Internet data sources for real estate market statistics

Maciej Beręsewicz
Department of Statistics, Poznań University of Economics

29 April 2015, Poznań
Meeting of Polish Statistical Society in Poznań
Faculty of Mathetmatics and Computer Science AMU

Outline

Outline

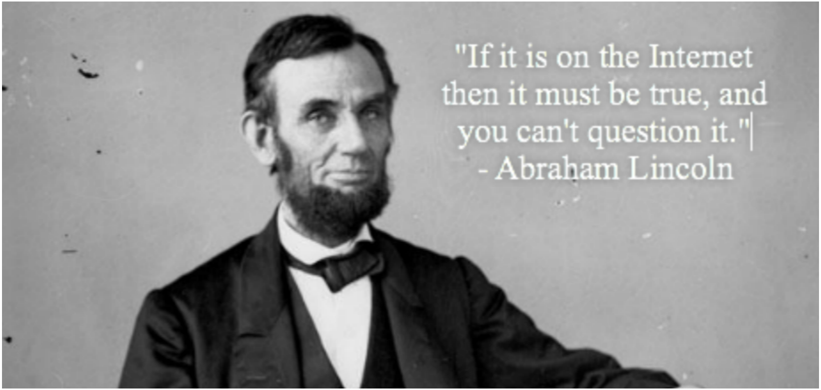
- Goals of the presentation
- Data sources for statistics
- New data sources for statistics
- Modelling bias in Internet data sources
- Representativeness of Internet data sources
- Summarising remarks
- References

The motto for the presentation

Statistics has value for its accuracy; without this essential quality it becomes null, even dangerous as it leads to error (Quetelet, 1846)

Like Olympic athletes, national statistical systems face unrelenting pressure for greater achievement, but unlike the Olympic motto—"Citius, Altius, Fortius"—the statistical challenge has rather more dimensions. The range of statistics needed grows ever wider, the level of geographical and other detail ever deeper, the timeliness ever quicker, and the demand for higher quality ever. All this, of course, and – with the relentless demand for greater effciency – ever cheaper. (Holt, 2007)

The motto for the presentation



Goals of the presentation

- The motivation for research on Internet data sources (IDS) in general, and in particular for the real estate market.
- Evaluation of bias and MSE of the selected variables obtained from IDS.
- Defintion and measurement of representativeness in the context of IDS.

Outline

Classical Data sources for (official) statistics

- Censuses – the classical one (full enumeration), dating back to 5000 BC,
- Surveys – probabilistic samples, complex surveys (complex designs), beginning of the 20th Century,
- Reporting – questionnaires filled by establishments (on monthly, quarterly basis; business surveys).

Classical data sources for (official) statistics - remarks

- Censuses - conducted every 5-10 years, full enumeration, time and cost consuming, results reported with delay, formally coverage of all units of the target population, not error-free,
- Surveys - conducted on monthly, quarterly basis, sample of the target population, selected topics (e.g. labour market, quality of life), sampling errors and, what is more important, non-sampling errors - non-response, high response burden, attrition (in panel surveys),
- Reporting - conducted on monthly, quarterly basis, formally full enumeration, high response burden, non-response, unwillingness to participate.

Outline

New data sources for statistics

- Official: Administrative sources – government register data, assumption: full coverage of the target population, in use since 1970s, multiple data sources (e.g., PESEL, REGON, NIP, VAT),
- Non-official – in particular the Internet and big data:
  - high volume, high variety, high velocity, high ...,
  - (wrong) assumption: full coverage of the population (coverage vary between data sources),
  - (wrong) assumption: $n \longrightarrow \infty \Rightarrow Bias(\breve{\theta}) \longrightarrow 0$,

New data sources for statistics - definitions

Definitions that can be found in statistical literature:

- Big data – *non-sampled data, characterized by the creation of databases from electronic sources which primary purpose is something other than statistical inference* (Horrigan, 2013).
- Organic Data – *collective assembling data by society on massive amounts of its behaviours which can be considered as aa ecosystem that is self-measuring in increasingly broad scope* (Groves, 2011).
- Internet data sources – *data collected and maintained by units external to statistical offices and administrative regulations available on the Internet (through web-based databases)*.

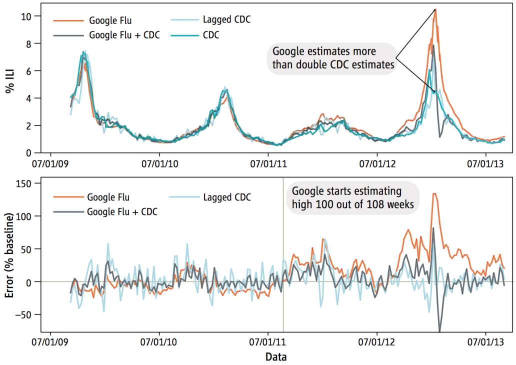
Official Statistics and new data sources

- Sensors and the Internet of Things.
- Mobile phone data (for tourism, population flows).
- Social media (Twitter postings).
- Web-scraping of prices to create price indexes.
- Use of Google Trends to predict official data (unemployment, tourism).

Selected research on new data sources

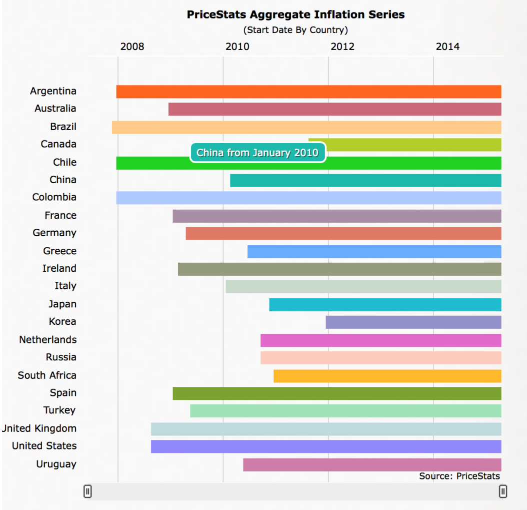
- Predicting unemployment - Fondeur & Karamé (2013), Xu, Li, Cheng, & Zheng (2012), Vicente, López-menéndez, & Pérez (2015)
- Opinions / Sentiment analysis - P. Daas, Roos, Ven, & Neroni (2012), P Daas & Puts (2014b), Miller (2011)
- Indexes - Vosen & Schmidt (2011), Cavallo (2012), Cavallo (2013)
- Representativeness and quality - Buelens, Daas, Burger, Puts, & Brakel (2014), P Daas & Puts (2014b)
- Source of information for small area estimation - Pratesi et al. (2013), Pratesi et al. (2014), Porter, Holan, Wikle, & Cressie (2013)
- General on new data sources - Choi & Varian (2012), Daas et al. (2011), P Daas & Puts (2014a), Hoekstra, Bosch, & Harteveld (2012), Ginsberg et al. (2008), Citro (2014), Ann Keller, Koonin, & Shipp (2012), Japec, Biemer, Decker, & Lane (2015)

Examples – Google Flu

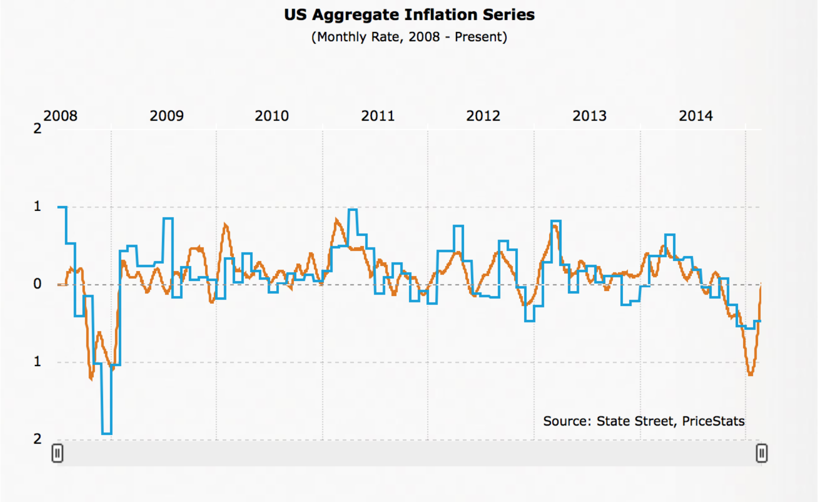


GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. **(Top)** Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. **(Bottom)** Error [as a percentage {[Non-CDC estmate)−(CDC estimate)]/(CDC) estimate)}. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

Examples – The Billion Price Project MIT (1)



Examples – The Billion Price Project MIT (2)



Examples – Twitter and Sentiment Analysis at Statistics Netherlands

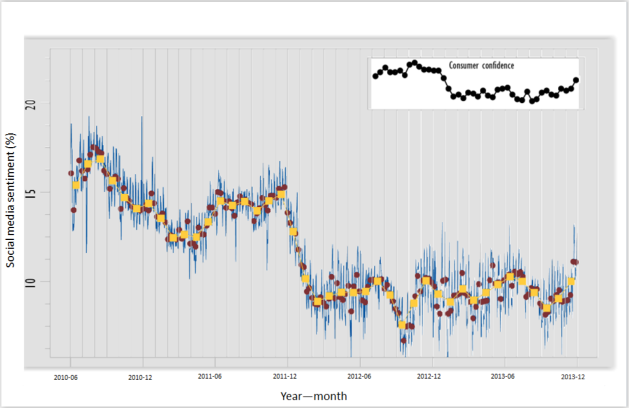


Figure 3. Social media sentiment (daily, weekly and monthly) in the Netherlands, June 2010 - November 2013. The development of consumer confidence for the same period is shown in the insert (Daas and Puts 2014).

