	2014 Q1	-4.86	203.03	41221.32

Continued on next page

Point estimates and variance of $Bias(\breve{\theta}_{kdt})$

Source	City	Quarter	Bias	SE Bias	Var Bias
Dom.Gratka.pl	WARSZAWA	2014 Q2	-105.65	208.40	43430.69
Dom.Gratka.pl	WARSZAWA	2014 Q3	-78.73	205.31	42150.64
Dom.Gratka.pl	WARSZAWA	2014 Q4	-47.12	202.71	41092.12
Dom.Gratka.pl	WROCŁAW	2012 Q1	164.36	145.30	21112.26
Dom.Gratka.pl	WROCŁAW	2012 Q2	119.91	150.40	22619.28
Dom.Gratka.pl	WROCŁAW	2012 Q3	140.86	149.38	22315.06
Dom.Gratka.pl	WROCŁAW	2012 Q4	97.15	149.36	22308.87
Dom.Gratka.pl	WROCŁAW	2013 Q1	108.49	149.53	22358.07
Dom.Gratka.pl	WROCŁAW	2013 Q2	26.16	147.07	21628.31
Dom.Gratka.pl	WROCŁAW	2013 Q3	-40.65	147.35	21711.59
Dom.Gratka.pl	WROCŁAW	2013 Q4	-169.17	147.60	21785.35
Dom.Gratka.pl	WROCŁAW	2014 Q1	-197.76	146.11	21349.12
Dom.Gratka.pl	WROCŁAW	2014 Q2	-2.37	149.93	22480.12
Dom.Gratka.pl	WROCŁAW	2014 Q3	-122.99	156.41	24465.59
Dom.Gratka.pl	WROCŁAW	2014 Q4	-180.59	151.16	22848.82
Nieruchomosci-online.pl	BIAŁYSTOK	2012 Q1	-214.88	93.92	8821.26
Nieruchomosci-online.pl	BIAŁYSTOK	2012 Q2	-378.24	75.95	5767.76
Nieruchomosci-online.pl	BIAŁYSTOK	2012 Q3	-2.04	81.41	6627.99
Nieruchomosci-online.pl	BIAŁYSTOK	2012 Q4	-307.31	71.70	5140.28
Nieruchomosci-online.pl	BIAŁYSTOK	2013 Q1	-373.11	56.59	3202.46
Nieruchomosci-online.pl	BIAŁYSTOK	2013 Q2	-278.26	58.43	3413.92
Nieruchomosci-online.pl	BIAŁYSTOK	2013 Q3	-299.10	58.36	3405.61
Nieruchomosci-online.pl	BIAŁYSTOK	2013 Q4	-267.51	59.46	3535.49
Nieruchomosci-online.pl	BIAŁYSTOK	2014 Q1	-223.74	55.77	3110.36
Nieruchomosci-online.pl	BIAŁYSTOK	2014 Q2	-203.63	58.64	3438.44
Nieruchomosci-online.pl	BIAŁYSTOK	2014 Q3	-89.60	60.94	3713.90
Nieruchomosci-online.pl	BIAŁYSTOK	2014 Q4	-227.48	58.90	3468.78
Nieruchomosci-online.pl	GDAŃSK	2012 Q1	-205.07	170.55	29087.89
Nieruchomosci-online.pl	GDAŃSK	2012 Q2	-164.59	162.28	26336.37
Nieruchomosci-online.pl	GDAŃSK	2012 Q3	13.41	160.02	25605.97
Nieruchomosci-online.pl	GDAŃSK	2012 Q4	-462.37	158.11	24999.36
Nieruchomosci-online.pl	GDAŃSK	2013 Q1	-648.51	154.15	23761.33
Nieruchomosci-online.pl	GDAŃSK	2013 Q2	-680.52	153.15	23454.39
Nieruchomosci-online.pl	GDAŃSK	2013 Q3	-418.78	151.88	23067.78
Nieruchomosci-online.pl	GDAŃSK	2013 Q4	-723.74	147.52	21761.80
Nieruchomosci-online.pl	GDAŃSK	2014 Q1	-519.20	144.72	20945.23
Nieruchomosci-online.pl	GDAŃSK	2014 Q2	-472.37	146.45	21448.96
Nieruchomosci-online.pl	GDAŃSK	2014 Q3	-306.19	148.30	21991.99
Nieruchomosci-online.pl	GDAŃSK	2014 Q4	-300.62	144.54	20890.72
Nieruchomosci-online.pl	KATOWICE	2012 Q1	-213.26	78.24	6121.60
Nieruchomosci-online.pl	KATOWICE	2012 Q2	-464.26	66.65	4441.95
Nieruchomosci-online.pl	KATOWICE	2012 Q3	-209.17	74.28	5517.14
Nieruchomosci-online.pl	KATOWICE	2012 Q4	-423.18	51.97	2700.71

