Measuring representativeness of Internet data sources through linking with register data

Maciej Beręsewicz

Department of Statistics
Poznan University of Economics and Business, Poland

Centre for Small Area Estimation
Statistical Office in Poznan, Poland

Madrid, May 31 - June 3

POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Q2016
European Conference on
Quality in Official Statistics

Outline of the presentation

1. Introduction
2. Internet data sources
3. The concept of representativeness
4. The conceptual framework of the study
5. Selected results
6. Conclusions

Introduction – motivation

- Increasing unit non-response in sample surveys.
- Growing information needs at a low level of (spatial) aggregation.
- *A change of paradigm in official statistics*, which involves the adoption of existing data sources instead of creating new ones.
- Internet data sources (IDSs) and big data are still not recognized and their *suitability as statistical sources is often unknown*.
- *New data sources*, in particular big data and the Internet have become *an important issue* in Official Statistics (Daas et al., 2015).
- 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 (Citro, 2014).
- The Internet plays an important role in the housing market, as a source of information for potential buyers, for price and demand forecasting.

Internet data sources – examples

- Social media,
- E-commerce / advertisement services,
- Price comparison websites,
- Google Trends,
- . . .

Internet data sources – an Internet/opt-in survey?

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.

- Despite its volume, an IDS should be treated as a sample.
- Unlike official statistics, which are based on probability selection mechanisms, IDS are the result of the self-selection process.
- The definition specifies that an IDS only refers to data created by Internet users or by private entities themselves.

Moreover, we argue that **big data** are *another* type secondary data source that have similar characteristics to an opt-in/self-selection Internet survey.

Internet data sources

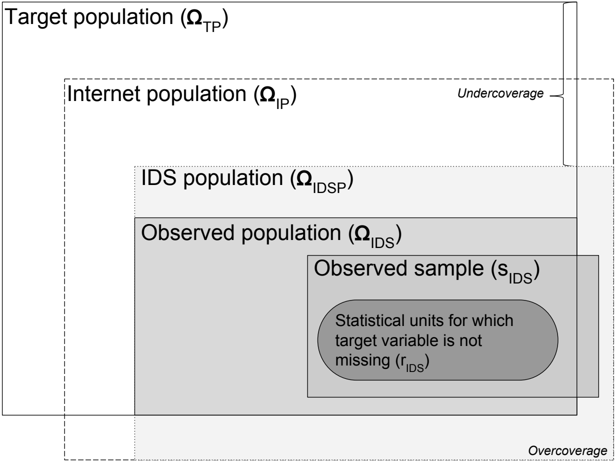
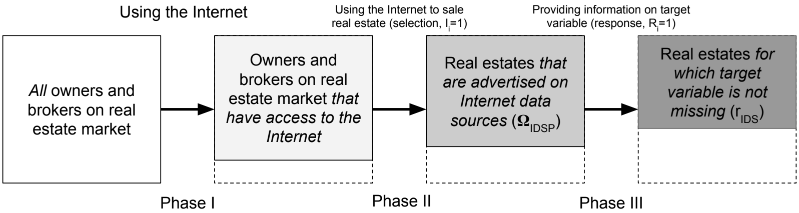

Target population ($\mathbf{\Omega}_{TP}$)
Internet population ($\mathbf{\Omega}_{IP}$)
Undercoverage
IDS population ($\mathbf{\Omega}_{IDSP}$)
Observed population ($\mathbf{\Omega}_{IDS}$)
Observed sample (s_{IDS})
Statistical units for which target variable is not missing (r_{IDS})
Overcoverage

Figure 1: The relation between the target and IDS population

Internet data sources – the self-selection mechanism



Using the Internet

Using the Internet to sale real estate (selection, I_i=1)

Providing information on target variable (response, R_i=1)

All owners and brokers on real estate market

Owners and brokers on real estate market *that have access to the Internet*

Real estates *that are advertised on Internet data sources* ($\mathbf{\Omega}_{IDSP}$)

Real estates *for which target variable is not missing* (r_{IDS})

Phase I

Phase II

Phase III

Figure 2: The self-selection mechanism underlying Internet data sources about the secondary real estate market

The concept of representativeness

It should be underlined that the concept of representativeness (Kruskal and Mosteller, 1979a,b,c) is still valid, even in the era of big data.

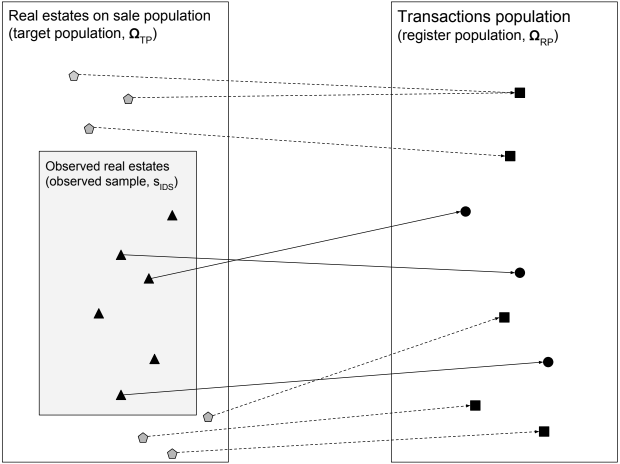
To measure representativeness we should identify **the self-selection mechanism**, and its results:

1. coverage of the target population,
2. discrepancies between sample and population distribution,
3. impact on estimation of finite population characteristics.

