Assessing selectivity and representativeness of Internet data sources for the real estate market in Poland

Maciej Beręsewicz

Department of Statistics
Poznan University of Economics

The Fourth Baltic-Nordic Conference on Survey Statistics
24-28 August 2015, Helsinki, Finland

POZNAŃ UNIVERSITY
OF ECONOMICS

Outline

Outline

The goals and scientific questions of the presentation

The goals of the presentation

- Assessment of the representativeness of the Internet data sources (IDS) for real estate market (REM) statistics,
- Assessment of the selection bias of the Internet data sources for real estate market,
- taking into account:
  - secondry real estate market as a hard-to-reach population,
  - multiple overlapping data sources on the Internet,
  - current official research as a „gold standard".

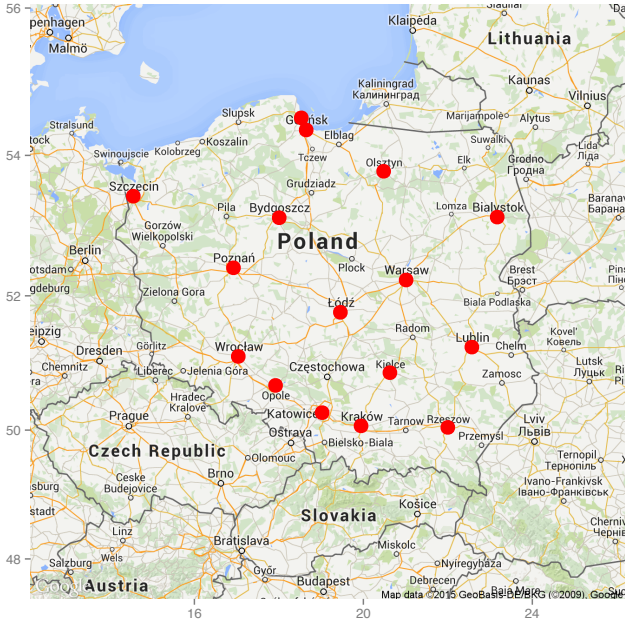
Questions that we would like to answer

- What kind of errors we can identify in the Internet data sources?
- What are the limitations of using Internet data sources for real estate market?
- Are Internet data sources representative of the secondary real estate market?
- Are there differences between Internet data sources for the secondary real estate market?

The research problem

Current research on real estate in Poland

- **Research**: The National Bank of Poland (NBP) and the Central Statistical Office (CSO) conduct a survey based on a sample of brokers in the biggest cities in Poland (red dots on the map).
- **Output**: Several characteristics of the primary and secondary market, including the offer and transaction price and the market structure are measured in this survey.
- **Drawbacks**: Data published quarterly, high response burden, results are published with a yearly delay.



Possible alternative data sources

- Internet is becoming the main data source for information about the real estate market.
- In order to sell real estate, brokers need to present information to a wider audience via e.g. the Internet.
- There are examples of the use of this source for statistics – Statistics Netherlands is using *Funda.nl* for housing market statistics.

Web portals dedicated to the real estate market in Poland

There are several big online advertising services for the real estate market in Poland. Internet data sources used in this study include:

Internet data sources

- **OtoDom** – owned by the Allegro Group, daily number of offers of flats for sale is over 378 000 (for Poznań over 9 500).
- **Gratka** – owned by Polska Presse, daily number of offers of flats for sale is over 366 000 (for Poznań over 13 000).
- **Morizon/Domy.pl/Oferty.net** – daily number of offers of flats for sale is over 355 000 (for Poznań over 9 200).
- **Szybko.pl** – daily number of offers of flats for sale is over 423 000 (for Poznań over 9 200).
- **Nieruchomosci-online.pl** – daily number of offers of flats for sale is over 170 000 (for Poznań over 5 000). This portal provides access to archived advertisements.

Remarks: these are numbers presented by the owners of these services and do not take into account multiple occurrences or misclassifications.

Secondary real estate market as a hard-to-reach population

Hard-to-reach populations – characteristics

Table: Comparison between a hard-to-reach population and the secondary real estate market

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

Source: based on Marpsat and Razafindratsima (2010).