Continued on next page

Point estimates and variance of $Bias(\breve{\theta}_{kdt})$

Source	City	Quarter	Bias	SE Bias	Var Bias
Nieruchomosci-online.pl	KATOWICE	2013 Q1	-191.37	61.32	3760.66
Nieruchomosci-online.pl	KATOWICE	2013 Q2	-355.30	41.57	1728.46
Nieruchomosci-online.pl	KATOWICE	2013 Q3	-238.40	48.10	2313.16
Nieruchomosci-online.pl	KATOWICE	2013 Q4	-319.93	35.96	1293.18
Nieruchomosci-online.pl	KATOWICE	2014 Q1	40.44	45.01	2026.05
Nieruchomosci-online.pl	KATOWICE	2014 Q2	-243.35	43.87	1925.00
Nieruchomosci-online.pl	KATOWICE	2014 Q3	-183.26	50.88	2589.10
Nieruchomosci-online.pl	KATOWICE	2014 Q4	-162.89	48.03	2307.29
Nieruchomosci-online.pl	KRAKÓW	2012 Q1	163.12	78.67	6188.47
Nieruchomosci-online.pl	KRAKÓW	2012 Q2	282.81	77.31	5977.21
Nieruchomosci-online.pl	KRAKÓW	2012 Q3	300.53	83.62	6992.19
Nieruchomosci-online.pl	KRAKÓW	2012 Q4	180.51	76.10	5790.59
Nieruchomosci-online.pl	KRAKÓW	2013 Q1	-72.83	81.46	6635.17
Nieruchomosci-online.pl	KRAKÓW	2013 Q2	35.64	79.04	6246.66
Nieruchomosci-online.pl	KRAKÓW	2013 Q3	226.22	78.99	6238.81
Nieruchomosci-online.pl	KRAKÓW	2013 Q4	12.25	77.53	6011.48
Nieruchomosci-online.pl	KRAKÓW	2014 Q1	150.54	79.25	6280.24
Nieruchomosci-online.pl	KRAKÓW	2014 Q2	-30.34	80.39	6462.23
Nieruchomosci-online.pl	KRAKÓW	2014 Q3	2.27	82.91	6874.66
Nieruchomosci-online.pl	KRAKÓW	2014 Q4	-197.31	79.37	6299.06
Nieruchomosci-online.pl	ŁÓDŹ	2012 Q1	-126.27	37.54	1409.32
Nieruchomosci-online.pl	ŁÓDŹ	2012 Q2	-202.10	43.93	1930.22
Nieruchomosci-online.pl	ŁÓDŹ	2012 Q3	5.88	37.29	1390.31
Nieruchomosci-online.pl	ŁÓDŹ	2012 Q4	-9.63	34.04	1159.03
Nieruchomosci-online.pl	ŁÓDŹ	2013 Q1	-166.89	58.35	3404.19
Nieruchomosci-online.pl	ŁÓDŹ	2013 Q2	-329.01	49.39	2439.42
Nieruchomosci-online.pl	ŁÓDŹ	2013 Q3	-159.80	56.68	3212.35
Nieruchomosci-online.pl	ŁÓDŹ	2013 Q4	-254.57	39.53	1562.80
Nieruchomosci-online.pl	ŁÓDŹ	2014 Q1	-41.40	38.27	1464.89
Nieruchomosci-online.pl	ŁÓDŹ	2014 Q2	-120.84	31.00	961.02
Nieruchomosci-online.pl	ŁÓDŹ	2014 Q3	17.65	45.21	2044.07
Nieruchomosci-online.pl	ŁÓDŹ	2014 Q4	-118.19	38.22	1460.97
Nieruchomosci-online.pl	LUBLIN	2012 Q1	-11.82	159.63	25480.92
Nieruchomosci-online.pl	LUBLIN	2012 Q2	-66.05	167.01	27892.83
Nieruchomosci-online.pl	LUBLIN	2012 Q3	-228.33	148.46	22040.69
Nieruchomosci-online.pl	LUBLIN	2012 Q4	92.69	226.35	51234.97
Nieruchomosci-online.pl	LUBLIN	2013 Q1	152.13	132.02	17430.15
Nieruchomosci-online.pl	LUBLIN	2013 Q2	-132.02	143.36	20551.20
Nieruchomosci-online.pl	LUBLIN	2013 Q3	-233.17	126.32	15956.29
Nieruchomosci-online.pl	LUBLIN	2013 Q4	-181.94	133.33	17777.32
Nieruchomosci-online.pl	LUBLIN	2014 Q1	-97.46	136.73	18694.37
Nieruchomosci-online.pl	LUBLIN	2014 Q2	-111.12	133.06	17705.14
Nieruchomosci-online.pl	LUBLIN	2014 Q3	-144.22	131.96	17414.70

Continued on next page

Point estimates and variance of $Bias(\breve{\theta}_{kdt})$

Source	City	Quarter	Bias	SE Bias	Var Bias
Nieruchomosci-online.pl	LUBLIN	2014 Q4	-65.42	142.52	20310.84
Nieruchomosci-online.pl	OLSZTYN	2012 Q1	-226.68	90.33	8158.82
Nieruchomosci-online.pl	OLSZTYN	2012 Q2	-198.10	79.83	6372.57
Nieruchomosci-online.pl	OLSZTYN	2012 Q3	298.69	134.38	18058.18
Nieruchomosci-online.pl	OLSZTYN	2012 Q4	1.53	80.84	6535.31
Nieruchomosci-online.pl	OLSZTYN	2013 Q1	-77.40	84.21	7091.65
Nieruchomosci-online.pl	OLSZTYN	2013 Q2	-189.15	83.79	7020.95
Nieruchomosci-online.pl	OLSZTYN	2013 Q3	-74.11	81.42	6628.69
Nieruchomosci-online.pl	OLSZTYN	2013 Q4	27.92	77.38	5987.62
Nieruchomosci-online.pl	OLSZTYN	2014 Q1	-244.27	66.09	4368.24
Nieruchomosci-online.pl	OLSZTYN	2014 Q2	-302.48	69.21	4790.71
Nieruchomosci-online.pl	OLSZTYN	2014 Q3	-192.65	76.42	5840.60
Nieruchomosci-online.pl	OLSZTYN	2014 Q4	-153.12	67.97	4619.82
Nieruchomosci-online.pl	OPOLE	2012 Q1	5.43	77.84	6059.23
Nieruchomosci-online.pl	OPOLE	2012 Q2	-41.95	78.73	6198.89
Nieruchomosci-online.pl	OPOLE	2012 Q3	31.92	81.29	6608.75
Nieruchomosci-online.pl	OPOLE	2012 Q4	54.49	60.37	3644.26
Nieruchomosci-online.pl	OPOLE	2013 Q1	45.79	59.33	3520.32
Nieruchomosci-online.pl	OPOLE	2013 Q2	-83.44	63.17	3990.85
Nieruchomosci-online.pl	OPOLE	2013 Q3	-290.62	68.36	4673.62
Nieruchomosci-online.pl	OPOLE	2013 Q4	-125.38	62.45	3900.56
Nieruchomosci-online.pl	OPOLE	2014 Q1	159.03	55.05	3030.31
Nieruchomosci-online.pl	OPOLE	2014 Q2	-169.32	54.68	2990.28
Nieruchomosci-online.pl	OPOLE	2014 Q3	59.32	67.55	4563.04
Nieruchomosci-online.pl	OPOLE	2014 Q4	52.12	57.58	3315.12
Nieruchomosci-online.pl	POZNAŃ	2012 Q1	-427.13	117.42	13787.31
Nieruchomosci-online.pl	POZNAŃ	2012 Q2	-288.23	120.33	14478.35
Nieruchomosci-online.pl	POZNAŃ	2012 Q3	-326.22	115.68	13382.16
Nieruchomosci-online.pl	POZNAŃ	2012 Q4	-236.43	116.18	13498.08
Nieruchomosci-online.pl	POZNAŃ	2013 Q1	-198.27	111.54	12441.83
Nieruchomosci-online.pl	POZNAŃ	2013 Q2	-292.65	118.74	14098.45
Nieruchomosci-online.pl	POZNAŃ	2013 Q3	-333.45	115.71	13389.90
Nieruchomosci-online.pl	POZNAŃ	2013 Q4	-465.29	118.61	14069.38
Nieruchomosci-online.pl	POZNAŃ	2014 Q1	-384.53	113.27	12829.91
Nieruchomosci-online.pl	POZNAŃ	2014 Q2	-371.96	111.28	12383.52
Nieruchomosci-online.pl	POZNAŃ	2014 Q3	-359.31	111.84	12507.46
Nieruchomosci-online.pl	POZNAŃ	2014 Q4	-308.53	114.02	13001.27
Nieruchomosci-online.pl	SZCZECIN	2012 Q1	-201.45	103.94	10804.03
Nieruchomosci-online.pl	SZCZECIN	2012 Q2	-63.73	102.18	10440.01
Nieruchomosci-online.pl	SZCZECIN	2012 Q3	-112.24	105.21	11069.78
Nieruchomosci-online.pl	SZCZECIN	2012 Q4	-13.52	101.45	10292.83
Nieruchomosci-online.pl	SZCZECIN	2013 Q1	-402.87	100.51	10102.81
Nieruchomosci-online.pl	SZCZECIN	2013 Q2	-137.38	99.45	9889.32