Missing at Random or **Not Missing at Random** pattern (Rubin, 1976)?

How we can identify the self-selection mechanism?

The conceptual framework of the empirical study



Real estates on sale population (target population, $\mathbf{\Omega}_{TP}$)

Transactions population (register population, $\mathbf{\Omega}_{RP}$)

Observed real estates (observed sample, s_{IDS})

Figure 3: The relation between the register and the IDS population

The conceptual framework of the empirical study

- The empirical study is based on data from one IDS (Nieruchomosci-Online.pl), which was selected because of the availability of individual data.
- Independent data source: *the Register of Real Estate Prices and Values*.
- Probabilistic methods should be applied to determine the selection mechanism.
- The following variables were used for the linkage:
  - floor area [m2],
  - the number of available rooms,
  - the floor number,
  - the number of floors in the building,
  - the offer price (from IDS)
  - the transaction price (from the register).
- The threshold of probability that two records refer to the same (or similar) unit was set at 60%.

Discrepancies of distribution of property prices for Warsaw

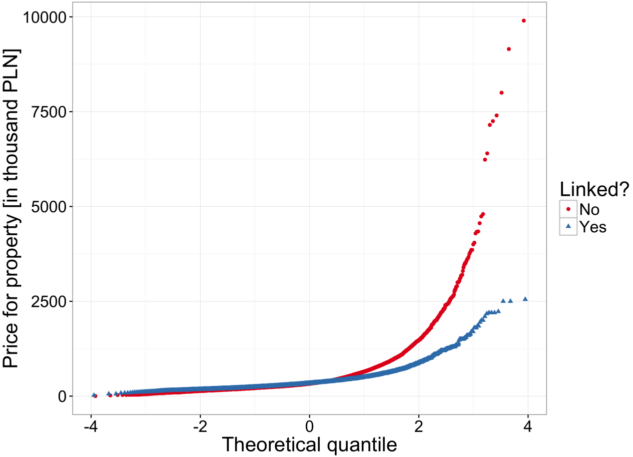


Figure 4: A comparison of the distributions of property prices for linked and non-lined units in the secondary real estate market in Warsaw between 2012 and 2014

Discrepancies of distribution of property prices for Poznań

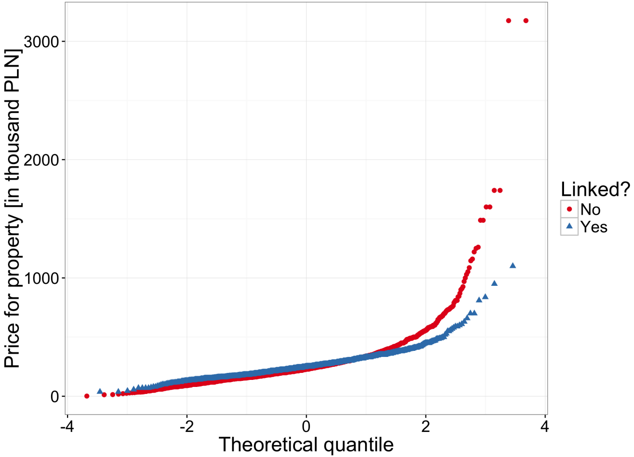


Figure 5: A comparison of the distributions of property prices for linked and non-lined units in the secondary real estate market in Poznań between 2012 and 2014

Comparison of basic descriptive statistics for price

Table 1: Descriptive statistics for price and price m2 for Warsaw and Poznań

City	Linked?	Mean price	Median price	Average price/m2
Warsaw	Yes	390 199	346 000	7 860
	No	476 602	355 000	8 284
	All	*429 271*	*350 000*	*8 067*
Poznań	Yes	265 333	253 000	5 294
	No	264 852	240 000	4 714
	All	*265 192*	*245 000*	*4 871*

Despite differences in distributions, the bias of the average price for m2 is small in both cases (does not exceed 10%).

Conclusions

- Despite the advantage of IDSs large size, they are still non-probabilistic, self-selected samples.
- Recognizing this problem is the starting point in the search for appropriate methods that can be applied in order to reduce the bias of estimates based on these sources.
- Internet data sources in both cases truncate the right tail of price distribution.
- Results show differences between properties observed online in Warsaw and Poznan in comparison to the register of transactions.

References

- Bethlehem, J. (2010). Selection bias in web surveys. International Statistical Review, 78(2), 161-188.
- Daas, P. J., Puts, M. J., Buelens, B., and van den Hurk, P. A. (2015). Big Data as a source for official statistics. Journal of Official Statistics, 31(2), 249-262.
- Kruskal, W., and Mosteller, F. (1979a). Representative sampling I: Non-scientific literature. International Statistical Review, 47, 13- 24.
- Kruskal, W., and Mosteller, F. (1979b). Representative sampling II: Scientific literature excluding statistics. International Statistical Review, 47, 111-123.
- Kruskal, W., and Mosteller, F. (1979c). Representative sampling III: Current statistical literature. International Statistical Review, 47, 245-265.
- Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?. Survey Methodology, 37(2), 115-136.
- Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.
- Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.
- Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the representativeness of survey response. Survey Methodology, 35(1), 101-113

Thank you for the attention!