Statistical challanges and opportunities in using IDS for REM statistics

Challenges

- Various web-services devoted to the real estate market in Poland.
- Web portals differ in the number of offers listed, the popularity or specialization and, more importantly, in quality.
- Non-official research indicates that brokers use from 4 to 9 web portals to place their offers. Therefore, there is a problem of multiple, overlapping and correlated data sources.
- Data cleaning process (e.g. semi-structured data, natural language processing).
- Distribution of sample characteristics on web portals may be different from the population.

Opportunities

- Decrease in respondent burden via automatic data collection (collect what is already available).
- Provide statistics at a more detailed level on a monthly basis.
- Possibility of integrating these data sources with register data on transactions (via probabilistic record linkage).

Sources of bias in Internet data sources for real estate market statistics

Generic errors

- Coverage error – not all brokers use the Internet; not all offers are placed on the Internet ; specialization of web-services (e.g. apartments, houses, improved lands, studios, lofts). In addition, coverage error may me highly correlated with Internet access ratio.
- Measurement error – the same, similar or different definitions; data can be rounded or erroneous, false (on purpose).
- Missing data – missing data in non required fields.

Specific errors

- Selection error, due to:
  - popularity and effectiveness of certain web-portals,
  - owners of the web-portal,
  - fees for the owner, advertising costs.
- Unit error – problem with the identification of units (multiple occurrences within and between data sources, lack of identifiers, limited variables to identify units, misclassifications),
- Errors connected with the nature of data source:
  - Multiple, overlapping data sources – high redundancy,
  - Semi-structured – information represented in a natural language.

Outline

Representativeness – definitions

Definitions

Kruskal and Mosteller (1979a, 1979b, 1979c) present an overview of the meanings of "representative sampling"

- General acclaim of data
- Absence of selective forces
- 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
- As permitting good estimation
- Good enough for a particular purpose

Bethlehem's (2009) definition of representativeness

A survey data set is defined to be *representative with respect to variables* $\boldsymbol{X}$ if the distribution of $\boldsymbol{X}$ in the data set is equal to the distribution of this variable in the population ($F(\hat{\boldsymbol{X}}) = F(\boldsymbol{X})$).
When, a weighting procedure is applied, then sample is *representative with respect to the auxilary variables*.

Representativeness – practical aspects

Practical issues

- Questions about *representativeness* are also valid in the case of sample surveys (e.g. in the presence of nonignorable nonresponse or estimation for unplaned (small) domains).
- It is an important issue in the case of self-selection web surveys.
- Register data are also evaluated in order to verify whether administrative data are representative of the target population (e.g. different definitions of units).
- Furthermore, this aspect is still valid in the world of new (massive) data sources (e.g. big data, the Internet).