Continued on next page

Point estimates and variance of $Bias(\breve{\theta}_{kdt})$

Source	City	Quarter	Bias	SE Bias	Var Bias
Nieruchomosci-online.pl	SZCZECIN	2013 Q3	-120.91	97.76	9557.69
Nieruchomosci-online.pl	SZCZECIN	2013 Q4	-196.60	100.52	10105.20
Nieruchomosci-online.pl	SZCZECIN	2014 Q1	-291.37	101.49	10299.26
Nieruchomosci-online.pl	SZCZECIN	2014 Q2	-267.21	100.16	10032.88
Nieruchomosci-online.pl	SZCZECIN	2014 Q3	-216.90	103.51	10713.79
Nieruchomosci-online.pl	SZCZECIN	2014 Q4	-276.04	99.51	9902.73
Nieruchomosci-online.pl	WARSZAWA	2012 Q1	-1036.78	191.30	36596.15
Nieruchomosci-online.pl	WARSZAWA	2012 Q2	-1062.60	207.78	43172.20
Nieruchomosci-online.pl	WARSZAWA	2012 Q3	-812.31	188.90	35684.71
Nieruchomosci-online.pl	WARSZAWA	2012 Q4	-721.03	189.35	35851.92
Nieruchomosci-online.pl	WARSZAWA	2013 Q1	-386.26	161.74	26161.26
Nieruchomosci-online.pl	WARSZAWA	2013 Q2	-440.26	171.45	29394.92
Nieruchomosci-online.pl	WARSZAWA	2013 Q3	-375.31	169.82	28839.82
Nieruchomosci-online.pl	WARSZAWA	2013 Q4	-572.92	158.11	25000.33
Nieruchomosci-online.pl	WARSZAWA	2014 Q1	-103.52	156.37	24452.96
Nieruchomosci-online.pl	WARSZAWA	2014 Q2	-456.15	163.44	26713.37
Nieruchomosci-online.pl	WARSZAWA	2014 Q3	-265.41	165.94	27536.53
Nieruchomosci-online.pl	WARSZAWA	2014 Q4	-354.80	159.12	25317.63
Nieruchomosci-online.pl	WROCŁAW	2012 Q1	-215.47	130.47	17023.21
Nieruchomosci-online.pl	WROCŁAW	2012 Q2	-277.86	130.31	16979.58
Nieruchomosci-online.pl	WROCŁAW	2012 Q3	-69.37	129.65	16809.04
Nieruchomosci-online.pl	WROCŁAW	2012 Q4	-152.35	123.77	15318.25
Nieruchomosci-online.pl	WROCŁAW	2013 Q1	-457.90	126.95	16115.37
Nieruchomosci-online.pl	WROCŁAW	2013 Q2	-530.35	121.66	14801.51
Nieruchomosci-online.pl	WROCŁAW	2013 Q3	-573.77	122.33	14964.12
Nieruchomosci-online.pl	WROCŁAW	2013 Q4	-460.01	117.87	13894.26
Nieruchomosci-online.pl	WROCŁAW	2014 Q1	-341.52	113.37	12853.89
Nieruchomosci-online.pl	WROCŁAW	2014 Q2	-125.82	119.86	14366.67
Nieruchomosci-online.pl	WROCŁAW	2014 Q3	-483.75	128.75	16575.99
Nieruchomosci-online.pl	WROCŁAW	2014 Q4	-450.96	120.56	14534.39
OtoDom.pl	BIAŁYSTOK	2012 Q1	-35.45	85.91	7379.83
OtoDom.pl	BIAŁYSTOK	2012 Q2	-43.31	81.77	6685.67
OtoDom.pl	BIAŁYSTOK	2012 Q3	5.04	81.89	6705.63
OtoDom.pl	BIAŁYSTOK	2012 Q4	-32.61	81.24	6599.14
OtoDom.pl	BIAŁYSTOK	2013 Q1	-40.13	82.08	6737.38
OtoDom.pl	BIAŁYSTOK	2013 Q2	-80.56	82.71	6841.26
OtoDom.pl	BIAŁYSTOK	2013 Q3	-78.94	81.27	6604.22
OtoDom.pl	BIAŁYSTOK	2013 Q4	-77.71	83.50	6972.58
OtoDom.pl	BIAŁYSTOK	2014 Q1	-86.71	87.56	7667.52
OtoDom.pl	BIAŁYSTOK	2014 Q2	-73.52	87.42	7641.56
OtoDom.pl	BIAŁYSTOK	2014 Q3	-71.31	85.78	7357.44
OtoDom.pl	BIAŁYSTOK	2014 Q4	-77.64	90.03	8105.54
OtoDom.pl	GDAŃSK	2012 Q1	-176.69	164.72	27133.72
OtoDom.pl	GDAŃSK	2012 Q2	-187.42	166.19	27617.88

Continued on next page

Point estimates and variance of $Bias(\breve{\theta}_{kdt})$

Source	City	Quarter	Bias	SE Bias	Var Bias
OtoDom.pl	GDAŃSK	2012 Q3	-180.12	165.96	27542.97
OtoDom.pl	GDAŃSK	2012 Q4	-268.02	181.75	33032.71
OtoDom.pl	GDAŃSK	2013 Q1	-545.11	188.41	35496.54
OtoDom.pl	GDAŃSK	2013 Q2	-407.73	165.19	27289.31
OtoDom.pl	GDAŃSK	2013 Q3	-278.52	160.90	25888.85
OtoDom.pl	GDAŃSK	2013 Q4	-381.24	158.33	25066.81
OtoDom.pl	GDAŃSK	2014 Q1	-303.67	160.40	25728.85
OtoDom.pl	GDAŃSK	2014 Q2	-278.94	160.83	25865.05
OtoDom.pl	GDAŃSK	2014 Q3	-66.36	159.54	25453.86
OtoDom.pl	GDAŃSK	2014 Q4	-69.05	157.59	24834.80
OtoDom.pl	KATOWICE	2012 Q1	-16.17	72.02	5186.75
OtoDom.pl	KATOWICE	2012 Q2	-129.49	71.55	5119.05
OtoDom.pl	KATOWICE	2012 Q3	-98.11	71.50	5112.75
OtoDom.pl	KATOWICE	2012 Q4	-200.76	73.44	5393.40
OtoDom.pl	KATOWICE	2013 Q1	-53.73	71.87	5164.78
OtoDom.pl	KATOWICE	2013 Q2	-80.03	67.48	4553.41
OtoDom.pl	KATOWICE	2013 Q3	-74.60	70.47	4966.19
OtoDom.pl	KATOWICE	2013 Q4	-43.92	69.21	4790.16
OtoDom.pl	KATOWICE	2014 Q1	2.51	71.86	5163.77
OtoDom.pl	KATOWICE	2014 Q2	-86.66	76.41	5838.98
OtoDom.pl	KATOWICE	2014 Q3	-92.93	75.49	5699.36
OtoDom.pl	KATOWICE	2014 Q4	73.65	82.43	6794.67
OtoDom.pl	KRAKÓW	2012 Q1	461.44	108.83	11844.37
OtoDom.pl	KRAKÓW	2012 Q2	386.24	105.98	11232.29
OtoDom.pl	KRAKÓW	2012 Q3	393.57	112.67	12695.01
OtoDom.pl	KRAKÓW	2012 Q4	302.47	107.25	11501.85
OtoDom.pl	KRAKÓW	2013 Q1	182.60	107.59	11576.23
OtoDom.pl	KRAKÓW	2013 Q2	263.50	106.85	11417.68
OtoDom.pl	KRAKÓW	2013 Q3	141.00	104.35	10889.72
OtoDom.pl	KRAKÓW	2013 Q4	259.38	106.07	11250.09
OtoDom.pl	KRAKÓW	2014 Q1	32.36	108.39	11749.23
OtoDom.pl	KRAKÓW	2014 Q2	137.36	105.15	11056.08
OtoDom.pl	KRAKÓW	2014 Q3	199.17	106.77	11400.64
OtoDom.pl	KRAKÓW	2014 Q4	-58.03	104.60	10941.43
OtoDom.pl	ŁÓDŹ	2012 Q1	111.82	67.40	4542.35
OtoDom.pl	ŁÓDŹ	2012 Q2	-2.71	66.60	4435.12
OtoDom.pl	ŁÓDŹ	2012 Q3	156.70	66.23	4386.88
OtoDom.pl	ŁÓDŹ	2012 Q4	30.08	64.32	4136.95
OtoDom.pl	ŁÓDŹ	2013 Q1	-166.88	65.67	4312.74
OtoDom.pl	ŁÓDŹ	2013 Q2	-272.29	66.30	4395.34
OtoDom.pl	ŁÓDŹ	2013 Q3	-273.70	65.49	4289.41
OtoDom.pl	ŁÓDŹ	2013 Q4	-245.46	65.45	4283.43

