On representativity and quality of Big Data for Real Estate Market Analysis

Maciej Beręsewicz [1]
maciej.beresewicz@ue.poznan.pl

[1]Department of Statistics, Poznan University of Economics

POZNAŃ UNIVERSITY OF ECONOMICS

Outline

Motivation

Key motivation points

- Shifting paradigms in Official Statistics (from surveys to registers to new data sources),
- Looking beyond administrative sources,
- Internet data sources & Big Data,
- Importance of real estate market in economy.

Motivation

Key motivation points

- Shifting paradigms in Official Statistics (from surveys to registers to new data sources),
- Looking beyond administrative sources,
- Internet data sources & Big Data,
- Importance of real estate market in economy.

Current work on internet data sources - selected topics

- Online auctions - Shmueli, Jank and Bapna (2005),
- E-commerce portals - Bapna et al (2006),
- Measure inflation - Cavallo (2012), MIT: the Billion Price Project,
- Measuring prices - Hoekstra, ten Bosch, Harteveld (2010), Daas et al (2011),
- Sentiment analysis - Daas et al (2013), Daas et al (2014a).
- Forecasting elections with non-representative polls - Wang W., Rothschild D., Goel S. and Galeman A. (under review in Journal of Forecasting).

First concept for measuring representativity of Big Data - Buelens et al (2014), Daas i Puts (2014b).

This presentation

The goals of the presentation:

This presentation

The goals of the presentation:

- to present real estate market analysis in Poland,

This presentation

The goals of the presentation:

- to present real estate market analysis in Poland,
- and problems in context of estimation,

This presentation

The goals of the presentation:

- to present real estate market analysis in Poland,
- and problems in context of estimation,
- to propose a definition of Big Data/internet data sources in context of existing definitions,

This presentation

The goals of the presentation:

- to present real estate market analysis in Poland,
- and problems in context of estimation,
- to propose a definition of Big Data/internet data sources in context of existing definitions,
- to present quality and representativity problems when analysing real estate using internet data sources,

This presentation

The goals of the presentation:

- to present real estate market analysis in Poland,
- and problems in context of estimation,
- to propose a definition of Big Data/internet data sources in context of existing definitions,
- to present quality and representativity problems when analysing real estate using internet data sources,
- to present first results of assessing representativity of real estate market in Poland.

Real estate market in Poland

Real estate market in Poland

Organisation

With the beginning of 2014 in Poland deregulation of several types of professions were made. All legal regulations for real estate broker, residential manager and appraiser were abolished.

Real estate market in Poland

Organisation

With the beginning of 2014 in Poland deregulation of several types of professions were made. All legal regulations for real estate broker, residential manager and appraiser were abolished.

Consequences

- lack of a full list (register) of brokers on real estate market,
- lack of a full list (register) of offered real estates (flats, houses...)
- estimation problems ...

Real estate market in Poland

Organisation

With the beginning of 2014 in Poland deregulation of several types of professions were made. All legal regulations for real estate broker, residential manager and appraiser were abolished.

Consequences

- lack of a full list (register) of brokers on real estate market,
- lack of a full list (register) of offered real estates (flats, houses...)
- estimation problems ...

Existing data sources

- Price and Value of Real Estates Register (pol. *Rejestr Cen i Wartości Nieruchomości*) - held by local governments in poviats (LAU1). Contains: transactions on primary and secondary market.
- Two research programs conducted by Central Statistical Office in Poland and National Bank of Poland:
  - residential resource management,
  - real estate trading and survey of prices of residential,
  - commercial property.

Real estate market as a hard-to-reach populaton

What is hard-to-reach population?

- Lack of list of all units,
- Problems when reaching and approaching units of this population,
- Problems when using classical approach to estimation (no design),
- Usage of non-probability samples (eg. respondent driven sampling, indirect sampling).

Real estate market as a hard-to-reach populaton

What is hard-to-reach population?

- Lack of list of all units,
- Problems when reaching and approaching units of this population,
- Problems when using classical approach to estimation (no design),
- Usage of non-probability samples (eg. respondent driven sampling, indirect sampling).