Examples – Road sensors at Statistics Netherlands

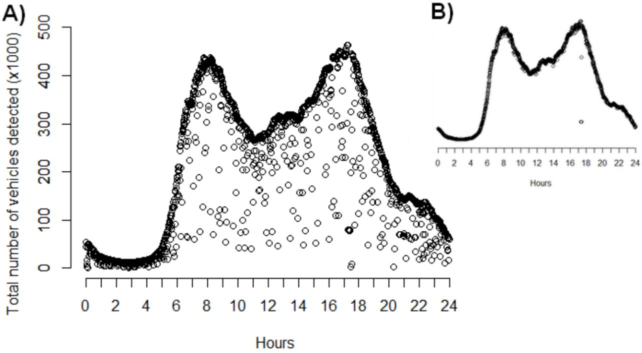


Figure 1. A) Total number of vehicles detected per minute in the Netherlands on December 1st, 2011. B) Results after correcting for missing data.

Statistical challenges in using Internet data sources and big data

- Formal and consistent approach to new data sources
- Coverage of target population (what fraction of population is observed in IDS?)
- Representativeness of these data (do we have reference/proxy official statistics?)
- Quality of these data (i.e. unit errors, missing data, outliers, measurement errors)
- Uncertainty of statistics based on these data (bias, variance)
- ...

The main question: Can we use these data to produce official statistics?

Statistical challenges - short classification

We classify big data and IDS problems into three groups:

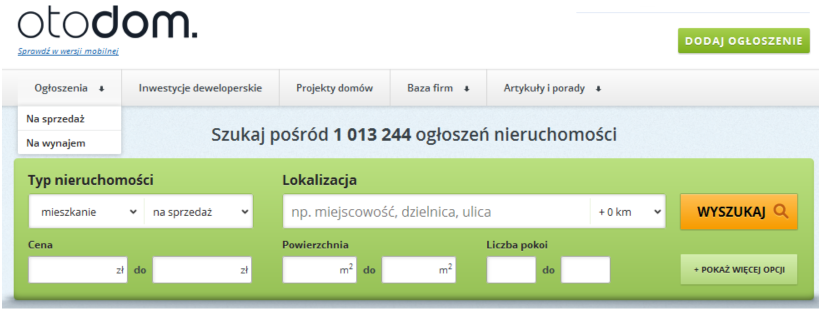
- Computational statistics – how to deal with big data sources; create efficient algorithms;
- Applied statistics – how to use Internet and big data sources for forecasting, creating and testing new methods or theories (e.g. socjology);
- Survey methodology – how to draw conclusions/interfere about general/target population using Internet and big data sources; measure bias and uncertainty; model-based approach to estimation;

Why real estate?

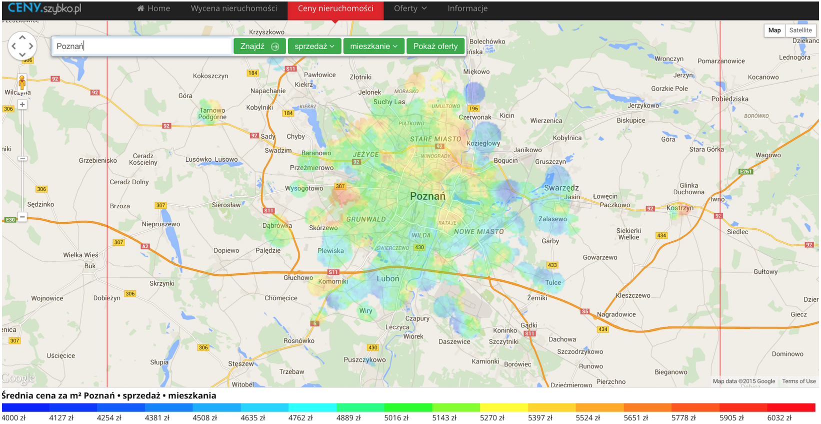
- Real estate market in Poland is only partially covered by official data – mainly surveys conducted by NBP and CSO; administrative data is also used (i.e. Register of real estate prices and values).
- Reports from these surveys are published with delay (i.e. report on 2015 will be published in autumn of 2016).
- Sources of data: survey of brokers, Register of prices and values of real estates (Pol. Rejestr Cen i Wartości Nieruchomości).
- Statistics Netherlands experiences in use of Internet data sources (in particular Funda.nl) for linking to register data and produce official statistics.

Possible source of information on real estate – the Internet.

The Internet portals (1)



The Internet portals (2)



Basic information about data sources

Data sources and variables

- **Data sources**: Dom.Gratka.pl, OtoDom.pl, Szybko.pl, Nieruchomości-Online.pl and Morizon.pl (Domy.pl/Oferty.net)
- **Reference data**: offer price obtained in research conducted by NBP with cooperation CSO in Poland.
- **Cities**: Białystok, Gdańsk, Katowice, Kraków, Lublin, Olsztyn, Opole, Poznań, Szczecin, Warszawa, Wrocław, Łódź (12 cities).
- **Time period**: 2012Q1 to 2014Q4 (12 periods).
- **Target population**: dwellings (flats) offered to sale
- **Target variables**: offer price per square meter, number of rooms (4 categories), floor area (4 categories) on the secondary market.

Basic information on data sources - number of observations

Table 1: Number of observations in selected IDS

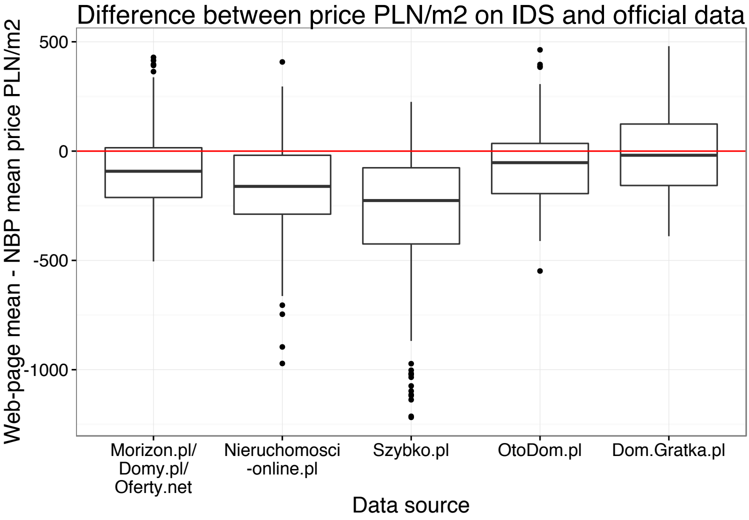
Quarter	Dom.Gratka	NieOnline	OtoDom
2011-1		10895	822597
2011-2		9229	807401
2011-3		8050	781936
2011-4		10686	808392
2012-1	685159	13955	856766
2012-2	683660	15135	847714
2012-3	689351	19946	864338
2012-4	692883	24448	956377
2013-1	661946	36312	994039
2013-2	503056	38171	967955
2013-3	610992	34459	904358
2013-4	565043	41585	872378
2014-1	531566	153562	890992
2014-2	518673	112121	1007559
2014-3	517590	58374	974099
2014-4	497451	98740	776500

Basic information on data sources - number of observations

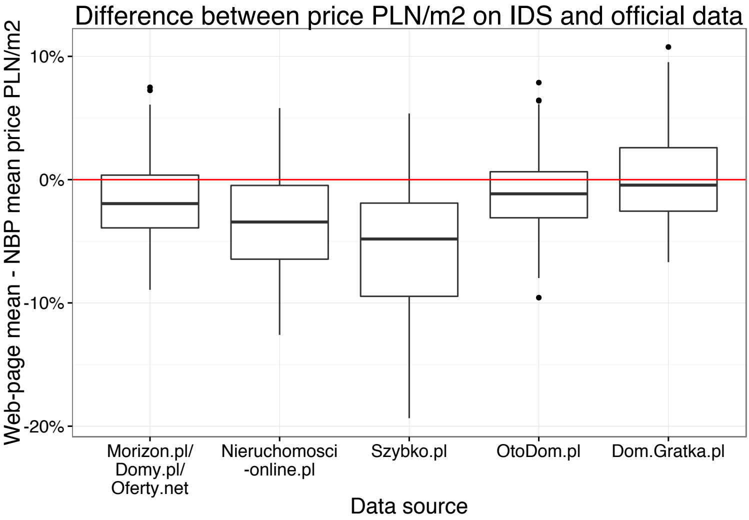
Table 2: Monthly distribution of number of observations for selected cities in Gratka, Otodom and Nieruchomosci-online

city	Min	Median	Mean	Max
BIAŁYSTOK	2	1232	2926	28705
GDAŃSK	46	7080	7285	22632
KATOWICE	21	1344	2643	12232
KRAKÓW	278	18561	19202	60007
LUBLIN	4	2282	3002	15054
OLSZTYN	9	1146	1754	6855
OPOLE	4	307	1026	5316
POZNAŃ	115	5678	7840	36600
SZCZECIN	75	5532	5684	16784
WARSZAWA	40	60623	53967	207395
WROCŁAW	112	13612	14618	52238
ŁÓDŹ	12	2662	4919	22821