Continued on next page

Point estimates and variance of $Bias(\breve{\theta}_{kdt})$

Source	City	Quarter	Bias	SE Bias	Var Bias
OtoDom.pl	ŁÓDŹ	2014 Q1	-260.31	65.71	4317.41
OtoDom.pl	ŁÓDŹ	2014 Q2	-197.20	62.57	3915.18
OtoDom.pl	ŁÓDŹ	2014 Q3	-254.49	63.09	3980.32
OtoDom.pl	ŁÓDŹ	2014 Q4	-208.71	63.24	3998.74
OtoDom.pl	LUBLIN	2012 Q1	62.96	139.58	19482.85
OtoDom.pl	LUBLIN	2012 Q2	1.79	153.36	23519.18
OtoDom.pl	LUBLIN	2012 Q3	-29.58	146.13	21353.60
OtoDom.pl	LUBLIN	2012 Q4	-42.21	139.11	19350.53
OtoDom.pl	LUBLIN	2013 Q1	389.27	139.40	19432.57
OtoDom.pl	LUBLIN	2013 Q2	2.00	151.21	22863.84
OtoDom.pl	LUBLIN	2013 Q3	-103.32	135.78	18435.81
OtoDom.pl	LUBLIN	2013 Q4	-76.33	142.36	20265.54
OtoDom.pl	LUBLIN	2014 Q1	-44.28	147.15	21652.11
OtoDom.pl	LUBLIN	2014 Q2	-46.37	141.26	19955.67
OtoDom.pl	LUBLIN	2014 Q3	-3.52	138.96	19309.34
OtoDom.pl	LUBLIN	2014 Q4	33.83	150.93	22779.01
OtoDom.pl	OLSZTYN	2012 Q1	-30.99	91.82	8431.00
OtoDom.pl	OLSZTYN	2012 Q2	-30.06	93.11	8668.89
OtoDom.pl	OLSZTYN	2012 Q3	-1.14	91.52	8376.11
OtoDom.pl	OLSZTYN	2012 Q4	28.23	89.90	8081.63
OtoDom.pl	OLSZTYN	2013 Q1	78.62	92.08	8478.04
OtoDom.pl	OLSZTYN	2013 Q2	11.29	91.26	8327.93
OtoDom.pl	OLSZTYN	2013 Q3	8.90	88.83	7891.00
OtoDom.pl	OLSZTYN	2013 Q4	-42.33	88.02	7747.69
OtoDom.pl	OLSZTYN	2014 Q1	-63.05	88.18	7775.49
OtoDom.pl	OLSZTYN	2014 Q2	-9.15	85.56	7320.10
OtoDom.pl	OLSZTYN	2014 Q3	19.62	84.53	7144.81
OtoDom.pl	OLSZTYN	2014 Q4	103.65	102.62	10531.38
OtoDom.pl	OPOLE	2012 Q1	190.74	76.78	5895.35
OtoDom.pl	OPOLE	2012 Q2	271.48	77.52	6009.48
OtoDom.pl	OPOLE	2012 Q3	162.76	78.94	6231.19
OtoDom.pl	OPOLE	2012 Q4	203.46	75.72	5733.44
OtoDom.pl	OPOLE	2013 Q1	159.42	78.25	6122.97
OtoDom.pl	OPOLE	2013 Q2	182.25	76.50	5852.50
OtoDom.pl	OPOLE	2013 Q3	-7.46	82.81	6857.64
OtoDom.pl	OPOLE	2013 Q4	48.08	81.05	6568.43
OtoDom.pl	OPOLE	2014 Q1	258.97	79.75	6359.50
OtoDom.pl	OPOLE	2014 Q2	22.83	78.18	6112.59
OtoDom.pl	OPOLE	2014 Q3	130.23	79.58	6332.75
OtoDom.pl	OPOLE	2014 Q4	110.86	76.60	5866.89
OtoDom.pl	POZNAŃ	2012 Q1	128.41	132.50	17556.08
OtoDom.pl	POZNAŃ	2012 Q2	147.29	134.32	18040.99
OtoDom.pl	POZNAŃ	2012 Q3	107.63	130.96	17151.65
OtoDom.pl	POZNAŃ	2012 Q4	84.93	132.09	17448.67