REM as a hard-to-reach population

- Lack of registers of brokers, real estates,
- On one hand brokers are not willing to give information concerning their work (offers),
- On the other hand in order to sell real estate they need to present detailed information - mainly on the Internet.

Real estate market in Poland

Internet data sources

When speaking about Internet data sources we think about:

- brokers' portals (np. HomeBroker.pl, Metrohouse.pl),
- brokers associations' portals (np. REAL NET - nieruchomista.pl, PFRN - fagora.pl),
- portals offering support in sale (np. otodom.pl, domiporta.pl),
- portals that aggregate other portals (np. dom.money.pl).

. . . of different quality.
Existing research indicates that from 80% to 90% of people who bought real estates used Internet in order to get information (non-official statistics, otodom.pl and TNS Poland (2014)).

Big Data and Internet data sources

Definition problems

Big Data and Internet data sources

Definition problems

- According to Economic Commission for Europe of the United Nations (UNECE) Internet data sources/Big Data can be a part of *administrative sources* which are defined as *data collected by sources external to statistical offices*.

Big Data and Internet data sources

Definition problems

- According to Economic Commission for Europe of the United Nations (UNECE) Internet data sources/Big Data can be a part of *administrative sources* which are defined as *data collected by sources external to statistical offices*.
- Eurostat defines *administrative sources* more precise as *A data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations* which does not include internet data sources . . . but some types of Big Data does.

Big Data and Internet data sources

Definition problems

- According to Economic Commission for Europe of the United Nations (UNECE) Internet data sources/Big Data can be a part of *administrative sources* which are defined as *data collected by sources external to statistical offices*.
- Eurostat defines *administrative sources* more precise as *A data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations* which does not include internet data sources . . . but some types of Big Data does.

Internet data sources and Big Data are not precisely defined in statistics. Terms are widely discussed in the context of Big Data characteristics pointed out by Bayler (2011) and Lanley (2012) in 3 (or more) Vs definition - volume, velocity and variety.

Big Data and Internet data sources

Internet data sources Sources containing data collected and maintained by units external to statistical offices and administrative regulations which are available (mainly) on the Internet (through web-based databases).

Big Data and Internet data sources

Internet data sources Sources containing data collected and maintained by units external to statistical offices and administrative regulations which are available (mainly) on the Internet (through web-based databases).

Which is also true for some types of Big Data.

Estimation and quality issues

Estimation and quality issues

- Conceptualisation

Estimation and quality issues

- Conceptualisation
- Representativity

Estimation and quality issues

- Conceptualisation
- Representativity
- Selectivity

Estimation and quality issues

- Conceptualisation
- Representativity
- Selectivity
- Measurement and other nonsampling errors (eg. unit error theory Zhang (2011))

Estimation and quality issues

- Conceptualisation
- Representativity
- Selectivity
- Measurement and other nonsampling errors (eg. unit error theory Zhang (2011))
- Measure uncertainty and precision

Estimation and quality issues

- Conceptualisation
- Representativity
- Selectivity
- Measurement and other nonsampling errors (eg. unit error theory Zhang (2011))
- Measure uncertainty and precision
- Sampling (stream Big Data, why download all data?)

Estimation and quality issues

- Conceptualisation
- Representativity
- Selectivity
- Measurement and other nonsampling errors (eg. unit error theory Zhang (2011))
- Measure uncertainty and precision
- Sampling (stream Big Data, why download all data?)
- Estimation

Estimation and quality issues

- Conceptualisation
- Representativity
- Selectivity
- Measurement and other nonsampling errors (eg. unit error theory Zhang (2011))
- Measure uncertainty and precision
- Sampling (stream Big Data, why download all data?)
- Estimation
- ...

Estimation and quality issues

- Conceptualisation
- Representativity
- Selectivity
- Measurement and other nonsampling errors (eg. unit error theory Zhang (2011))
- Measure uncertainty and precision
- Sampling (stream Big Data, why download all data?)
- Estimation
- ...
- Place in statistical information system.