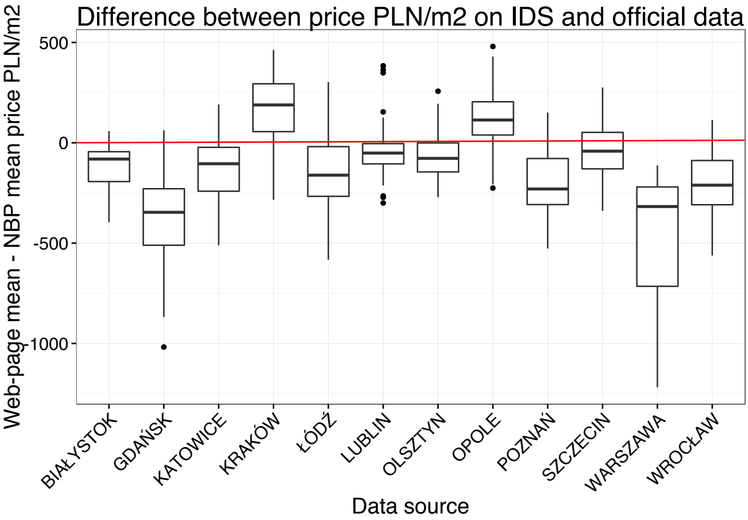
Basic information on data sources (DS) - bias in price m2



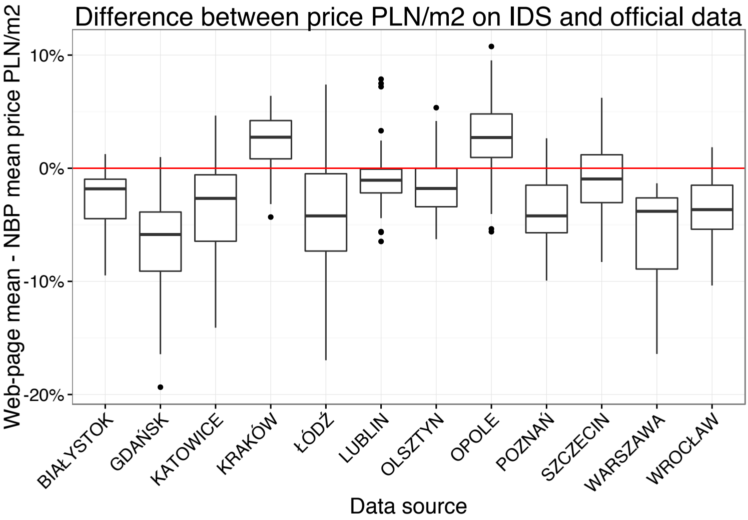
Basic information on data sources (DS) - relative bias in price m2



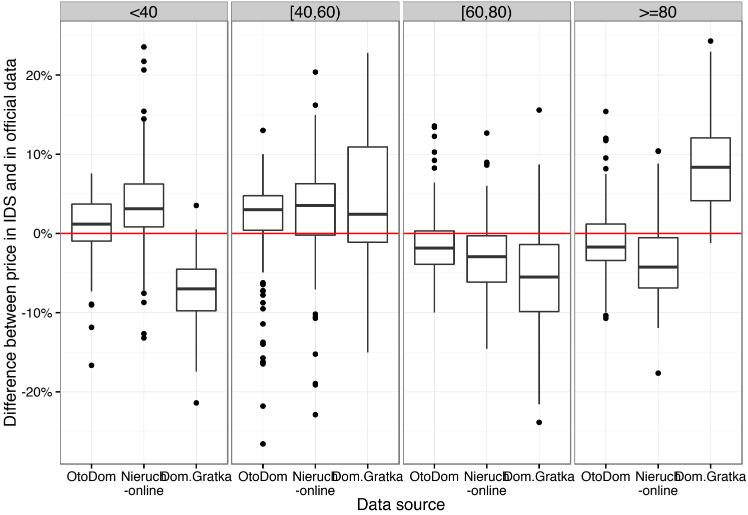
Basic information on data sources (DS) - bias in price m2 for cities



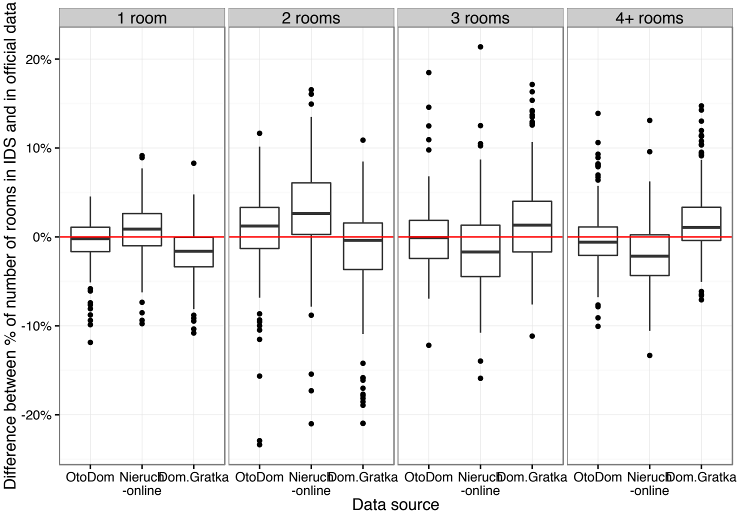
Basic information on data sources (DS) - relative bias in price m2 for cities



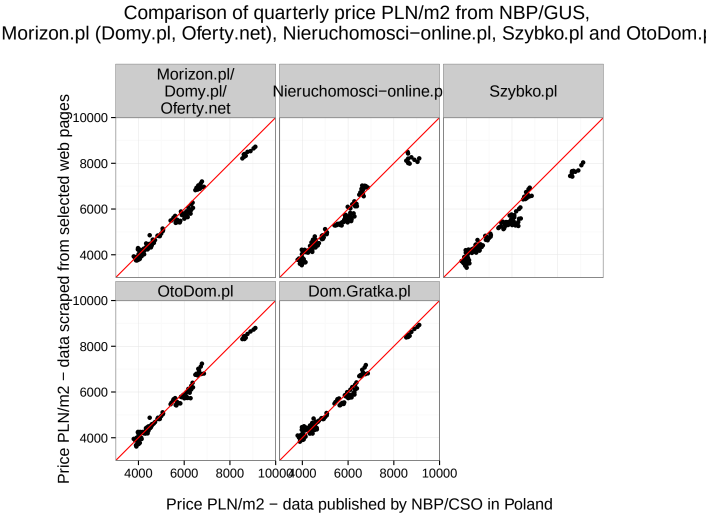
Basic information on DS - bias in % floor area ($m2$)



Basic information on DS - bias in % rooms



Basic information on data sources – price m2



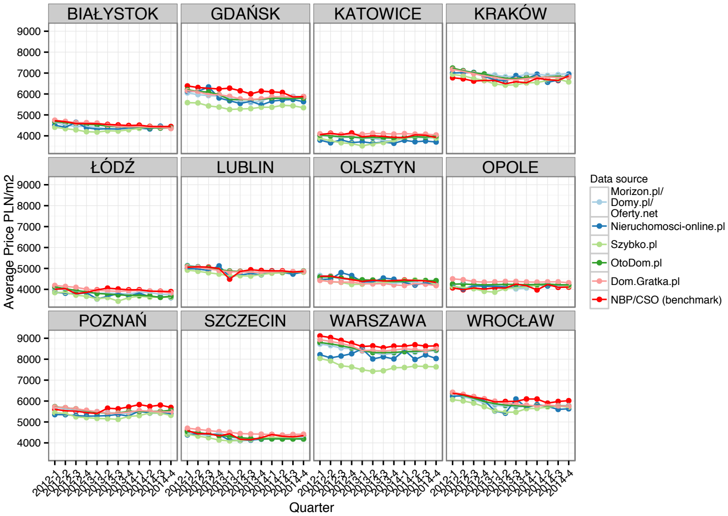
Basic information on data sources – price m2 (correlations)