Continued on next page

Point estimates and variance of $Bias(\breve{\theta}_{kdt})$

Source	City	Quarter	Bias	SE Bias	Var Bias
OtoDom.pl	POZNAŃ	2013 Q1	57.80	126.65	16039.13
OtoDom.pl	POZNAŃ	2013 Q2	-241.97	129.63	16803.90
OtoDom.pl	POZNAŃ	2013 Q3	-196.74	126.92	16108.07
OtoDom.pl	POZNAŃ	2013 Q4	-207.52	132.02	17428.15
OtoDom.pl	POZNAŃ	2014 Q1	-329.34	130.60	17056.97
OtoDom.pl	POZNAŃ	2014 Q2	-227.60	128.17	16426.98
OtoDom.pl	POZNAŃ	2014 Q3	-301.42	127.07	16145.84
OtoDom.pl	POZNAŃ	2014 Q4	-131.23	127.69	16304.41
OtoDom.pl	SZCZECIN	2012 Q1	-35.79	111.02	12326.49
OtoDom.pl	SZCZECIN	2012 Q2	64.04	110.88	12294.94
OtoDom.pl	SZCZECIN	2012 Q3	-20.37	115.32	13297.80
OtoDom.pl	SZCZECIN	2012 Q4	23.31	112.19	12587.61
OtoDom.pl	SZCZECIN	2013 Q1	-129.36	111.36	12400.27
OtoDom.pl	SZCZECIN	2013 Q2	81.10	110.80	12277.08
OtoDom.pl	SZCZECIN	2013 Q3	59.44	109.58	12007.19
OtoDom.pl	SZCZECIN	2013 Q4	-55.94	113.16	12805.73
OtoDom.pl	SZCZECIN	2014 Q1	-201.98	117.40	13782.32
OtoDom.pl	SZCZECIN	2014 Q2	-125.11	114.78	13175.14
OtoDom.pl	SZCZECIN	2014 Q3	-101.99	115.17	13265.17
OtoDom.pl	SZCZECIN	2014 Q4	-145.06	113.36	12851.59
OtoDom.pl	WARSZAWA	2012 Q1	-326.22	185.08	34253.45
OtoDom.pl	WARSZAWA	2012 Q2	-318.57	189.42	35879.70
OtoDom.pl	WARSZAWA	2012 Q3	-264.81	192.44	37034.25
OtoDom.pl	WARSZAWA	2012 Q4	-240.09	189.39	35867.51
OtoDom.pl	WARSZAWA	2013 Q1	-193.58	188.44	35510.86
OtoDom.pl	WARSZAWA	2013 Q2	-327.28	197.01	38811.57
OtoDom.pl	WARSZAWA	2013 Q3	-239.88	197.22	38895.75
OtoDom.pl	WARSZAWA	2013 Q4	-313.18	187.74	35245.45
OtoDom.pl	WARSZAWA	2014 Q1	-266.98	187.27	35071.49
OtoDom.pl	WARSZAWA	2014 Q2	-324.57	192.14	36916.43
OtoDom.pl	WARSZAWA	2014 Q3	-229.55	192.22	36949.93
OtoDom.pl	WARSZAWA	2014 Q4	-208.15	187.96	35327.90
OtoDom.pl	WROCŁAW	2012 Q1	32.31	138.72	19242.50
OtoDom.pl	WROCŁAW	2012 Q2	-42.34	143.67	20640.97
OtoDom.pl	WROCŁAW	2012 Q3	-48.75	140.74	19808.38
OtoDom.pl	WROCŁAW	2012 Q4	-124.99	138.71	19240.77
OtoDom.pl	WROCŁAW	2013 Q1	-85.43	139.47	19452.71
OtoDom.pl	WROCŁAW	2013 Q2	-175.59	135.12	18257.66
OtoDom.pl	WROCŁAW	2013 Q3	-212.68	134.64	18129.01
OtoDom.pl	WROCŁAW	2013 Q4	-354.63	135.24	18289.18
OtoDom.pl	WROCŁAW	2014 Q1	-315.07	135.61	18389.79
OtoDom.pl	WROCŁAW	2014 Q2	-124.83	137.68	18954.73
OtoDom.pl	WROCŁAW	2014 Q3	-212.38	145.28	21106.22

Continued on next page

Point estimates and variance of $Bias(\breve{\theta}_{kdt})$

Source	City	Quarter	Bias	SE Bias	Var Bias
OtoDom.pl	WROCŁAW	2014 Q4	-260.56	139.22	19382.87

A.1.2 Description of data and model

TABLE A.4: Correlation between bias and cities

City	OtoDom_Gratka	Gratka_NieOnline	OtoDom_NieOnline
BIAŁYSTOK	0.93	0.10	0.26
GDAŃSK	0.94	0.69	0.77
KATOWICE	0.74	0.66	0.68
KRAKÓW	0.99	0.69	0.64
ŁÓDŹ	0.97	0.42	0.36
LUBLIN	0.99	0.70	0.72
OLSZTYN	0.19	0.37	0.11
OPOLE	0.97	0.66	0.74
POZNAŃ	0.98	0.46	0.39
SZCZECIN	0.98	0.79	0.82
WARSZAWA	0.77	0.39	0.48
WROCŁAW	0.94	0.52	0.56

A.2 Distribution of number of rooms for all domains

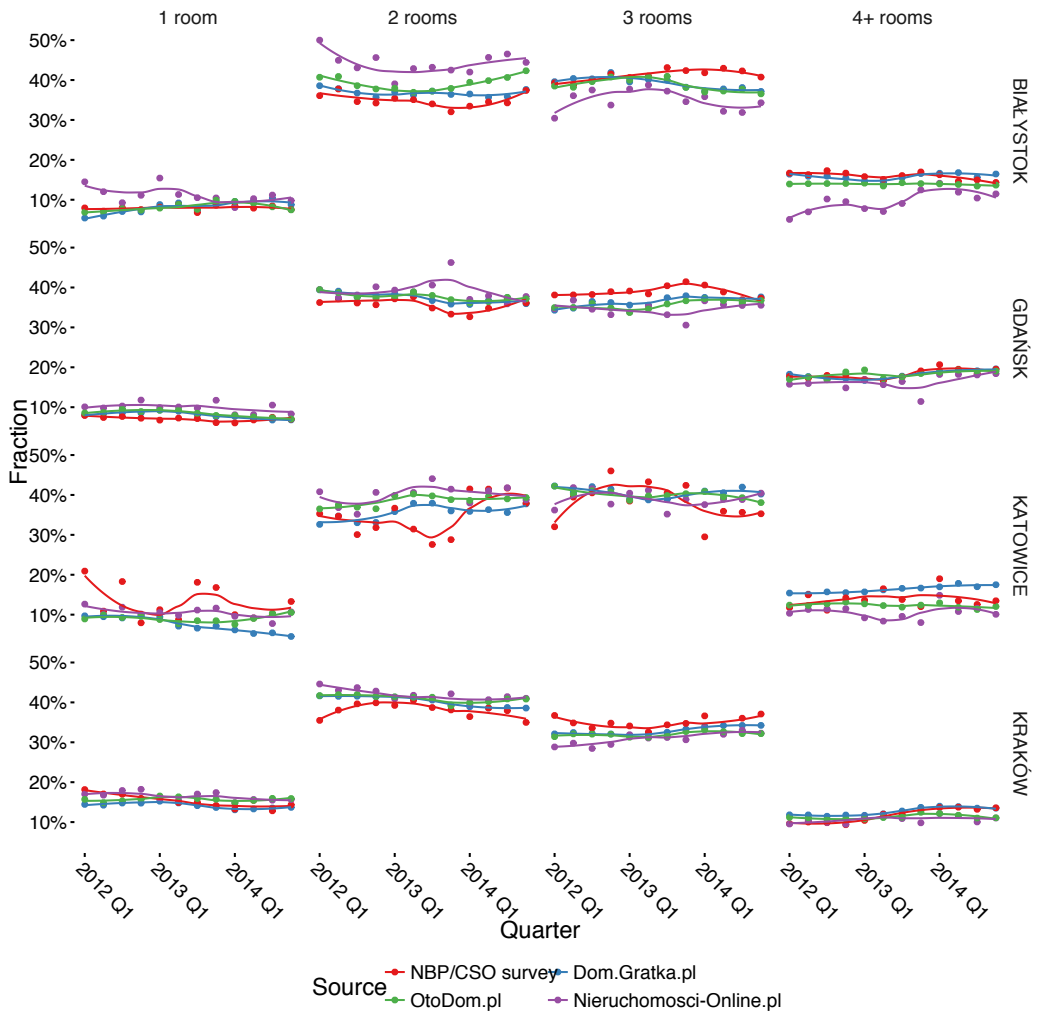


FIGURE A.1: Comparison of number of rooms distribution in NBP/CSO survey and IDS in Białystok, Gdańsk, Katowice and Kraków