Concept of representativity I

Kruskal and Mosteller (1979a, 1979b, 1979c) definitions

The key elements of statistical data sources are representativity and quality. In literature many definitions can be found, but none is given explicitly. Kruskal and Mosteller (1979a, 1979b, 1979c) made a complex literature review and presented nine definitions of representativity. In the papers authors refer to the following aspects:

- general acclaim about data,
- lack of selective forces,
- miniature of population,
- typical/ideal cases,
- reflects variability of population,
- definition without explaining what it means,
- refers to specific sampling methods (equality of probability of inclusion),
- provides good estimation,
- fit to specific purposes.

Bethlehem definition

Bethlehem (2009) defined representativity with respect to sample when *relative distributions are the same in sample and population*. It means that sample is representative when **characteristics of sample and populations are the same**. This statement can be understood that representative sample is the same as miniature of population by Kruskal and Mosteller (1979a, 1979b, 1979c).

Concept of representativity

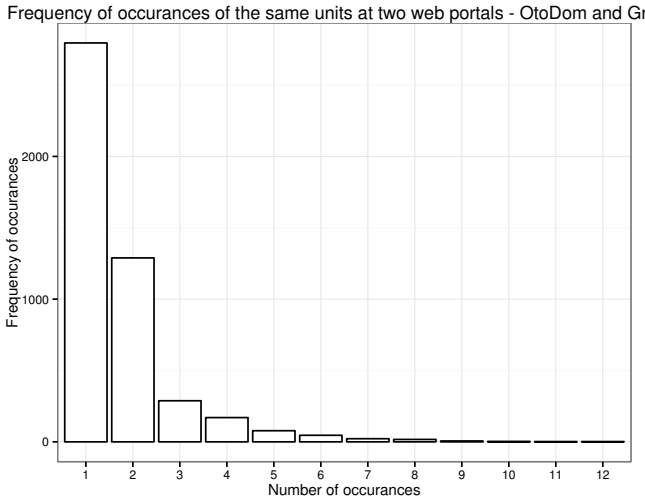
In context of which representativity definition should we assess representativity? Or we need another definition?

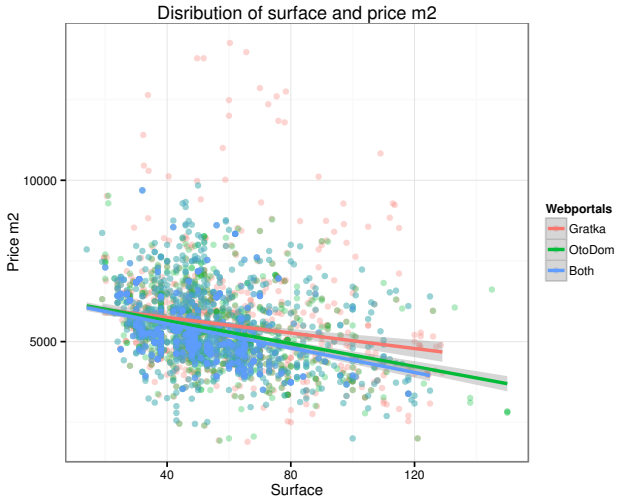
- miniature of population ...
- or more precisely definition proposed by Bethlehem (2009) - characteristics of sample is the same as in the population