Table 3: Correlations for price m2 between data sources

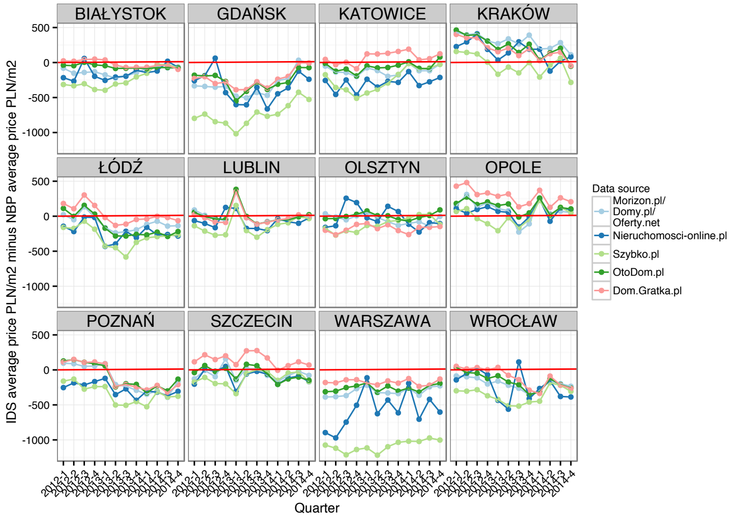
	nbp	morizon	nieonline	szybko	otodom	gratka_2
nbp	1.00	0.99	0.99	0.98	0.99	0.99
morizon	0.99	1.00	0.99	0.99	1.00	1.00
nieonline	0.99	0.99	1.00	0.99	0.99	0.99
szybko	0.98	0.99	0.99	1.00	0.99	0.99
otodom	0.99	1.00	0.99	0.99	1.00	1.00
gratka_2	0.99	1.00	0.99	0.99	1.00	1.00

Where morizon = Morizon.pl/Domy.pl/Oferty.net, nieonline = Nieruchomosci-online.pl, szybko = Szybko.pl, otodom = OtoDom.pl and gratka_2 = Dom.gratka.pl

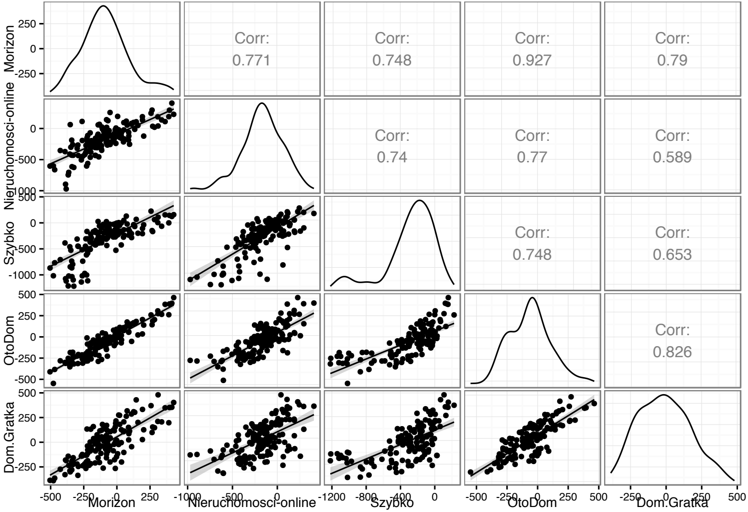
Basic information on data sources – Average quarterly price m^2



Basic information on data sources – differences between price m2 in IDS and in official data



Basic information on data sources – correlation between data sources and bias for price m2



Why there are differences?

Table 4: Descriptive statistics of sample sizes from NBP/CSO

City	Min	Mean	Median	Max
Warszawa	776	4912	5299	7909
Gdańsk	754	1423	1160	2493
Białystok	646	923	899	1164
Poznań	16	871	874	2383
Lublin	315	698	541	1668
Olsztyn	95	598	313	1692
Kraków	111	444	430	809
Wrocław	284	383	368	525
Łódź	81	375	155	1423
Szczecin	165	362	341	593
Katowice	114	170	169	240
Opole	15	32	33	51

Why there are differences?

- The NBP/CSO conduct a survey of brokers, who may give more information flats on offer.
- The most expensive flats are rarely placed on the Internet.
- Szybko.pl is not as popular as OtoDom or Morizon.
- Prices on secondary market in Warsaw are higher than on primary market which may indicate that offers placed on Szybko.pl may include misclassified units.

NBP/CSO reports on: mean price m^2, hedonic price index, flats by number of rooms and floor area, both on primary and secondary market (offers and transactions)

Outline

Why to model bias?

Why to model bias?

- Evaluate what is the systematic bias across all data sources (the IDS-based bias).
- Takes into account uncertainty of IDS-based statistics and survey based statistics, especially in case of small sample size.
- Include the estimated model-based bias in MSE of IDS-based statistics.

We extend approach proposed by Fosen & Zhang (2011) and L.-C. Zhang (2012) by taking into account time series, multiple data sources and we adopt the approach for the Internet data sources (which could also be useful for big data).

Modelling bias – notation

- U denotes the target population, ${}_RU$ denotes population U observed in the the administrative records (and we assume ${}_RU \subseteq U$), ${}_sU$ denotes population U observed in sample s and ${}_{IDS}U$ denote population U observed in the IDS. We assume ${}_{IDS}U \subset U$ where , $n_{s,t} \ll n_{IDS,t} < N_{U,t}$ where $n_{s,t}$ denotes sample size in time t, $n_{IDS,t}$ sample size for IDS in time t and $N_{U,t}$ denotes the population size in time t.
- $y, y_t, y_{d,t}$ – denote the target variable of interest (continuous, binary, ordinal etc.),
- $z, z_t, z_{d,t}$ – denote the IDS-based variable of interest $(z = y)$ or proxy variable with similar definition as $y, y_t, y_{d,t}$,
- $v, v_t, v_{d,t}$ – denote register-based variable or proxy variable with similar definition as $y, y_t, y_{d,t}$,

Modelling bias – notation

- We assume that for $y_t, y_{d,t}$ we have $z_t, z_{d,t}$ and $v_t, v_{d,t}$,
- $x_t, x_{d,t}$ – denote auxiliary variables that are found in all data sources (e.g. age, sex, floor area).
- $p_U(x_t, y_t)$ is the empirical cumulative distribution function (ECDF) of variable of interest in population at t
- $p_{RU}(x_t, v_t)$ is the ECDF based on the administrative sources which could be used for adjusting the $p_s(x_t, y_t)$ or $p_{IDS}(x_t, z_t)$.
- $p_s(x_t, y_t)$ is the ECDF based on the sample data in time,
- $p_{IDS}(x_t, z_t)$ is the ECDF based on the IDS,
- θ_t denotes the target characteristic of y in time t (e.g. mean, median, proportion, count).

Modelling bias – notation

- $\hat{\theta}_t(p_s(x_t, y_t); p_U(x_t, y_t))$ denotes the estimator of θ_t based on the sample data with the population data $p_U(x_t, y_t)$ as auxiliary information used for selecting the sample or adjusting weights (e.g. calibration).
- Now we consider the $\breve{\theta}_t$ as the estimator of θ_t based on IDS. We can consider following settings for this estimator:
  - $\breve{\theta}_t(p_{IDS}(x_t, z_t); p_s(x_t, y_t))$ – in the case when sample data $p_s(x_t, y_t)$ is used,
  - $\breve{\theta}_t(p_{IDS}(x_t, z_t); p_U(x_t, y_t))$ – in the case when population data $p_U(x_t, y_t)$ is used,
  - $\breve{\theta}_t(p_{IDS}(x_t, z_t); p_{U^*}(x_t, v_t))$ – in the case when administrative population data $p_{U^*}(x_t, v_t)$ is used,
  - $\breve{\theta}_t(p_{IDS}(x_t, z_t))$ – in the case when purely IDS-based statistics are considered with IDS data $p_{IDS}(x_t, z_t)$.
- In addition, we remember that $cov(\hat{\theta}_t, \hat{\theta}_{t-1}) > 0$ and $cov(\breve{\theta}_t, \breve{\theta}_{t-1}) > 0$.

Modelling bias – variance

- In general the variance of $\breve{\theta}_t$ can be given by following decomposition that take into account setting described on previous slide:
  - $V(\breve{\theta}_t) = E_s(V(\breve{\theta}_t|p_s(x_t, y_t))) + V_s(E(\breve{\theta}_t|p_s(x_t, y_t)))$,
- However, we will consider the conditional variance in the situation when the $p_{IDS}(x_t)$ is being fixed which is given by $V(\breve{\theta}_t) = V(\breve{\theta}_t(p_{IDS}(x_t, z_t)|p_{IDS}(x_t))) > 0$.
- We remember that $n \to +\infty \Rightarrow V(\theta_t) \to 0$ which in our case means that $V_s(\breve{\theta}_t) < V(\hat{\theta}_t)$ because $n_{s,t} \ll n_{IDS,t} < N_{U,t}$. However the bias of the $Bias(\hat{\theta}_t) \leq Bias(\breve{\theta}_t)$ could be large due to systematic errors, selectivity, undercoverage.

Modelling bias – bias

In general bias of $\breve{\theta}_t$ can be written as

$$Bias(\breve{\theta}_t) = E(\breve{\theta}_t) - \theta_t, \tag{1}$$

- We assume $cov(Bias(\breve{\theta}_t), Bias(\breve{\theta}_{t-1})) > 0$,
- Following L.-C. Zhang (2012) we have conditional and unconditional bias,
- We assume unconditional bias $p_{IDS}(x_t)$ is being fixed and $p_{IDS}(z_t)$ is random,
- If we have an sample-based or bias-adjusted register-based estimator $\hat{\theta}$ the estimator of $Bias(\breve{\theta}_t)$ is given by:

$$\widehat{Bias}(\breve{\theta}_t) = \breve{\theta}_t - \hat{\theta}_t, \tag{2}$$

- Estimator of MSE is given by $\widehat{MSE}(\breve{\theta}_t) = (\breve{\theta}_t - \hat{\theta}_t)^2 + V(\breve{\theta}_t)$

Modelling bias – why?