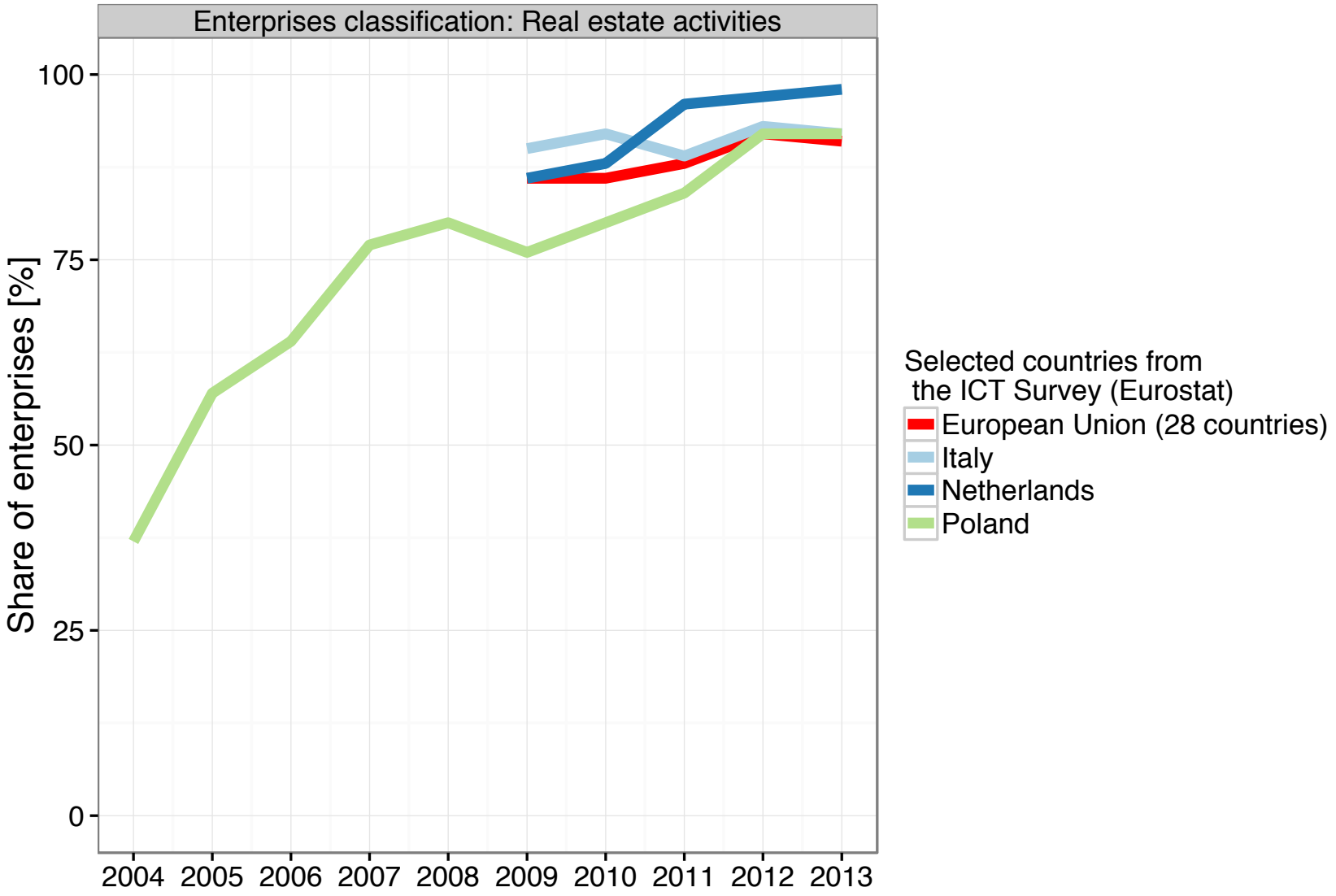
How to measure representativeness in the case of IDS for REM?

We need to ask the following questions concerning representativeness of Internet data sources for real estate market statistics:

- What is the target population? For example, are we interested in the number of brokers that use IDS or real estates offered for sale? – this may lead to different answers
- Do we have population quantities for comparison? Do we have an offical data source (survey or register based) that provides unbiased estimates as a reference?
- With respect to the coverage of flats – what share of flats are published on-line; drawback – there is a lack of official research on this matter

Therefore, how we can measure representativeness of Internet data sources?

Access to the Internet in enterprises



ICT usage in enterprises in Poland based on the ICT survey

ICT usage in enterprises in Poland based on the ICT survey

Specification	2010	2011	2012	2013	2014
Enterprises having a website or homepage in % of total enterprises in a group					
Total	65.5	64.7	67.6	66.0	65.3
Real estate activities	67.0	63.3	70.9	74.9	73.9
Product catalogues or price lists in % of total enterprises in a group					
Total	48.8	46.9	51.4	51.5	60.4
Real estate activities	26.8	25.0	32.8	35.7	51.4

Source: based on the ICT survey in Poland.

ICT usage in enterprises in Poland based on the ICT survey – remarks

- Target population: the ICT survey was addressed to companies with 10 and more employees, while brokers in Poland are othen self-employed
- Questions stated: brokers do not need to have their own webpage to offer properties for sale, they often use advertising web services
- Domain (Real estate activities): contains several different types of enterprises, which can only be partially connected to the sale process.

Therefore fractions in these two tables may be underestimated.

Representativeness – measures of representativeness

Population R-Indicators

Schouten et al. (2009) proposed two types of indicators for the representativeness of a survey response – population and response-based R-indicators. Population based indicators assume that we know the population size and/or fraction of mean propensities $\bar{\rho}$ and is given by (when response-level ρ_i are unknown)

$$R(\rho) \geq 1 - 2\sqrt{\bar{\rho}(1-\bar{\rho})}. \tag{1}$$

Maximum bias is equal to $|B_{max}| = S(Y)\sqrt{1/\bar{\rho} - 1}$. When domain propensities are known ρ_d we apply

$$R(\rho) \approx 1 - 2\sqrt{\sum_{d=1}^{D} f_d(\bar{\rho}_d - \bar{\rho})^2}. \tag{2}$$

When, ρ_i and $1/\pi_i$ are known for each i unit, response-based R indicator is given by:

$$R(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}(\rho_i - \bar{\rho})^2} \tag{3}$$

Proposed measures of representativeness of IDS

Generalized population R-Indicators

I propose a generalized population R-indicators that do not assume known total or domain population size, instead it can be estimated using capture-recapture methods.