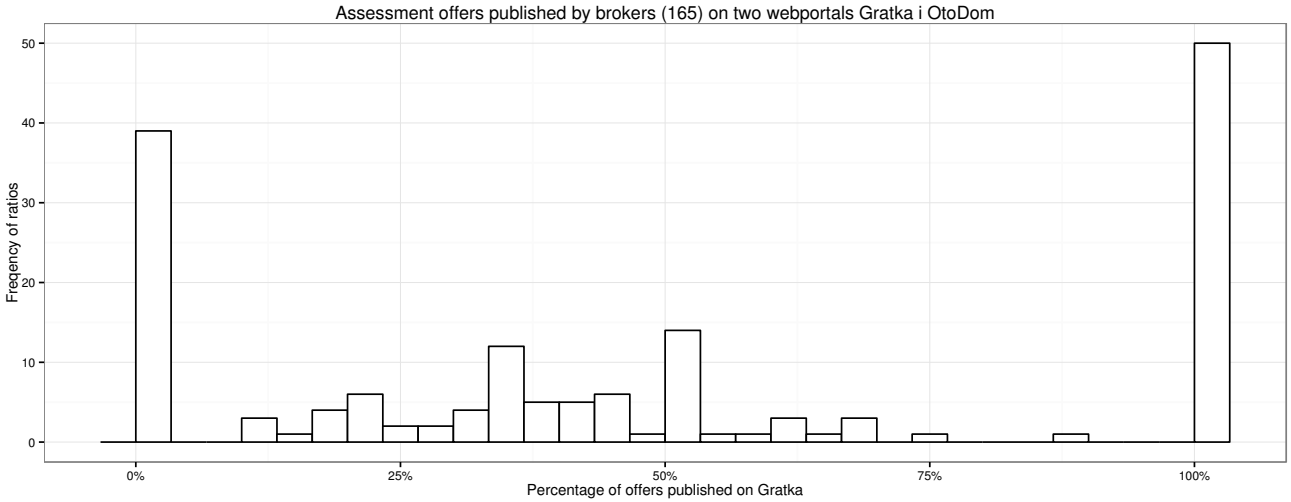
Research

- Web scraping algorithm was developed (in R with RCurl, XML and httr packages),
- Three web portals were choosen - otodom.pl, gratka.pl and nieruchomosci-online.pl (for the last was the longest time period),
- Population - flats offered on secondary market in Poznań, Poland.
- Reference research by National Bank of Poland and Central Statistical Office was taken.