Direct estimates ($\hat{\theta}_t$) obtained from large samples (e.g. at the country level) could be reliable while for small domains ($\hat{\theta}_{d,t}$) could be unreliable due to high variance ($V_s(\hat{\theta}_{d,t})$) caused by small simple size ($n_{s,d,t}$). Relation between $V(\breve{\theta}_t)$ and $V(\hat{\theta}_t)$ at country level $V(\breve{\theta}_t) < V(\hat{\theta}_t)$ however, at the domain level could be even $V(\breve{\theta}_t) \ll V(\hat{\theta}_t)$. Therefore the modelling approach is needed in order to take into account uncertainty in the direct estimates in small domains.

Modelling bias – decomposition

We can decompose estimators $\breve{\theta}_t$ and $\hat{\theta}_t$ into:
In the case of time series data

$$\breve{\theta}_t = \underbrace{\theta_t}_{\text{True value}} + \underbrace{b_t}_{\text{bias}} + \zeta_t, \zeta_t \sim N(0, \sigma^2_{\zeta,t}) \quad (3)$$

$$\hat{\theta}_t = \theta_t + \epsilon_t, \epsilon_t \sim N(0, \sigma^2_{\epsilon,t}) \quad (4)$$

In the case of space-time data

$$\breve{\theta}_{d,t} = \underbrace{\theta_{d,t}}_{\text{True value}} + \underbrace{b_{d,t}}_{\text{bias}} + \zeta_{d,t}, \zeta_{d,t} \sim N(0, \sigma^2_{\zeta,d,t}) \quad (5)$$

$$\hat{\theta}_{d,t} = \theta_{d,t} + \epsilon_{d,t}, \epsilon_{d,t} \sim N(0, \sigma^2_{\epsilon,d,t}) \quad (6)$$

Modelling bias – decomposition

In order to estimate bias we calculate u_t and $u_{d,t}$

$$u_t = \breve{\theta}_t - \hat{\theta}_t = \theta_t + b_t + \zeta_t - (\theta_t + \epsilon_t) = b_t + \zeta_t - \epsilon_t = b_t + \zeta_t + e_t, \tag{7}$$

$$\begin{aligned} u_{d,t} = \breve{\theta}_{d,t} - \hat{\theta}_{d,t} = \\ \theta_{d,t} + b_{d,t} + \zeta_{d,t} - (\theta_{d,t} + \epsilon_{d,t}) = \\ b_{d,t} + \zeta_{d,t} - \epsilon_{d,t} = b_{d,t} + \zeta_{d,t} + e_{d,t}, \end{aligned} \tag{8}$$

where $e_t = -\epsilon_t$ and $e_{d,t} = -\epsilon_{d,t}$.

Modelling bias – decomposition

In the matrix notation we have

$$\breve{\theta} - \hat{\theta} = \boldsymbol{u} = \boldsymbol{b} + \zeta + \boldsymbol{e} \tag{9}$$

Following and extending L.-C. Zhang (2012) we consider $\boldsymbol{b}$ as mixed linear model which could be written as

$$\boldsymbol{b} = \boldsymbol{X}\beta + \boldsymbol{Z}\nu + \xi \tag{10}$$

in result we have

$$\breve{\theta} - \hat{\theta} = \boldsymbol{u} = \boldsymbol{X}\beta + \boldsymbol{Z}\nu + \xi + \zeta + \boldsymbol{e} \tag{11}$$

Modelling bias - I case

We start with a single data source with multiple domains. Let $\breve{\theta}_d$ denote estimator of θ for d domain and $\widehat{\theta}_d$ direct estimator for d domain. The model 11 will be:

$$u_d = \breve{\theta}_d - \widehat{\theta}_d = X_{i,d}\boldsymbol{\beta} + \nu_d + \xi_d + \zeta_u + e_d \tag{12}$$

Which in case of no auxiliary variables reduces to random intercept model given by:

$$u_d = \breve{\theta}_d - \widehat{\theta}_d = \beta_0 + \nu_d + \xi_d + \zeta_d + e_d = \beta + \nu_d + \omega_d \tag{13}$$

where ν_d refers to domain effect. $\widehat{\beta}$ could have interpretation as a systematic error caused by Internet data sources.

Modelling bias - I case

In result $\widehat{Bias}_{eblup}(\breve{\theta})$ will has the following form:

$$\widehat{Bias}_{eblup}(\breve{\theta}_d) = \hat{\gamma}(\breve{\theta}_d - \hat{\theta}_d) + (1 - \hat{\gamma})\hat{\beta} \tag{14}$$

where $\hat{\gamma} = \hat{\sigma}^2_{\nu}/(\hat{\sigma}^2_{\nu} + \hat{\sigma}^2_{\zeta,d} + \hat{\sigma}^2_{\epsilon,d})$ where $\sigma^2_{\zeta,d}$ denote estimator of variance for each $\breve{\theta}_d$ and $\sigma^2_{\epsilon,d}$ is estimated variance for each direct estimate $\hat{\theta}_d$.

Modelling bias - II case

Now, we consider multiple IDS denoted by k and for each we have information on d domain. In case of IDS we should note that ${}_{IDS_k}U \cap {}_{IDS_{k-1}}U \neq \emptyset$ therefore $cov(\breve{\theta}_{k,d}, \breve{\theta}_{k-1,d}) > 0$ which also leads to $cov(\breve{\theta}_{k,d} - \hat{\theta}_d, \breve{\theta}_{k-1,d} - \hat{\theta}_d) > 0$. In result, in case of no auxiliary variables, we obtain the following model

$$u_{k,d} = \breve{\theta}_{k,d} - \hat{\theta}_d = \beta_0 + \nu_d + \upsilon_k + \xi_{k,d} + \zeta_{k,d} + e_d = \hat{\beta} + \nu_d + \upsilon_k + \omega_{k,d}, \quad (15)$$

where ν_d denotes random effect for domain and $\nu \sim N(0, \sigma^2_\nu)$, υ_k denotes random for data source and $\boldsymbol{\upsilon} \sim N(0, \sigma^2_\upsilon \mathbf{D})$ and ω_d has the same interpretation as before.

Modelling bias - II case

Taking above into account $\widehat{Bias}_{eblup}(\breve{\theta}_{k,d})$ will be defined as in equation 14 however, $\hat{\gamma}$ will take into account variance of random effect for data source:

$$\hat{\gamma} = (\hat{\sigma}^2_{\nu} + \hat{\sigma}^2_{\upsilon})/(\hat{\sigma}^2_{\nu} + \hat{\sigma}^2_{\upsilon} + \hat{\sigma}^2_{\zeta,k,d} + \hat{\sigma}^2_{\epsilon,d}). \tag{16}$$

For $\hat{\sigma}^2_{\zeta,k,d}$ and $\hat{\sigma}^2_{\epsilon,d}$, we face similar situation as in the first case. EBLUP estimator of bias will be given by:

$$\widehat{Bias}_{eblup}(\breve{\theta}_{k,d}) = \hat{\gamma}(\breve{\theta}_{k,d} - \hat{\theta}_{k,d}) + (1 - \hat{\gamma})\hat{\beta} \tag{17}$$

Modelling bias - III case

The last case is when we have multiple data sources k and domains d that are observed at time t. Now we need to consider situation when not only intercept vary between data sources and domains but also slope. We could also take into account that interaction between data source and domain could be valid. Therefore, the model for u will be given by:

$$
\begin{aligned}
u_{k,d,t} = & \\
& = \breve{\theta}_{k,d,t} - \hat{\theta}_{d,t} = \\
& = \beta_0 + x_t\beta_1 + \nu_{d,t} + \upsilon_{k,t} + x_t\beta_{1,\nu} + x_t\beta_{1,\upsilon} + \xi_{k,d,t} + \zeta_{k,d,t} + e_{d,t} = \\
& = \beta_0 + \nu_{d,t} + \upsilon_{k,t} + (\beta_1 + \beta_{1,\nu} + \beta_{1,\upsilon})x_t + \omega_{k,d,t}.
\end{aligned} \tag{18}
$$

Modelling bias - III case

β_0 is a overall constant that could be explained as overall bias of using Internet as a data source, x denotes trend ($x = (1, 2, \ldots, t)^T$) and β_1 is a overall slope. β_1 can be interpreted as change in bias over time due to using Internet data sources.