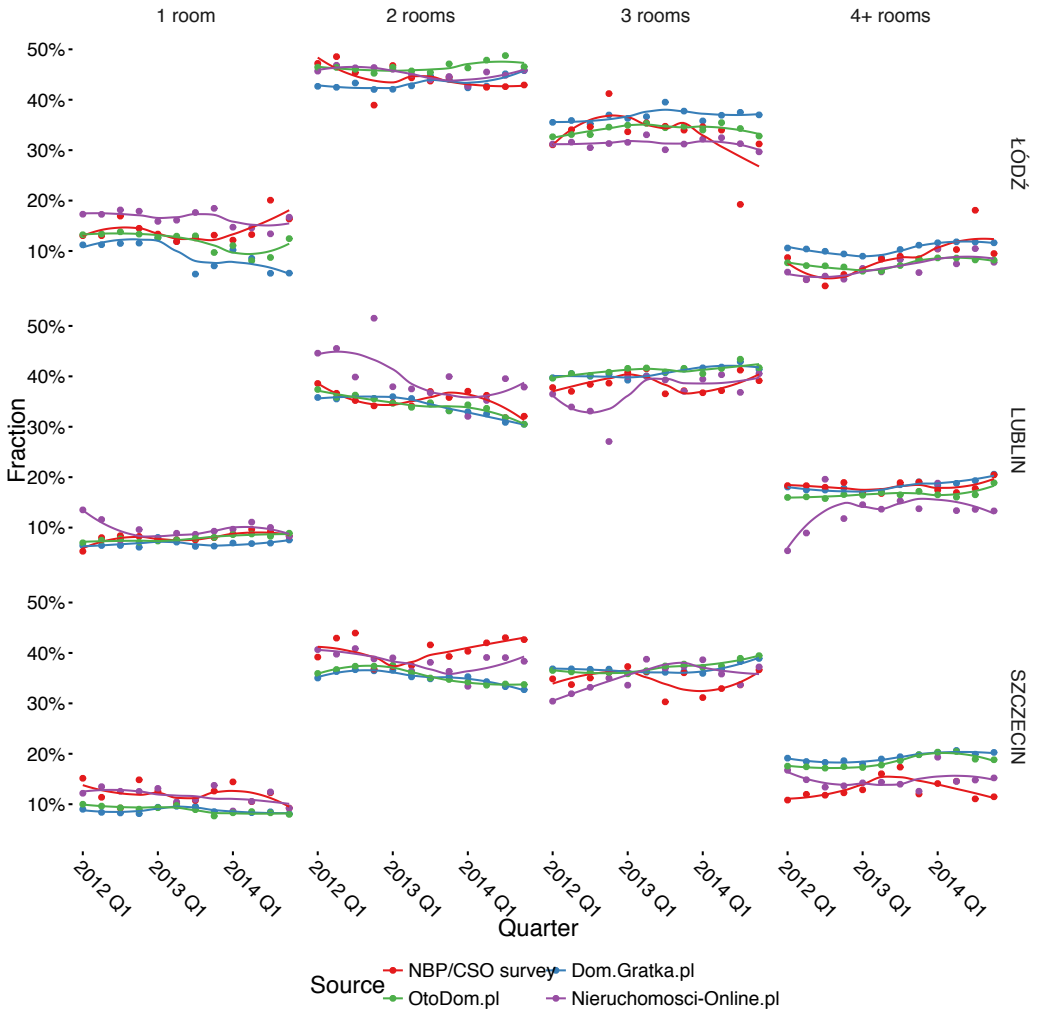


FIGURE A.2: Comparison of number of rooms distribution in NBP/CSO survey and IDS in Lublin, Szczecin and Łódź

A.3 Map of Poland

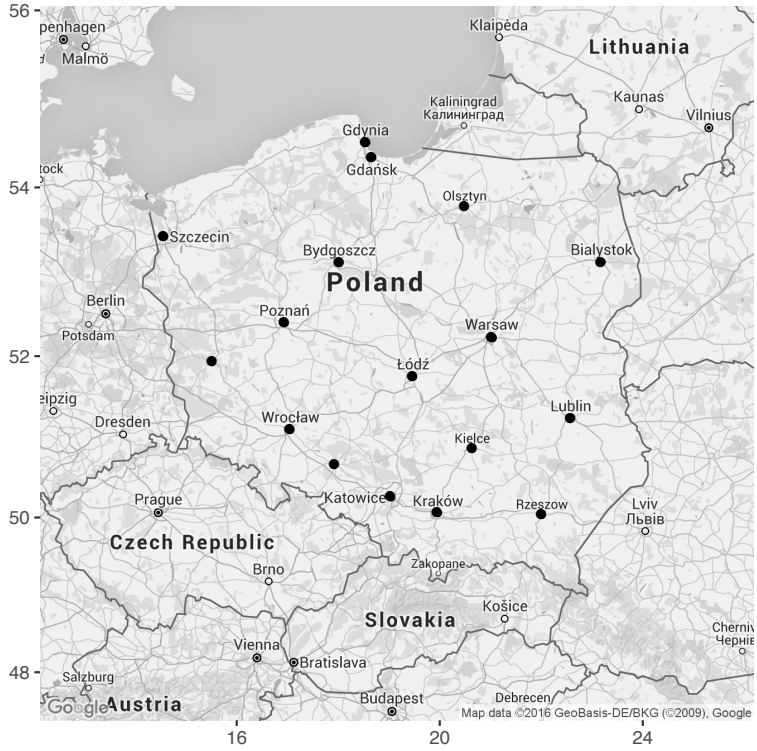


FIGURE A.3: Map of Poland with cities that are under research by NBP/CSO

A.4 R routines

Other R routines are available from the author at request.

A.4.1 ICT survey data

R Example A.4.1.

```
library(eurostat)

library(dplyr)

library(stringi)


get_eurostat(id = "isoc_ci_in_en2") %>% eurostat::label_eurostat(.) %>% filter(
    stri_detect(sizen_r2,
    regex = "Real estate activities"), unit == "Percentage of enterprises", year(
        time) %in%
    2010:2015) %>% filter(geo %in% c("Austria", "Bulgaria", "Cyprus", "Czech
        Republic",
    "Germany (until 1990 former territory of the FRG)", "Estonia", "Spain", "
        France",
    "Hungary", "Ireland", "Italy", "Lithuania", "Latvia", "Netherlands", "Poland
        ",
    "Portugal", "Romania", "Sweden", "Slovenia", "Slovakia", "United Kingdom", "
        European Union (28 countries)")) %>%
    mutate(geo = gsub(" \\(until 1990 former territory of the FRG\\)", "", geo),
        time = year(time)) %>% select(geo, time, values) %>% spread(time, values)
            %>%
    xtable(., digits = 0) %>% print(., include.rownames = F, comment = F,
        timestamp = F)
```

A.4.2 Estimated models

R Example A.4.2.

```
Multivariate Meta-Analysis Model (k = 432; method: REML)

Variance Components:

                estim      sqrt  nlvls  fixed  factor
sigma^2    18804.8469  137.1308     12     no    city

Test for Heterogeneity:
Q(df = 431) = 1974.6929, p-val < .0001

Model Results:

 estimate        se      zval      pval     ci.lb     ci.ub
 -99.6768   39.9934   -2.4923    0.0127 -178.0625  -21.2911         *

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R Example A.4.3.