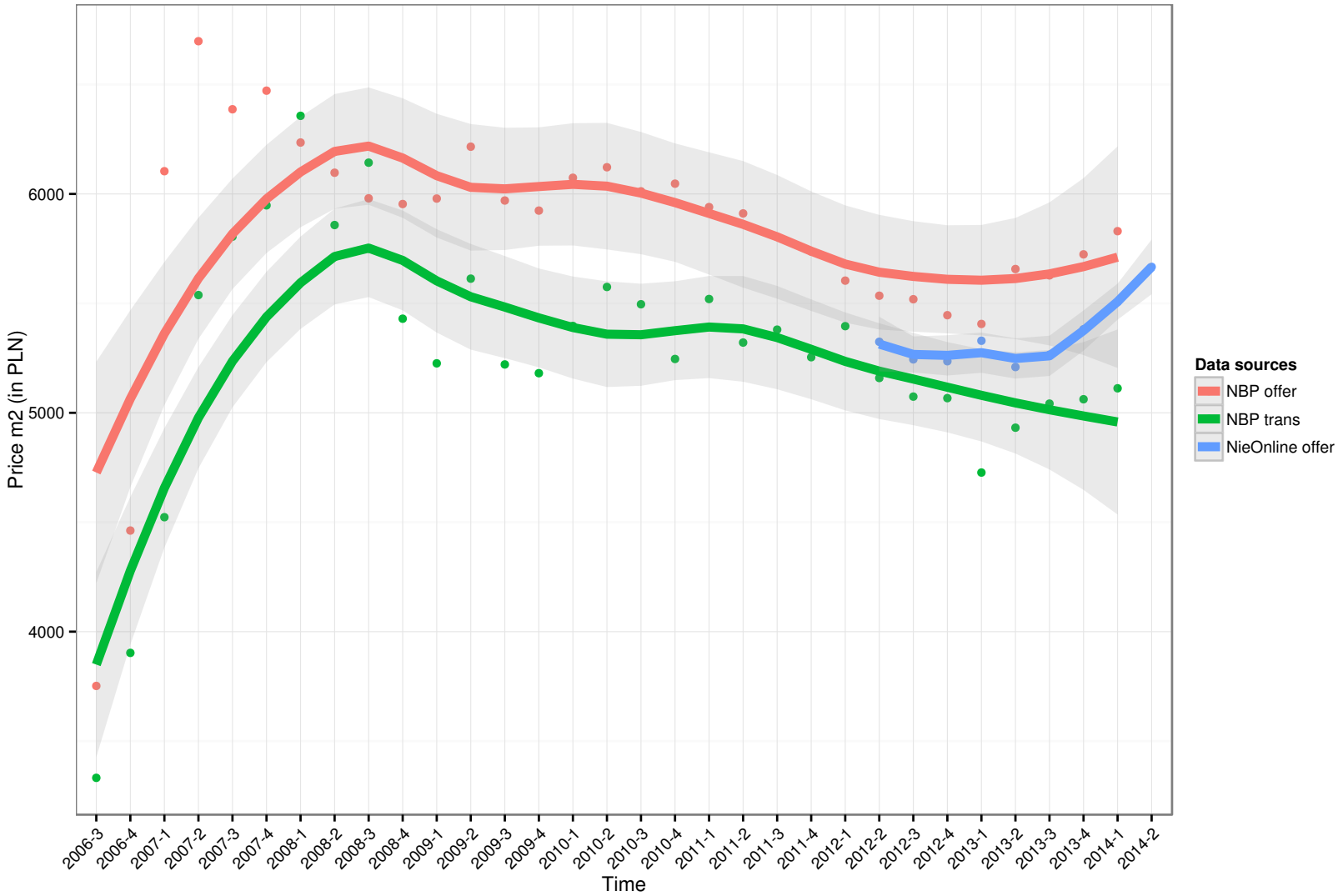
Selectivity



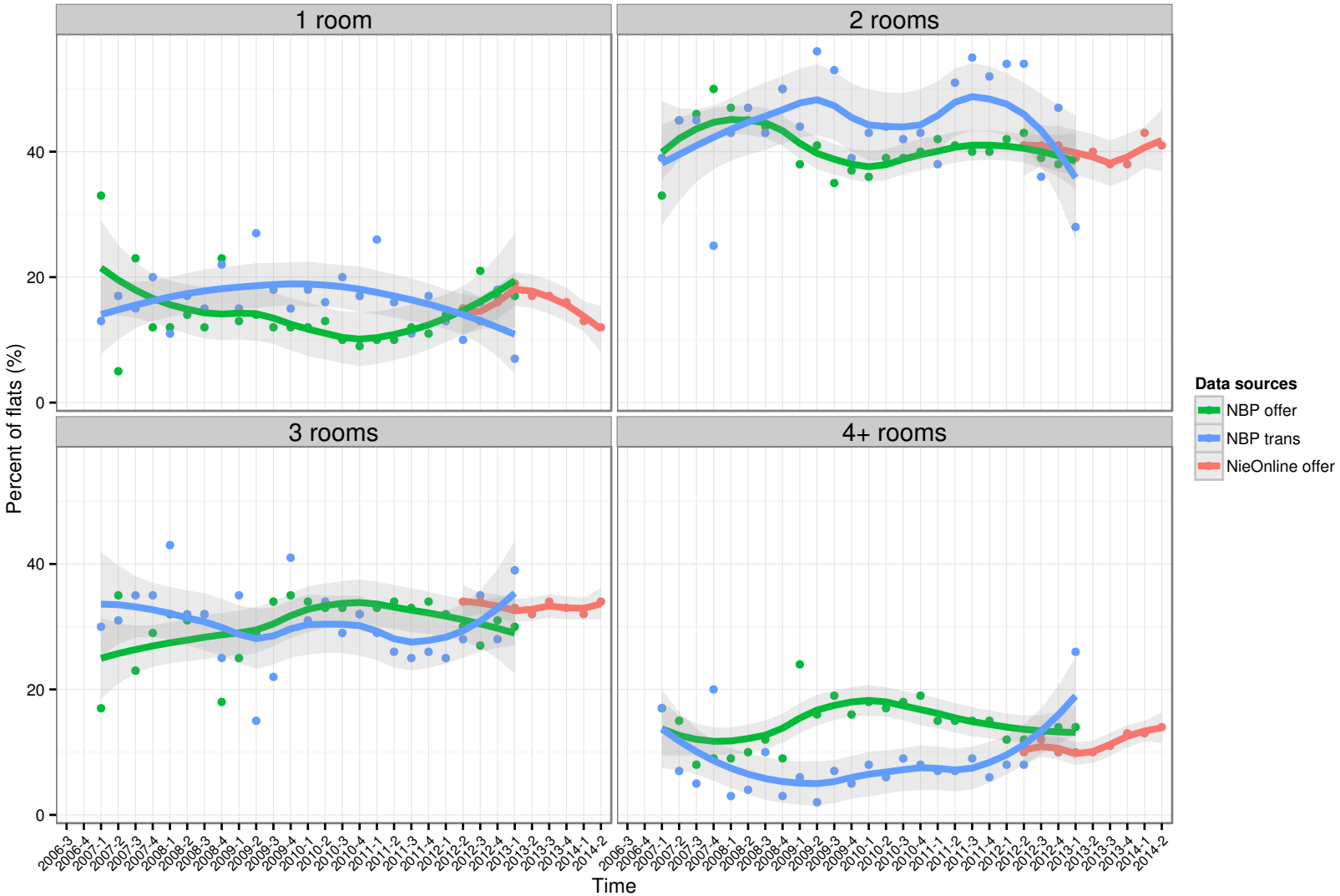