Comparison of marginal distributions

I propose the following approach to comparing marginal distributions:

1. Visual comparison of marginal distributions along with standard errors; if a time series is available, we compare trends over time.
2. Test differences between marginal distributions of Internet data sources and official reference data under a linear mixed model that takes into account known or smoothed sampling variances.

Generalized population R-indicators

Generalized population R-indicator

A generalized population R-indicator (GPR indicator) is given by:

$$R(\tilde{\rho}_k) = 1 - 2S(\tilde{\rho}_k) \tag{4}$$

where

$$S(\tilde{\rho}_k) \leq \sqrt{\tilde{\bar{\rho}}_k(1 - \tilde{\bar{\rho}}_k)} \tag{5}$$

where $\tilde{\bar{\rho}}_k$ is given for each k data source by:

$$\tilde{\bar{\rho}}_k = \frac{N_k}{\hat{N}} \tag{6}$$

where N_k is the size of k Internet data source and $\hat{N}$ is a population size estimate based on the capture-recapture procedure.

Generalized population R-indicator

Capture-recapture for R-indicator

The simpliest estimator of $\hat{N}$ based on Petersen estimator (Lavallée and Rivest 2012) / Dual System Estimation is given by:

$$\hat{N} = \frac{N_k \times N_{k'}}{N_{kk'}} \tag{7}$$

where $N_{k'} = max\{N_1, N_2, ..., N_k\}$, $k \neq k'$ and $N_{kk'}$ is number of units that occur in both data sources. If we would like to calculate GPR indicator for domain we estimate $\hat{N}_d$ by

$$\hat{N}_d = \frac{N_{kd} \times N_{k'd}}{N_{kk'd}}, \tag{8}$$

where $N_{k'd} = max\{N_{1d}, N_{2d}, ..., N_{kd}\}$ and then $\tilde{\bar{\rho}}_d = N_{kd}/\hat{N}_d$. In addition, equation (4) can be expressed as a conditional generalized R-indicator given by:

$$R(\tilde{\rho}_k|kk') = 1 - 2S(\tilde{\rho}_k) \tag{9}$$

In order to estimate confidence intervals of (9) we use parametric boostrap under multinomial distribution to estimate $\hat{N}$.

Representativeness – Comparison of marginal distributions

Comparison of marginal distributions

For the sake of comparison of the marginal distribution we propose applying a linear mixed model given by:

$$\breve{\theta} - \theta = \beta + \boldsymbol{Zv} + \epsilon \tag{10}$$

or when θ is estimated based on the sample

$$\breve{\theta} - \hat{\theta} = \beta + \boldsymbol{Zv} + \epsilon \tag{11}$$

where θ is a vector of population marginal distribution of $\boldsymbol{X}$ variables, $\breve{\theta}$ is a vector of marginal distribution estimated based on Internet data sources, $\hat{\theta}$ is a sample-based estimate of the marginal distribution. $\boldsymbol{Zv}$ is a matrix of random effects for $\boldsymbol{X}$ and ϵ denotes known sampling variance from the Internet data source ξ and, in the case of sample-based population totals ψ.

Outline

Definition of selectivity in IDS context

For the sake of the study we propose the following definition of `selectivity`:

Selectivity

Selectivity is observed when $\breve{\theta} \neq \theta$ under the assumption that $F(\breve{\boldsymbol{X}}) = F(\boldsymbol{X})$ where $\breve{\theta}$ denotes a target statistic of target variable estimated based on new data source and known quantity for the target population θ and $\boldsymbol{X}$ denotes auxiliary variables. $F(\breve{\boldsymbol{X}}), F(\boldsymbol{X})$ denotes the distribution of $\boldsymbol{X}$ in new data source and population.