Modelling bias - III case

$$\hat{\gamma} = (\hat{\sigma}^2_{\nu} + \hat{\sigma}^2_{\upsilon} + \hat{\sigma}^2_{\nu\upsilon})/(\hat{\sigma}^2_{\nu} + \hat{\sigma}^2_{\upsilon} + \hat{\sigma}^2_{\nu\upsilon} + \hat{\sigma}^2_{\zeta,k,d,t} + \hat{\sigma}^2_{\epsilon,d,t}). \quad (19)$$

Where $\hat{\sigma}^2_{\nu} = \hat{\sigma}^2_{\nu,int} + \hat{\sigma}^2_{\nu,slope}$, $\hat{\sigma}^2_{\upsilon} = \hat{\sigma}^2_{\upsilon,int} + \hat{\sigma}^2_{\upsilon,slope}$ and $\hat{\sigma}^2_{\nu\upsilon} = \hat{\sigma}^2_{\nu\upsilon,int} + \hat{\sigma}^2_{\nu\upsilon,slope}$. For $\hat{\sigma}^2_{\zeta,k,d,t}$ and $\hat{\sigma}^2_{\epsilon,d,t}$, we face similar situation as in the first case. EBLUP estimator of bias will be given by:

$$\widehat{Bias}_{eblup}(\breve{\theta}_{k,d,t}) = \hat{\gamma}(\breve{\theta}_{k,d,t} - \hat{\theta}_{k,d,t}) + (1-\hat{\gamma})\hat{\boldsymbol{\beta}}\mathbf{X}_t \quad (20)$$

where $\mathbf{X}_t$ denotes matrix of $\left[\mathbf{1}\boldsymbol{x}\right]$ where $x = (1, 2, \ldots, t)^T$.

Results of estimation – model

Estimated hierarchical model $(Y \sim 1 + (1 + T|city/page))$

Table 5: Model fitted

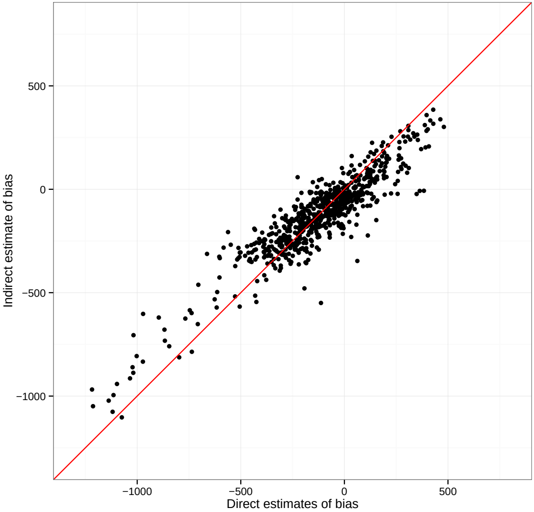
	Model 1
(Intercept)	-153.96^{**}
	(47.14)
AIC	9048.25
BIC	9084.89
Log Likelihood	-4516.13
Num. obs.	720
Num. groups: page:city	60
Num. groups: city	12
Variance: page:city.(Intercept)	39079.01
Variance: page:city.lp	71.47
Variance: city.(Intercept)	55167.33
Variance: city.lp	262.05
Variance: Residual	11535.59

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

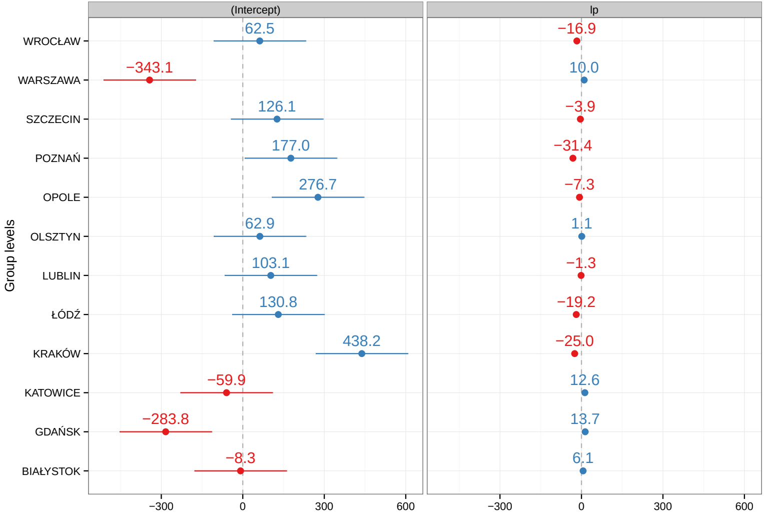
Results of estimation – model

- $\hat{\beta} = -153.95678$
- $\hat{\sigma}^2_{\nu,city} = \hat{\sigma}^2_{\nu,city,int} + \hat{\sigma}^2_{\nu,city,slope} = 5.5167334 \times 10^4 + 262.0470513 = 5.5429381 \times 10^4$
- $\hat{\sigma}^2_{\nu,page:city} = \hat{\sigma}^2_{\nu,page:city,int} + \hat{\sigma}^2_{\nu,page:city,slope} = 3.9079008 \times 10^4 + 71.4689079 = 3.9150477 \times 10^4$
- $\hat{\sigma}^2_{\omega} = 1.1535588 \times 10^4$
- $\hat{\gamma} = 0.8912921$ (when assuming constant $\hat{\sigma}^2_{\nu\upsilon} + \hat{\sigma}^2_{\zeta,k,d,t} + \hat{\sigma}^2_{\epsilon,d,t} \approx \hat{\sigma}^2_{\omega}$)

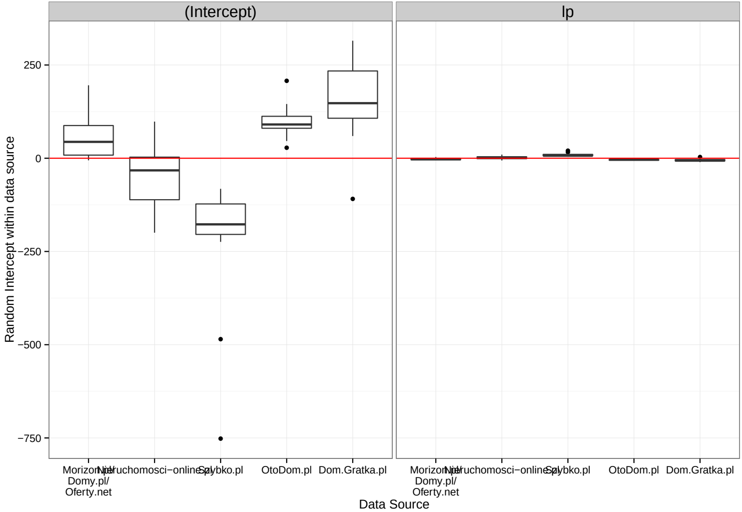
Results of estimation – fitted model



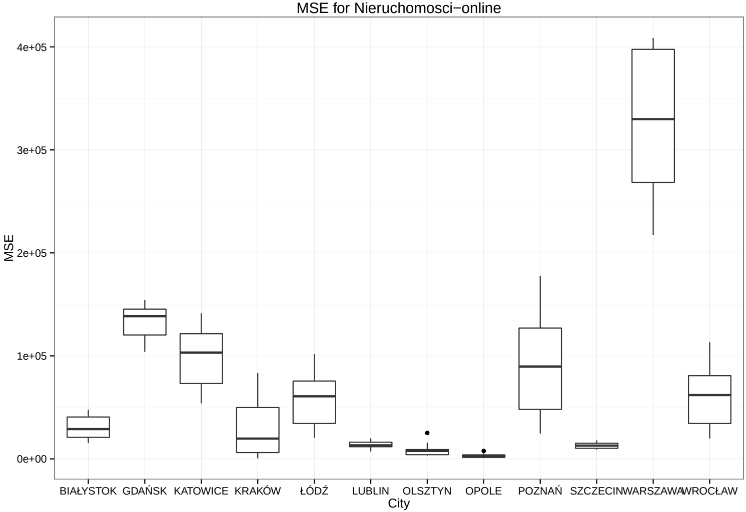
Random intercept and slope for city



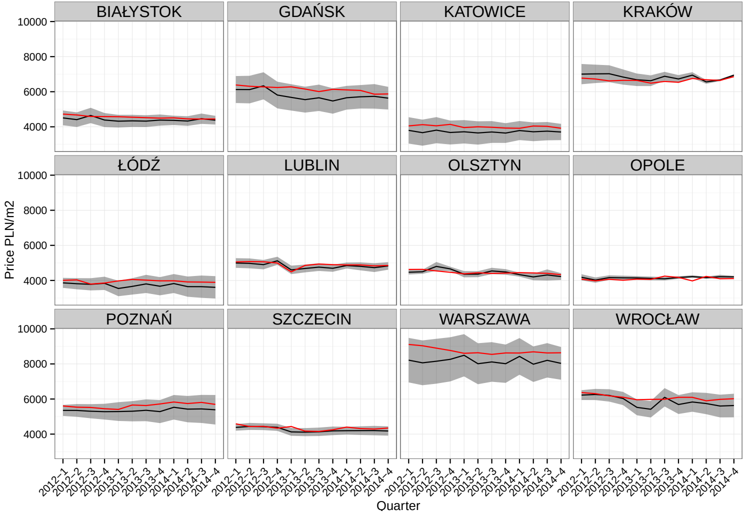
Random intercept and slope for page and city