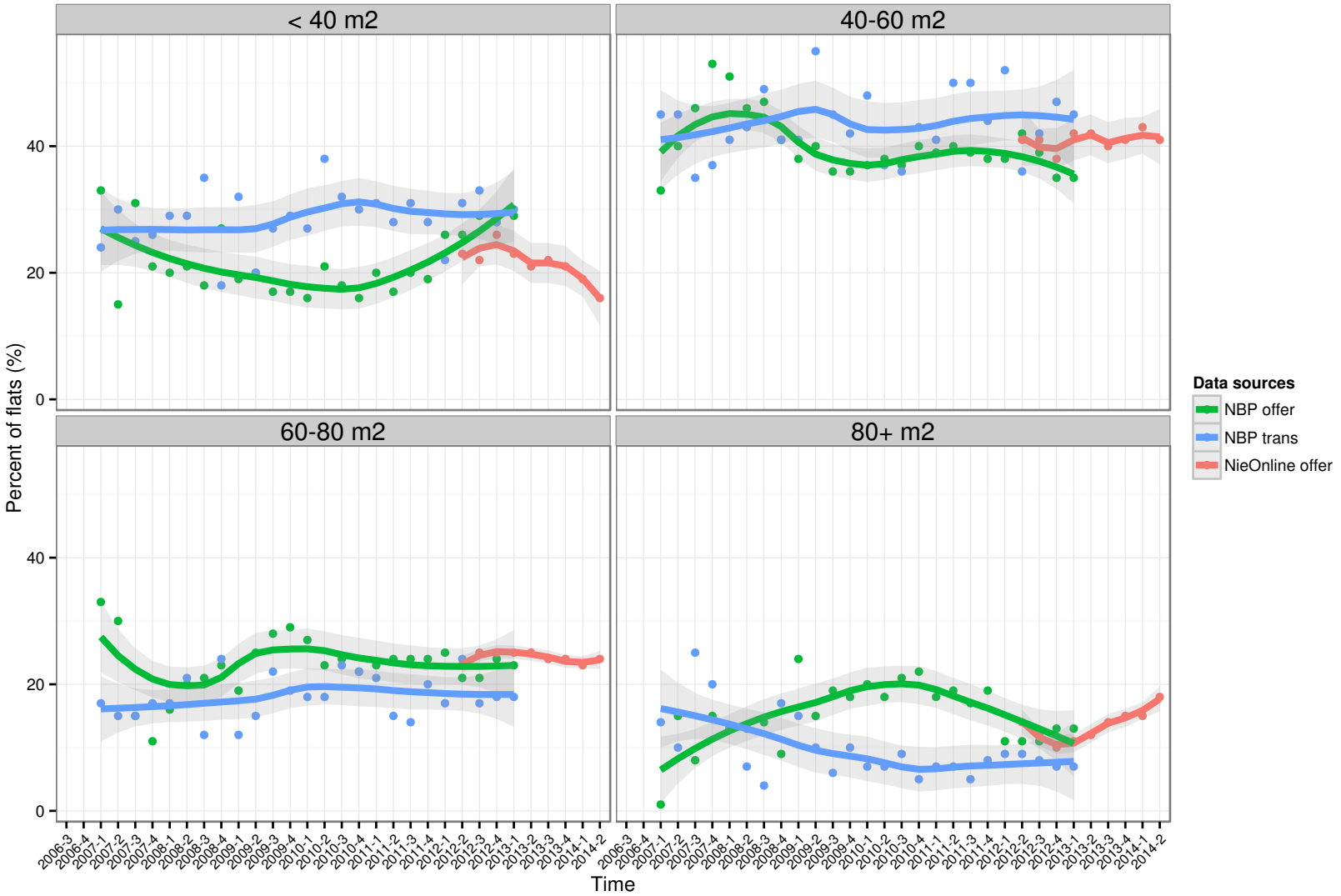
First results - Price m2 in Poznań, Poland