Selectivity

When $\boldsymbol{Y}$ is unknown and needs to be estimated, we propose the following definition:
Selectivity is observed when $\breve{\theta} \neq \hat{\theta}$ under the assumption that $F(\breve{\boldsymbol{X}}) = F(\hat{\boldsymbol{X}})$ where where $\breve{\theta}$ denotes a target statistic of target variable estimated based on new data source, θ is known from survey and $\boldsymbol{X}$ denotes auxiliary variables. $F(\breve{\boldsymbol{X}}), F(\boldsymbol{X})$ denotes distribution of $\boldsymbol{X}$ in new data source and sample.

Weak and strong selectivity at domain level in IDS

In addition, for the sake of research we propose the following definition of weak and strong selectivity in the case of multiple data sources. We start with a linear mixed model given by:

$$\breve{\theta} - \theta = \beta + \boldsymbol{Zv} + \boldsymbol{\epsilon} \tag{12}$$

where $\breve{\theta}$ is a statistic (mean, proportion) of target variable $\boldsymbol{y}$ estimated based on k Internet data sources, θ is a true value of statistic of $\boldsymbol{y}$ target population (under assumption that $F(\breve{\boldsymbol{X}}) = F(\boldsymbol{X})$), $\boldsymbol{Zv}$ is a matrix of random effects (domain, data source) and $\boldsymbol{\epsilon}$ is random error with known sampling variance from k Internet data sources $\boldsymbol{\epsilon} \sim N(0, \xi)$.
Under this model we define weak and strong selectivity:

- We assume a basic model with only one random component – domain effect
- weak selectivity – occurs when a random effect for data source significantly improves the model in comparison to the model with only domain effect,
- strong selectivity – occurs when a random effect for interaction between domain and data sources significantly improves the model in comparison to the model defined for weak selectivity.

Outline

Web scraping – the idea

Web scraping definition

Web scraping (web harvesting, web data extraction) is a computer software technique of extracting information from websites. Usually such software programs simulate human exploration in the World Wide Web by either implementing low-level Hypertext Transfer Protocol (HTTP), or embedding a fully-fledged web browser (such as IE, Mozilla, Chrome or Safari). (Wikipedia, 2015)

Web scraping as a mode of data collection

- Decreases respondent burden; limits the number of questions
- Collects what is already available (on the Internet)
- Can be part of a mixed mode data collection
- Can be used for the creation of sampling frames

Already used by NSIs (Barcaroli 2015; Buelens, Boonstra and Daas 2012; Daas et al. 2011; Griffioen, de Haan, Willenborg 2014; Hoekstra, ten Bosch, Harteveld 2012) and Economists (The Billion Price Project, Cavallo 2013) for statistical purposes.

Web scraping – drawbacks

- Blocking of scrapers
- Owners of services can limit available data
- Demanding services (e.g. flight reservation services)

Co-operation between enterprises and University

- For the purpose of the study co-operation between University and data owners was established
- Historical, aggregated data was acquired (free of charge) with predefined classifications
- Current data are downloaded directly via API or web-scraping.

Skills needed for data capture

For the project there ways of data capture was applied which was a result of different technologies that these companies use:

- Web-scraping technique was used in the co-operation enabling direct data capture from the webpage - Python + R
- Access to internal API whereby data are downloaded directly from the server by sending queries - JSON, PHP + R

Outline

Assessment of representativeness of two data sources

Data sources

Data were obtained via a web scraping technique from two Internet data sources `OtoDom` and `Gratka`. Only flats that were offered for sale between 01.08.2013 and 14.08.2013 were analyzed. After the cleaning procedure, for `Gratka` (N_1) we obtained 2532 flats (initially 2 780), for `OtoDom` (N_2) there were 2187 and overlap (N_{12}) between these two data sources was equal to 1974. Therefore, the overlap ratio for `OtoDom` was equal to 90.26% and for `Gratka` 77.96%.