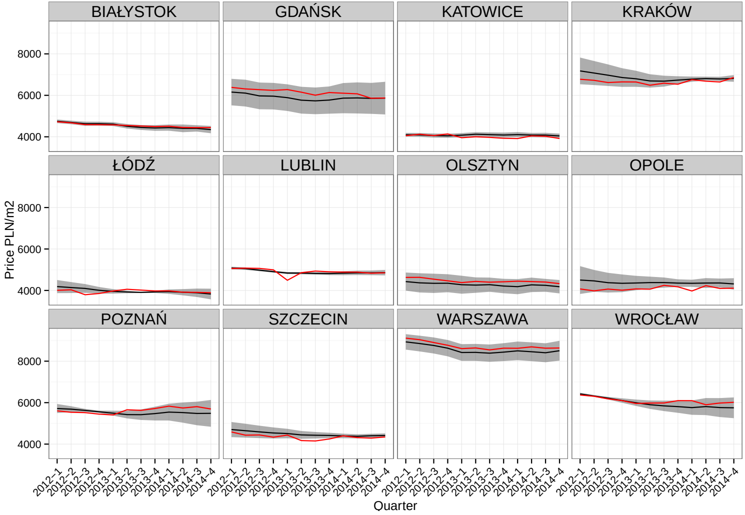
Results of estimation – model



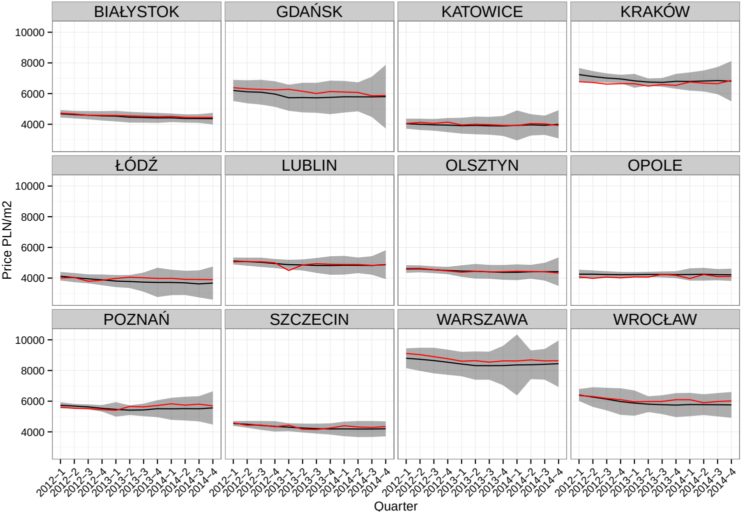
Results of estimation – model (nieruchomosci-online) CI



Results of estimation – model (gratka) CI



Results of estimation – model (otodom) CI



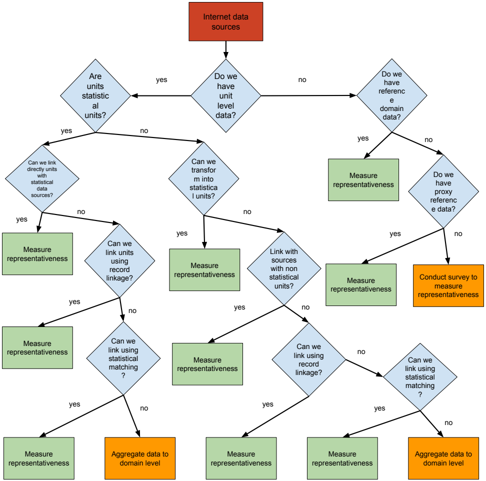
Outline

Concept of representativeness

The definitions of representative sampling listed by (Kruskal & Mosteller, 1979a, 1979b, 1979c) include:

- a general statement about data,
- lack of selective forces,
- miniature of population,
- typical elements of the population/representatives,
- reflects the variation in the population,
- a term used without an explanation,
- refers to a particular sampling method,
- enables good estimation,
- suitable for a particular purpose.

Concept of trend representativeness - diagram flow

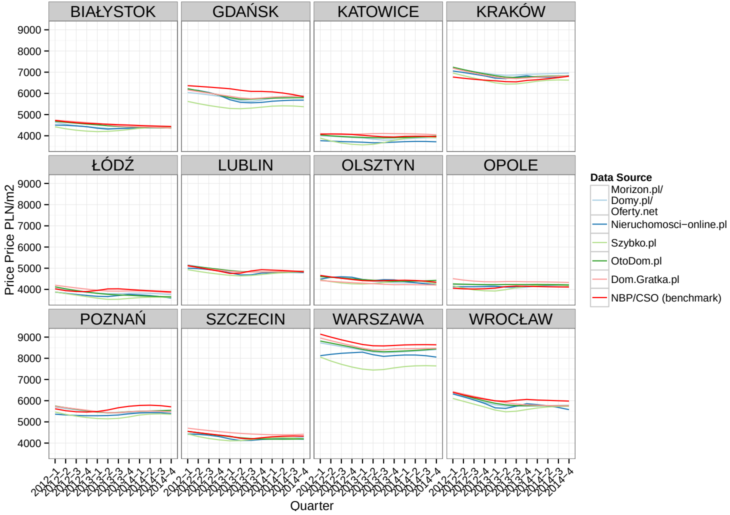


Representativeness - ggregated data

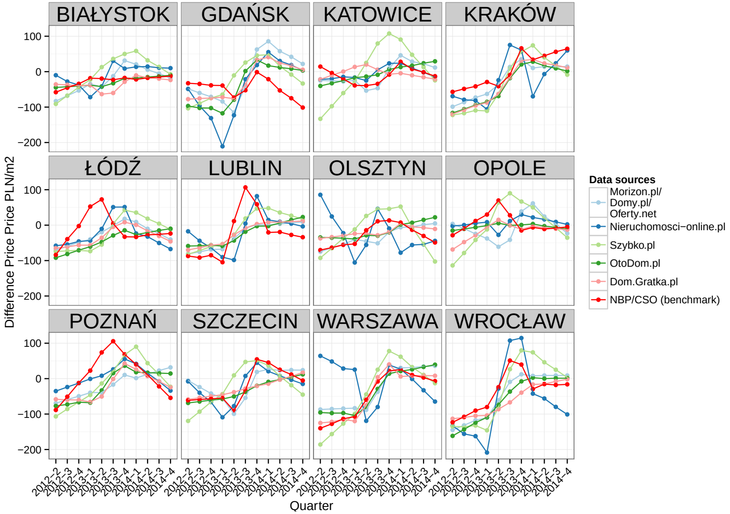
Due to bias of $\breve{\theta}$ and high uncertainty of $\hat{\theta}$ we propose measuring representativeness by comparing estimated {trends} in IDS and reference data. To achieve this purpose we will follow these steps:

- If possible reweight (poststratify, calibrate) to known totals/structures from official data.
- Estimate trends in $\breve{\theta}$ and $\hat{\theta}$ – method depends on the availability of data (how long is time series), we can use `loess`, `STL` or family of `ARIMA/X11-ARIMA`.
- Use temporal correlation and visual diagnostics to detect which time series correlate with official data.
- Use co-integration of trends to anwser the question whether the trends are representative in time.

Comparison of estimated LOESS trends



Comparison of first differences from estimated LOESS trends



Trend representativeness - correlation of differences

We use correlation between first differences proposed by Chouakria & Nagabhushan (2007) given by

$$CORT(\mathbf{X}_t, \mathbf{Y}_t) = \frac{\sum_{t=1}^{T-1}(X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{t=1}^{T-1}(X_{t+1} - X_t)^2}\sqrt{\sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2}}.$$

CORT for price - overall

Table 6: Correlation between data sources

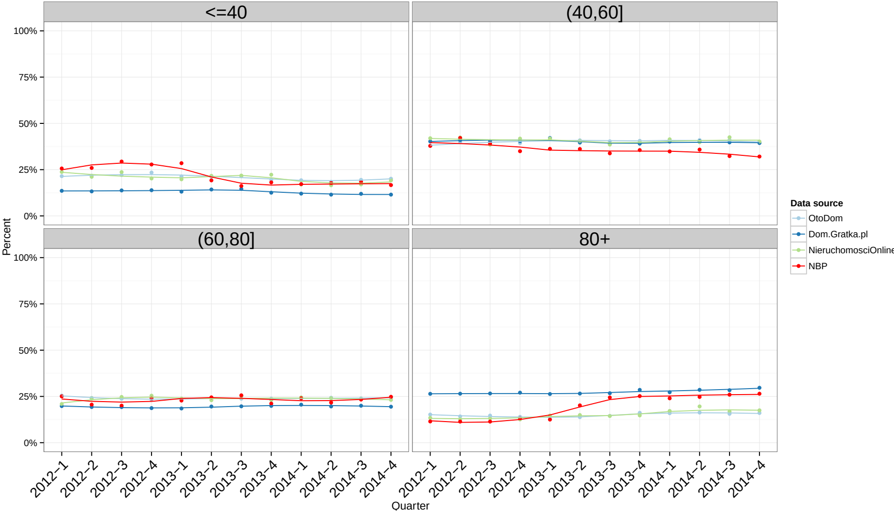
Morizon	Nieruchomosci	Szybko	OtoDom	Dom.Gratka	NBP
1.00	1.00	1.00	1.00	1.00	0.99
1.00	1.00	1.00	1.00	0.99	0.99
1.00	1.00	1.00	1.00	0.99	0.98
1.00	1.00	1.00	1.00	1.00	0.99
1.00	0.99	0.99	1.00	1.00	0.99
0.99	0.99	0.98	0.99	0.99	1.00