First results - Rooms in Poznań, Poland



First results - Surface in Poznań, Poland



Comments

- Selectivity is visible between portals,
- Undercoverage of small flats (<40 m2),
- Differences between offer prices reported by NBP and one of the portals (nieruchomosci-online.pl),
- Further assessment and comparison to other two datasets is needed,
- Data from 2013 is still not available ...

Thank you for your attention!

Contact

Maciej Beresewicz
maciej.beresewicz@ue.poznan.pl

Bibliography I

- Bapna R., Goes P., Gopal R., Marsden J.R., 2006, Moving from Data-Constrained to Data-Enabled Research: Experiences and Challenges in Collecting, Validating and Analyzing Large-Scale e-Commerce Data, Statistical Science, vol 21, nr 2, p.113-298
- Beimer P.P., Lyberg L.E (2003), Introduction to Survey Quality, Wiley Series in Survey Methodology, John Wiley & Sons, Inc, New York.
- Bethlehem, J. (2009). Applied Survey Methods, Wiley Series in Survey Methodology, John Wiley & Sons, Inc, New York.
- Bethlehem, J.G. (2008): Representativity of web surveys - an Illusion? In: Stoop, I. & Wittenberg, M., Access Panels and Online Research, Panacea or Pitfall? DANS Symposium Publications. Aksant, Amsterdam, pp. 19-44.
- Bethlehem, J.G., Biffignandi, S., 2012, Handbook of Web Surveys. John Wiley & Sons, Hoboken, NJ, USA.
- Beyer, M., 2011,. Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data. Gartner.
- Buelens, B., Daas, P., Burger, J., Puts, M., van den Brakel, J., 2013, Selectivity of Big Data. Internal report, Statistics Netherlands, Heerlen, The Netherlands.
- Buelens, B., et al (2012). Shifting paradigms in official statistics (pp. 1–21). The Hague/Herleen: Statistics Netherlands.
- 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. The Hague/Herleen: Statistics Netherlands.
- Daas, P.J.H., 2012, Secondary data collection, The Hague/Herleen: Statistics Netherlands.
- Daas, P.J.H., Puts, M.J.H., Buelens, B., van den Hurk, P.A.M., 2013, Big Data and Official Statistics, Konferencja NTTS, Bruksela, Belgia.