Results

Results of estimation:

- Estimated population size $\hat{N}$ was equal to 2805,
- Estimated propensity score for `Gratka` was 0.9 and for `OtoDom` was 0.78,
- Maximum bias for `Gratka` was $S(Y) \times 0.33$ and for `OtoDom` was $S(Y) \times 0.53$,
- Estimated Generalized R-indicator for `Gratka` was equal to 0.41 and for `OtoDom` was 0.17 which indicates that service `Gratka` is more representative than `OtoDom`.
- Estimated 95% confidence intervals (based on bootstrap procedure for $\hat{N}$) for both measures are given below
  - `Gratka` – $(0.38, 0.41)$
  - `OtoDom` – $(0.16, 0.17)$

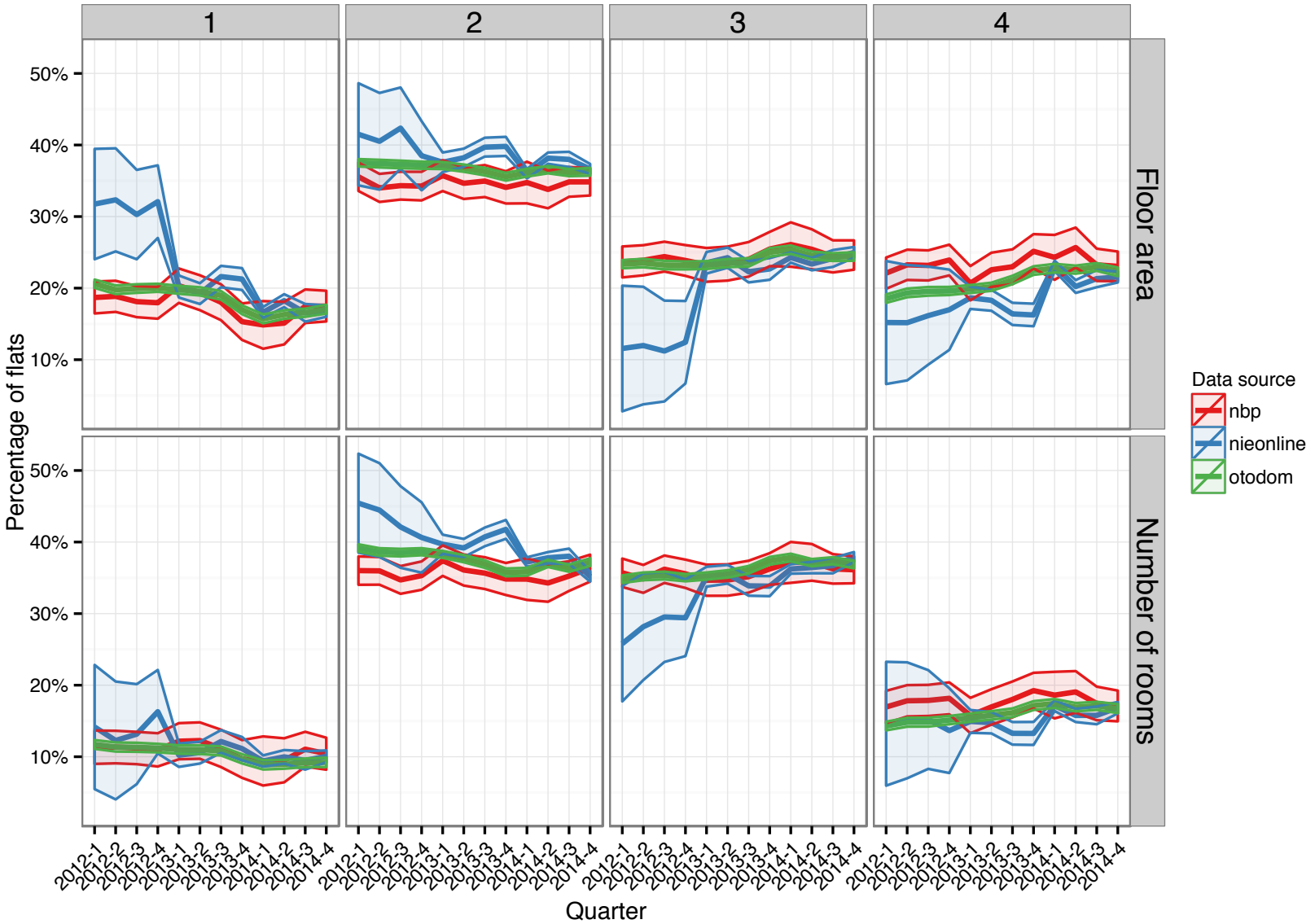
Comparison of distributions – selected results (Warsaw)

Marginal distributions

Data from the IDS was harmonized for the comparison purposes. In result, the following variables were taken into account:

- Rooms – 1 room (1), 2 rooms (2), 3 rooms (3), 4 and more rooms (4)
- Floor area – to 40 m2 (1), 40-60 m2 (2), 60-80 m2 (3), 80 and more m2 (4)

Comparison of distributions – selected results (Warsaw)



Comparison of distributions – selected results (Warsaw)

Model

The following model was estimated for these two data sources:

$$\breve{\theta} - \hat{\theta} = \beta + \boldsymbol{u} + \boldsymbol{e} \tag{13}$$

where ***u*** denotes random effect representing each level of floor area and number of rooms, ***e*** denotes random error with known variace from IDS data source (***xi***) and sample survey (***psi***).