CORT for price - cities

Table 7: Correlation in time between data sources and NBP/CSO data for price m2

City	Morizon	Nieruchomosci	Szybko	OtoDom	Dom.Gratka
BIAŁYSTOK	0.80	0.36	0.42	0.94	0.84
GDAŃSK	0.25	0.43	0.40	0.46	0.44
KATOWICE	0.76	0.49	-0.04	0.27	-0.46
KRAKÓW	0.77	0.61	0.77	0.72	0.78
ŁÓDŹ	0.32	0.32	-0.06	0.28	0.22
LUBLIN	0.72	0.60	0.70	0.60	0.70
OLSZTYN	0.73	0.16	0.82	0.56	0.78
OPOLE	-0.75	-0.44	0.48	0.41	0.41
POZNAŃ	0.12	0.84	0.62	0.38	0.33
SZCZECIN	0.85	0.83	0.64	0.78	0.73
WARSZAWA	0.92	-0.24	0.96	0.94	0.99
WROCŁAW	0.91	0.92	0.85	0.88	0.77

Trend comparison for % floor area (Poznań)

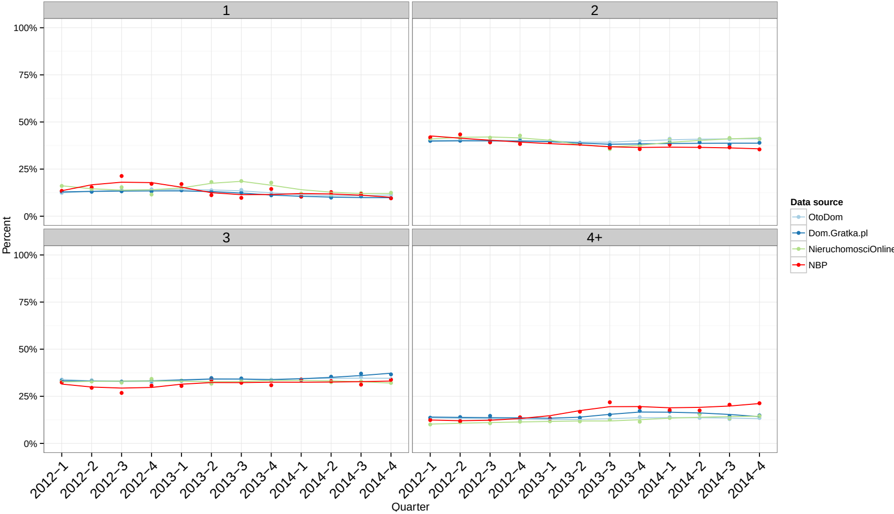


CORT for Poznań – % floor area

Table 8: CORT for floor area in Poznań

PANEL	OtoDom	Nieruchomosci	Dom.Gratka
<=40	-0.17	-0.03	-0.46
(40,60]	0.53	-0.09	0.35
(60,80]	0.22	0.11	0.30
80+	0.17	-0.13	0.30

Trend comparison for % number of rooms (Poznań)



CORT for Poznań – % floor rooms

Table 9: CORT for floor area in Poznań

PANEL	OtoDom	Nieruchomosci	Dom.Gratka
1	0.01	0.15	0.11
2	0.07	0.34	0.43
3	0.12	0.04	0.16
4+	-0.42	-0.04	0.03

Outline

Summary remarks

- Overall bias of estimation of average price m2 is $\hat{\beta} = -153.95678$ and slightly change over time.
- Modell based approach allowed to estimate MSE for selected IDS.
- The smallest bias can be observed in OtoDom and Dom.Gratka.pl, while the highest differences can be observed for Warsaw (~ -340 PLN/m2) and Kraków (~ 440).
- IDS are representative for price m2, however not representative for the fraction of flat area (4 groups) and number of rooms (4 groups).

Thank you for your attention!

Contact details

Maciej Beręsewicz
Department of Statistics
Poznań University of Economics
maciej.beresewicz@ue.poznan.pl

Outline

References I

Ann Keller, S., Koonin, S. E., & Shipp, S. (2012). Big data and city living–what can it do for us? *Significance*, *9*(4), 4–7.

Buelens, B., Daas, P., Burger, J., Puts, M., & Brakel, J. van den. (2014). Selectivity of big data. Statistics Netherlands. Retrieved from `http://www.pietdaas.nl/beta/pubs/pubs/Selectivity_Buelens.pdf`

Cavallo, A. (2012). Scraped data and sticky prices. *MIT Sloan Research Paper*. Retrieved from `http://www.mit.edu/%7Eafc/papers/Cavallo-Scraped.pdf`

Cavallo, A. (2013). Online and official price indexes: Measuring argentina's inflation. *Journal of Monetary Economics*, *60*(2), 152–165.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic Record*, *88*(s1), 2–9.

Chouakria, A. D., & Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, *1*(1), 5–21.

References II

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Meth*, *40*(2), 137–161.

Daas, P., & Puts, M. (2014a). Big data as a source of statistical information. *The Survey Statistician*, *69*, 22–31. Retrieved from `http://pietdaas.nl/beta/pubs/pubs/Big_data_survey_stat.pdf`

Daas, P., & Puts, M. (2014b). Social media sentiment and consumer confidence. European Central Bank. Retrieved from `http://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.pdf`

Daas, P., Roos, M., Blois, C. de, Hoekstra, R., Bosch, O. ten, & Ma, Y. (2011). New data sources for statistics: Experiences at statistics netherlands. In *Paper for the 2011 european new technique and technologies for statistics conference, february* (pp. 22–24).

Daas, P., Roos, M., Ven, M. van de, & Neroni, J. (2012). Twitter as a potential data source for statistics. Statistics Netherlands. Retrieved from `http://pietdaas.nl/beta/pubs/pubs/DiscPaper_Twitter.pdf`

References III

Fondeur, Y., & Karamé, F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling*, *30*, 117–125. http://doi.org/10.1016/j.econmod.2012.07.017

Fosen, J., & Zhang, L.-c. (2011). *The approach to quality evaluation of the micro-integrated employment statistics*.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014.

Groves, R. M. (2011). Three Eras of Survey Research. *Public Opinion Quarterly*, *75*(5), 861–871. http://doi.org/10.1093/poq/nfr057

Hoekstra, R., Bosch, O. ten, & 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.

Holt, D. T. (2007). The Official Statistics Olympic Challenge. *The American Statistician*, *61*(1), 1–8. http://doi.org/10.1198/000313007X168173

References IV

Horrigan, M. W. (2013). Big data: A perspective from the bLS. Big Data: A Perspective from the BLS.

Japec, L., Biemer, P., Decker, P., & Lane, J. (2015). AAPOR Report on Big Data AAPOR Big Data Task Force.

Kruskal, W., & Mosteller, F. (1979a). Representative sampling i: Non-scientific literature. *International Statistical Review*, *47*, 13–24. Retrieved from `http://www.jstor.org/stable/1402564`

Kruskal, W., & Mosteller, F. (1979b). Representative sampling iI: Scientific literature excluding statistics. *International Statistical Review*, *47*, 111–123. Retrieved from `http://www.jstor.org/stable/1402564`

Kruskal, W., & Mosteller, F. (1979c). Representative sampling III: The current statistical literature. *International Statistical Review*, *47*, 245–265. Retrieved from `http://www.jstor.org/stable/1402647`

Miller, G. (2011). Social scientists wade into the tweet stream. *Science*, *333*(6051), 1814–1815.

References V

Porter, A. T., Holan, S. H., Wikle, C. K., & Cressie, N. (2013). Spatial fay-herriot models for small area estimation with functional covariates. *ArXiv Preprint ArXiv:1303.6668*.

Pratesi, M., Giannotti, F., Giusti, C., Marchetti, S., Pedreschi, D., & Salvati, N. (2014). *Area level sae models with measurement errors in covariates: An application to sample surveys and big data sources*. Retrieved from `http://sae2014.ue.poznan.pl/SAE2014_book.pdf`

Pratesi, M., Pedreschi, D., Giannotti, F., Marchetti, S., Salvati, N., & Maggino, F. (2013). *Small area model-based estimators using big data sources*. Retrieved from `http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_208.pdf`

Quetelet, A. (1846). *Lettres à sAR le duc régnant de saxe-cobourg et gotha, sur la théorie des probabilités: Appliquée aux sciences morales et politiques*. Hayez.

Vicente, M. R., López-menéndez, A. J., & Pérez, R. (2015). Technological Forecasting & Social Change Forecasting unemployment with internet search data : Does it help to improve predictions when job destruction is

References VI

skyrocketing ? *Technological Forecasting & Social Change*, *92*, 132–139. http://doi.org/10.1016/j.techfore.2014.12.005

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: Survey-based indicators vs. google trends. *Journal of Forecasting*, *30*(6), 565–578.

Xu, W., Li, Z., Cheng, C., & Zheng, T. (2012). Data mining for unemployment rate prediction using search engine query data. *Service Oriented Computing and Applications*, *7*(1), 33–42. http://doi.org/10.1007/s11761-012-0122-2

Zhang, L.-C. (2012). *On the accuracy of register-based census employment statistics*. Retrieved from http://www.q2012.gr/articlefiles/sessions/23.4_Zhang_AaccuracyRegisterStatistics.pdf