Bibliography II

- Daas, P.J.H., Puts, M.J.H., 2014a, Big Data as a Source of Statistical Information. The Survey Statistician 69, 22-31.
- Daas, P.J.H., Puts, M.J.H., 2014b, Sociale Media Sentiment and Consumer Confidence. Paper for the Workshop on using Big Data for Forecasting and Statistics, Frankfurt, Germany.
- Daas, P.J.H., Roos, M., van de Ven, M., Neroni, J., 2012, Twitter as a potential data source for statistics. Discussion paper 201221, Statistics Netherlands, The Hague/Heerlen, The Netherlands.
- Fellegi, I. Sunter, A., 1969, A Theory for Record Linkage, Journal of the American Statistical Association 64 (328): pp. 1183–1210
- Gołata E., 2009, Metodologia integracji danych w spisie opartym na rejestrach, W: Statystyka w praktyce społeczno-gospodarczej, red. J. Kolonko, W. Gamrot, AE Katowice.
- Groves, R.M., 2011, Three Eras of Survey Research. Public Opinion Quarterly 75 (5), 861-871.
- GUS, Warszawa, 2012, Społeczeństwo informacyjne w Polsce. Wyniki badań statystycznych z lat 2008-2012.
- GUS, Warszawa, 2014, Program badań statystycznych statystyki publicznej
- Hoekstra R., ten Bosch O., Harteveld F., 2010, Automated Data Collection from Web Sources for Official Statistics: First Experiences, Statistics Netherlands, Heerlen, The Netherlands.
- Kruskall W., Mosteller F., 1979a, Representative sampling, II: Scientific literature, excluding statistics, International Statistical Review, Vol. 47, No. 2 (Aug., 1979), pp. 111-127.
- Kruskall W., Mosteller F., 1979b, Representative sampling, III: The current statistical literature, International Statistical Review, Vol. 47, No. 3 (Dec., 1979), pp. 245-265.
- Kruskall W., Mosteller F., 1979c, Representative sampling, IV: The history of the concept in statistics, 1895-1939, International Statistical Review, Vol. 47, No. 3 (Dec., 1979), pp. 245-265.
- Laney, D., 2012, The Importance of 'Big Data': A Definition. Gartner.

Bibliography III

- Miller, G., 2011, Social Scientists Wade Into the Tweet Stream. Science 333 (6051), 1814-1815.
- NBP, Warszawa, 2013, Raport o sytuacji na rynku nieruchomości mieszkaniowych i komercyjnych w Polsce w 2012, Cykliczne raporty analityczne.
- Paradysz J., 2004, Zasilanie statystyki regionalnej za pomocą estymacji dla małych obszarów w perspektywie wykorzystania rejestrów administracyjnych, Wiadomości Statystyczne, nr 3, s. 1-9.
- Paradysz J., 2007, Rejestry administracyjne jako źródło zasilania w statystyce regionalnej w: Statystyka regionalna w jednoczącej się Europie (red. Paradysz), Poznań
- PBI/Gemius 2013, MegaPanel, stan na kwiecień 2014.
- Roszka W., 2013, Statystyczna integracja danych w badaniach społeczno - ekonomicznych, praca doktorska, UE Poznań.
- Rozporządzenie z dnia 29 marca 2001 r., Ministra Rozwoju Regionalnego i Budownictwa w sprawie ewidencji gruntów i budynków (Dz.U.2001.38.454).
- Schouten, B., Cobben, F., Bethlehem, J. (2009), Indictators of Representativeness of Survey Nonresponse. Survey Methodology 35, pp. 101-113.
- Shmueli, G., W. Jank, and R. Bapna, 2005, Sampling eCommerce Data from the Web: Methodological and Practical Issues, JSM, Minneapolis, MA, Proceedings of the American Statistical Association, Statistical Computing Section.
- United Nations, Geneva, 2000, Terminology on statistical metadata.
- Ustawa z dnia 17 maja 1989r. Prawo geodezyjne i kartograficzne, tekst jedn. Dz. U. z 2010 r. Nr 193, poz. 1287, z późn. zm.
- Wallgren A., Wallgren B., 2007, Register-based Statistics. Administrative Data for Statistical Purposes, John Wiley & Sons Ltd.

Bibliography IV

- Zhang, L.-C., 2011, A Unit-Error Theory for Register-Based Household Statistics, Journal of Official Statistics, vol. 27, no. 3, s. 415–432.
- Zhang L.-C., 2012, Topics of statistical theory for register-based statistics and data integration, Statistica Neerlandica, vol. 66, no. 1, s. 41–63.
- Zhang, L.-C., 2013, Population size estimation based on multiple lists. Uncertainty analysis for categorical data fusion, wykład w ramach projektu KdG, UE Poznań