Results

- Otodom – $\beta\,(SE) = 0.0000(0.0058)$, $\sigma^2\,(SE) = 0.0003(0.0161)$
- Nieruchomosci-pl – $\beta\,(SE) = 0.0005(0.0095)$, $\sigma^2\,(SE) = 0.0007(0.0265)$

Where σ^2 is the variance of the random effect of the number of rooms and floor area. Differences between marginal distributions are not significant, therefore we can say that IDS are representative with respect to the $\mathbf{X}$ which are the number of rooms (1,2,3,4+) and floor area (<=40,40-60,60-80,80+).

Outline

Discussion

- A proposal of generalized population R-indicator which take into account estimated population size was presented; in addition method for comparing marginal distributions taking into account sampling variances was proposed.
- We observe big difference in terms of representative "response" in Internet data sources.
- The results indicates that two presented Internet data sources are representative with respect to floor area and number of rooms.
- Research indicates that Internet data sources could be used for describing secondary real estate market in Poland.
- There is a need for a official research in order to further representativeness and selectivity assessment.

Acknowledgements

The paper and the research has been financed by National Science Centre Poland, Preludium 7 grant no. 2014/13/N/HS4/02999.

Contact

Department of Statistics
Poznan University of Economics
e-mail: maciej.beresewicz@ue.poznan.pl

Thank you for your attention!

Outline

Literature I

- Barcaroli, G. (2015). Internet as Data Source in the Istat Survey on ICT in Enterprises. Austrian Journal of Statistics, 44(2), 31-43.
- Bethlehem, J. (2009). Applied survey methods: A statistical perspective (Vol. 558). John Wiley & Sons.
- Bethlehem, J. (2010). Selection Bias in Web Surveys. International Statistical Review, 78(2), 161–188. doi:10.1111/j.1751-5823.2010.00112.x
- Buelens, B., Boonstra, H. J., & Daas, P. J. H. (2012). Shifting paradigms in official statistics (No. 18). Statistics Netherlands. The Hague/Herleen: Statistics Netherlands.
- Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. Journal of Monetary Economics, 60(2), 152–165. doi:10.1016/j.jmoneco.2012.10.002
- Cavallo, A. (2013). Scraped Data and Sticky Prices, (May).
- Daas, P. J. H., Roos, M., de Blois, C., Hoekstra, R., ten Bosch, O., & Ma, Y. (2011). New data sources for statistics: experiences at Statistics Netherlands (No. 9). The Hague/Herleen: Statistics Netherlands.
- Griffioen, R., de Haan, J., & Willenborg, L. (2014). Collecting clothing data from the Internet. Heerlen.

Literature II

- Hoekstra, R., ten Bosch, O., & Harteveld, F. (2012). Automated data collection from web sources for official statistics: First experiences. Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 28(3), 99-111.
- Kruskal, W., & Mosteller, F. (1979a). Representative sampling I: Non-scientific literature. International Statistical Review, 47, 13-24.
- Kruskal, W., & Mosteller, F. (1979b). Representative sampling II: Scientific literature, excluding statistics. International Statistical Review, 47, 111-123.
- Kruskal, W., & Mosteller, F. (1979c). Representative sampling III: The current statistical literature. International Statistical Review, 47, 245-265.
- Lavallée, P., & Rivest, L.-P. (2012). Capture – Recapture Sampling and Indirect Sampling. Journal of Official Statistics, 28(1), 1–27.
- Marpsat, M., & Razafindratsima, N. (2010). Survey methods for hard-to-reach populations : introduction to the special issue. Methodological Innovations Online, 5(2), 3–16. doi:10.4256/mio.2010.0014
- Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. Survey Methodology, 35(1), 101–113.