```
Multivariate Meta-Analysis Model (k = 432; method: REML)

Variance Components:

                estim      sqrt  nlvls  fixed  factor    R
sigma^2.1  20305.3160  142.4967     12     no    city   no
sigma^2.2  37779.2640  194.3689      3     no  source  yes

Test for Heterogeneity:
Q(df = 431) = 1974.6929, p-val < .0001

Model Results:

 estimate        se      zval      pval     ci.lb     ci.ub
 -94.2299  147.9889   -0.6367    0.5243 -384.2828  195.8230

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R Example A.4.4.

```
Multivariate Meta-Analysis Model (k = 432; method: REML)

Variance Components:

                estim      sqrt  nlvls  fixed       factor    R
sigma^2.1  18174.4430  134.8126     12     no         city   no
sigma^2.2  36062.5169  189.9013      3     no       source  yes
sigma^2.3   3660.0794   60.4986     36     no  page_source   no

Test for Heterogeneity:
Q(df = 431) = 1974.6929, p-val < .0001

Model Results:

 estimate        se      zval      pval     ci.lb     ci.ub
 -92.7689  144.6238   -0.6414    0.5212 -376.2264  190.6886

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R Example A.4.5.

```
Multivariate Meta-Analysis Model (k = 432; method: REML)

Variance Components:

                estim      sqrt  nlvls  fixed       factor    R
sigma^2.1  13074.1752  114.3424     12     no         city   no
sigma^2.2  35975.0347  189.6709      3     no       source  yes
sigma^2.3   3804.0854   61.6773     36     no  page_source   no

outer factor: city    (nlvls = 12)
inner factor: quarter (nlvls = 12)

                estim      sqrt  fixed
tau^2      12107.8688  110.0358     no
rho            0.6919               no

Test for Heterogeneity:
Q(df = 431) = 1974.6929, p-val < .0001

Model Results:

 estimate        se      zval      pval     ci.lb     ci.ub
 -89.0714  144.1932   -0.6177    0.5368 -371.6849  193.5422

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A.4.3 Normality tests

R Example A.4.6.

```
        Shapiro-Wilk normality test

data:  rsd$z
W = 0.99268, p-value = 0.03295


        One-sample Kolmogorov-Smirnov test

data:  rsd$z
D = 0.039846, p-value = 0.4867
alternative hypothesis: two-sided
```

R Example A.4.7.

```
        Shapiro-Wilk normality test

data:  x
W = 0.95141, p-value = 0.6577


        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.15708, p-value = 0.8846
alternative hypothesis: two-sided
```

R Example A.4.8.

```
        Shapiro-Wilk normality test

data:  x
W = 0.97538, p-value = 0.5891


        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.07918, p-value = 0.9644
alternative hypothesis: two-sided
```

A.5 Definitions regarding real estate used in official statistics

- Residential building – A building designated towards residential purposes and fully occupied by residential premises or a building in which at least one half of the overall space constitutes residential space with the remaining part constituting other space (except for residential-animal buildings or residential-workshop buildings)[1].

- Premises – The room or set of rooms separated with the durable walls within the building dedicated to the permanent stay of people, which together with the auxiliary rooms serve the purpose of fulfilling their housing needs or which are used according to their dedication for the purposes other than residential purposes.

- Business premises – Premises used only for economic activity on rental terms and in the buildings owned by condominiums also premises constituting a separate ownership (retail outlets, service premises, production premises, studios, premises rented as office spaces, premises used for cultural and social activity etc.), which are a source of profit for the entity that rents them[2].

- Dwelling – A premise for a permanent residence of persons - built or remodelled for residential purposes structurally separated by fixed walls within a building, into which a separate access leads from a staircase, a passage, a common hall or directly from the street, courtyard or garden, comprising one or several rooms and auxiliary spaces.

  – Dwellings owned by a gmina – Dwellings (1) in buildings that are in whole owned by a gmina and individual dwellings owned by a gmina, but located in buildings constituting a shared real property, (2) owned by the State Treasury but administrated by a municipality, (3) handed over to a gmina, but remaining at the disposal of public use units[3].

  – Dwellings owned by housing co-operatives – Dwellings in buildings which are the property or joint property of housing co-operatives, excluding dwellings for which, on the grounds of the Act of December 15, 2000 on housing co-operatives, a separate ownership title was established, for the benefit of one or more natural persons[4].

  – Dwellings owned by natural persons – Dwellings, the ownership title to which belongs to a natural person (one or more, e.g. spouses), while

[1] See http://stat.gov.pl/en/metainformations/glossary/terms-used-in-official-statistics/1018,term.html.

[2] See http://stat.gov.pl/en/metainformations/glossary/terms-used-in-official-statistics/181,term.html.

[3] See http://stat.gov.pl/en/metainformations/glossary/terms-used-in-official-statistics/745,term.html.

[4] See http://stat.gov.pl/en/metainformations/glossary/terms-used-in-official-statistics/204,term.html.

this person: - can be the owner of the whole real estate in which the dwelling is situated; for example an individual single family house, - can have a share in a shared real estate, as the right related to the separate ownership title to a dwelling, e.g. located in a building included in a condominium[5].

– Dwellings owned by other entities – Dwellings constituting the property of: (1) private organizations involved in the building or buying of dwellings for profit: for sale or for rental; (2) trade unions, associations, foundations, political parties, professional and economic self-governments; (3) the Roman-Catholic Church and other churches and religious associations, including catholic universities and church institutes, (4) other entities, not covered by terms: "Dwellings owned by natural persons" "Dwellings owned by housing co-operatives" "Dwellings owned by companies" "Dwellings owned by a gmina" "Dwellings owned by the State Treasury" "Dwellings owned by public building societies" [6].

[5]See http://stat.gov.pl/en/metainformations/glossary/terms-used-in-official-statistics/203,term.html.

[6]See http://stat.gov.pl/en/metainformations/glossary/terms-used-in-official-statistics/1017,term.html.

TABLE A.5: Basic statistics on bias broken down by source

Source	Min Bias	Mean Bias	Median Bias	Max Bias	Std Bias	RAB [%]
Nieruchomosci-online.pl	-1062.60	-214.22	-202.87	300.53	229.32	4.04
Dom.Gratka.pl	-349.29	88.54	82.48	504.91	174.36	1.67
OtoDom.pl	-545.11	-58.62	-54.84	461.44	178.34	1.11

TABLE A.6: Basic statistics on bias broken down by city

City	Min Bias	Mean Bias	Median Bias	Max Bias	Std Bias	RAB [%]
GDAŃSK	-723.74	-286.15	-273.27	13.41	185.19	4.66
WARSZAWA	-1062.60	-292.17	-252.45	47.95	267.65	3.35
OPOLE	-290.62	135.04	148.65	495.34	167.55	3.30
WROCŁAW	-573.77	-170.02	-160.76	164.36	197.90	2.80
KRAKÓW	-197.31	173.15	175.13	461.44	155.29	2.60
BIAŁYSTOK	-378.24	-98.39	-66.61	92.06	124.15	2.16
POZNAŃ	-465.29	-98.45	-163.98	372.59	248.24	1.75
SZCZECIN	-402.87	-55.52	-59.84	209.13	150.35	1.28
KATOWICE	-464.26	-32.87	-64.17	333.92	215.48	0.82
OLSZTYN	-302.48	13.88	15.46	298.69	146.35	0.31
ŁÓDŹ	-329.01	-12.23	-6.17	504.91	213.88	0.31
LUBLIN	-233.17	-13.49	-29.68	427.27	132.57	0.28

TABLE A.7: Descriptive statistics of model-based estimated bias classified by Internet data sources and domains

Source	City	Minimum	Q1	Median	Mean	Q3	Maximum	RB [%]	ARB [%]
OtoDom.pl	BIAŁYSTOK	-80.05	-75.58	-71.73	-59.58	-41.58	-22.04	-1.32	1.32
OtoDom.pl	GDAŃSK	-318.20	-294.50	-252.20	-252.40	-219.00	-171.70	-4.11	4.11
OtoDom.pl	KATOWICE	-139.40	-85.60	-63.57	-66.30	-45.87	8.62	-1.64	1.68
OtoDom.pl	KRAKÓW	69.99	132.50	201.90	214.20	290.10	349.80	3.22	3.22
OtoDom.pl	ŁÓDŹ	-247.60	-229.80	-201.40	-127.60	8.68	83.23	-3.21	3.99
OtoDom.pl	LUBLIN	-33.54	-22.23	6.80	6.97	22.95	93.20	0.15	0.56
OtoDom.pl	OLSZTYN	-31.67	-16.06	1.30	4.10	20.06	47.66	0.10	0.46
OtoDom.pl	OPOLE	69.29	108.20	151.30	143.20	183.40	220.00	3.52	3.52
OtoDom.pl	POZNAŃ	-218.10	-191.30	-140.80	-93.88	23.87	57.49	-1.63	2.12
OtoDom.pl	SZCZECIN	-109.50	-104.40	-30.89	-47.89	-10.39	2.57	-1.10	1.12
OtoDom.pl	WARSZAWA	-274.80	-266.10	-263.20	-261.30	-254.00	-243.90	-2.99	2.99
OtoDom.pl	WROCŁAW	-239.30	-201.30	-180.90	-159.10	-108.60	-64.95	-2.63	2.63
Dom.Gratka.pl	BIAŁYSTOK	-21.54	-16.19	-4.69	8.80	34.34	54.65	0.18	0.56
Dom.Gratka.pl	GDAŃSK	-217.80	-203.00	-171.80	-165.90	-133.20	-93.88	-2.70	2.70
Dom.Gratka.pl	KATOWICE	135.40	153.60	203.60	192.90	222.20	242.50	4.83	4.83
Dom.Gratka.pl	KRAKÓW	97.19	153.70	203.30	215.80	269.60	335.80	3.24	3.24
Dom.Gratka.pl	ŁÓDŹ	118.90	142.30	152.50	211.20	315.70	365.40	5.36	5.36
Dom.Gratka.pl	LUBLIN	9.07	23.06	48.44	51.45	74.74	139.20	1.06	1.06
Dom.Gratka.pl	OLSZTYN	96.18	108.70	147.20	144.20	179.70	185.20	3.23	3.23
Dom.Gratka.pl	OPOLE	199.10	235.10	267.00	283.50	333.30	397.70	6.95	6.95
Dom.Gratka.pl	POZNAŃ	1.25	17.57	87.91	112.90	223.90	242.80	2.04	2.04
Dom.Gratka.pl	SZCZECIN	2.62	10.91	108.10	79.02	123.80	140.80	1.82	1.82
Dom.Gratka.pl	WARSZAWA	-79.72	-73.34	-65.34	-66.95	-62.77	-51.03	-0.77	0.77
Dom.Gratka.pl	WROCŁAW	-87.30	-74.88	-10.78	-3.85	69.71	83.05	-0.08	1.07
NieruchomosciOnline.pl	BIAŁYSTOK	-333.50	-283.10	-254.40	-243.50	-205.40	-143.30	-5.35	5.35
NieruchomosciOnline.pl	GDAŃSK	-519.30	-482.60	-415.20	-409.50	-345.60	-277.40	-6.68	6.68
NieruchomosciOnline.pl	KATOWICE	-367.90	-301.00	-263.20	-253.00	-201.40	-50.71	-6.28	6.28
NieruchomosciOnline.pl	KRAKÓW	-105.10	23.69	71.18	81.77	149.50	223.40	1.23	1.56
NieruchomosciOnline.pl	ŁÓDŹ	-273.00	-171.90	-120.40	-125.50	-61.47	-19.91	-3.15	3.15
NieruchomosciOnline.pl	LUBLIN	-149.40	-127.00	-112.30	-108.80	-96.63	-41.50	-2.21	2.21
NieruchomosciOnline.pl	OLSZTYN	-246.60	-184.30	-134.30	-130.00	-80.31	-16.27	-2.92	2.92
NieruchomosciOnline.pl	OPOLE	-182.90	-83.22	3.72	-24.56	21.44	57.06	-0.57	1.29
NieruchomosciOnline.pl	POZNAŃ	-371.40	-344.10	-323.70	-320.60	-297.10	-260.20	-5.68	5.68
NieruchomosciOnline.pl	SZCZECIN	-246.20	-237.60	-193.30	-193.10	-160.00	-126.50	-4.45	4.45
NieruchomosciOnline.pl	WARSZAWA	-677.00	-590.30	-448.00	-490.10	-386.20	-362.70	-5.59	5.59
NieruchomosciOnline.pl	WROCŁAW	-449.90	-390.10	-361.00	-339.20	-263.60	-223.20	-5.60	5.60

TABLE A.8: Spearman coefficient correlation matrix between variables presented in Table 5.2

	Offer	Trans	Ratio	No.Trans	RegonL	RanEff	Abs RanEff
Offer	1.00	0.98	-0.05	0.68	0.64	-0.48	0.58
Trans	0.98	1.00	0.03	0.73	0.69	-0.48	0.56
Ratio	-0.05	0.03	1.00	-0.17	-0.31	0.12	-0.17
No.Trans	0.68	0.73	-0.17	1.00	0.80	-0.36	0.58
RegonL	0.64	0.69	-0.31	0.80	1.00	-0.46	0.51
RanEff	-0.48	-0.48	0.12	-0.36	-0.46	1.00	0.10
Abs RanEff	0.58	0.56	-0.17	0.58	0.51	0.10	1.00