Measuring representativeness of Internet data sources through linkage with register data

Maciej Beręsewicz[1,2]

[1] *Poznań University of Economics and Business, Poznań, Poland*, *maciej.beresewicz@ue.poznan.pl*
[2] *Statistical Office, Poznań, Poland*

Abstract

The main goal of the paper is to assess the representativeness of Internet data sources through probability linkage with register data. To achieve this aim selected websites publishing ads for residential real estates available in Poland were used and linked with the Register of Prices and Values of the Real Estate Market. The main underlying assumption was that an independent source that fully covers a similar target population is available. In addition, the adoption of R-indicators is discussed. A detailed description and results of the study are presented using data for the capital city of Poland Warsaw and the city of Poznań.
The paper and the research have been financed by the National Science Centre, Poland, Preludium 7 grant no. 2014/13/N/HS4/02999.

Keywords: Internet data sources, representativeness, probabilistic record linkage.

1. Introduction

New data sources, in particular big data and the Internet have become an important issue in Official Statistics (Daas et al., 2015). The non-statistical character of these data sources requires that they should be assessed before they can be incorporated in the statistical system. Moreover, despite their size, new data sources are in fact non-probability samples that suffer from the self-selection error. Nonetheless, the crucial question from the viewpoint of survey methodology is the concern about the representativeness of these data sources. The answer is often challenging and the measurement of representativeness is not widely discussed in the

literature. This paper aims to bridge this gap by proposing an approach to measure the representativeness of new data sources, in particular Internet data sources.

2. Internet data sources

The paper focuses on the Internet as a new data source for statistics, in particular real estate market statistics. For the sake of clarity, the term *Internet data source* (IDS) is defined below.

Definition 1. 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.

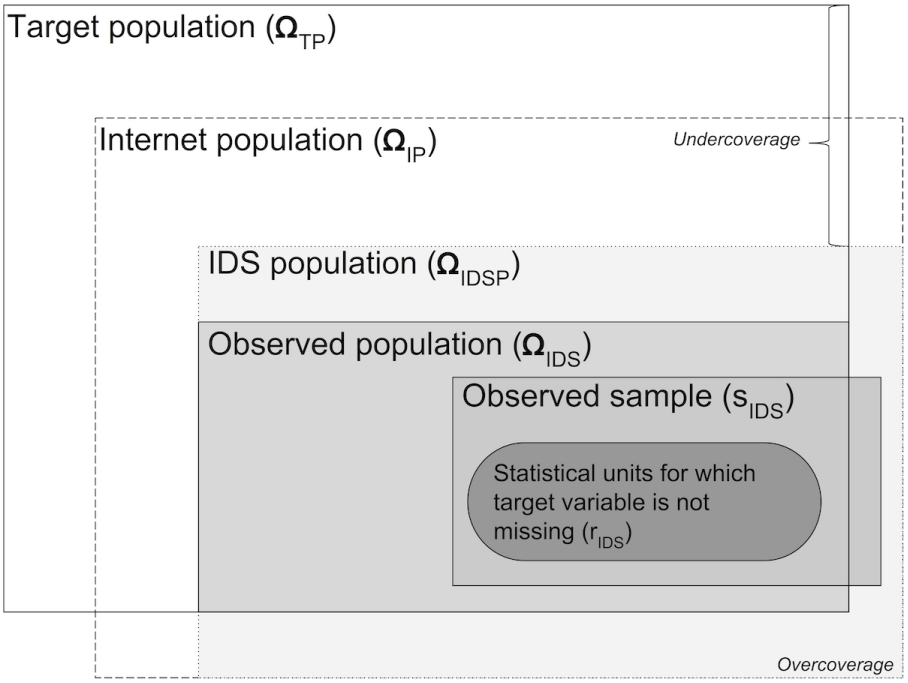

Fig 1 The relation between the target and IDS population

The definition emphasises a number of aspects. First, despite its volume, an IDS should be treated as a sample. The reason for that is an IDS does not contain all units from the target population. Figure 1 presents the relation between the IDS and the target population.

Secondly, unlike official statistics, which are based on probability selection mechanisms, IDSs are the result of the self-selection process: the decision whether to provide information to an IDS is left to individuals/entities, which reflects their non-probabilistic character. The definition explicitly states that data are not collected by statistical institutions or public agencies but by private/commercial entities. The definition specifies that an IDS only refers to data created by Internet users or by private entities themselves. Finally, IDSs are created via the Internet without the support of on-line questionnaires used in on-line surveys. In fact, IDSs are a new type of Internet surveys, where data are collected directly from a given on-line service. This type of survey could be called an IDS survey or an IDS-based survey. IDS data are the result of interactions of individuals and enterprises with the Internet. Therefore, before these data can be used for statistics, their quality assessment should be done to determine bias and its sources.

3. The concept of representativeness

Representativeness is a concept widely discussed in survey methodology, particularly in official statistics. There is, however, no straightforward definition of representativeness, which was already noted by Kruskal and Mosteller (1979a,b,c); the authors provide a list of denotations of representativeness that can be found in statistical and non-statistical literature of that time: (1) general, unjustified acclaim for the data, (2) absence of selective forces, (3) mirror or miniature of the population, (4) typical or ideal case(s), (5) coverage of the population, (6) a vague term to be made precise, (7) representative sampling as a specific sampling method, (8) representative sampling as permitting good estimation, (9) representative sampling as good enough for a particular purpose.

It should be underlined that the concept of representativeness is still valid, even in the era of big data. Hence, in the case of IDSs the following definitions should be considered and applied: the self-selection mechanism, and its results (1) coverage of the target population, (2) comparison of sample and population distributions and finally (3) impact on estimation of finite population characteristics. Figure 2 presents a possible self-selection mechanism in the real estate market.

In order to verify the self-selection mechanism, it is possible to use propensity score (Rosenbaum and Rubin, 1983). In addition, missing data patterns, discussed by Rubin (1976), should be considered: (1) missing completely at random (MCAR), (2) missing at random (MAR) and (3) not missing at random (NMAR).

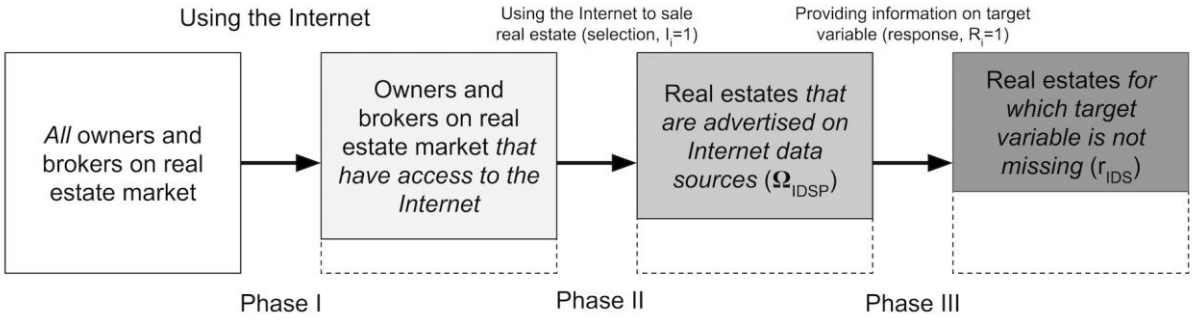

Fig 2 The self-selection mechanism underlying Internet data sources about the secondary real estate market

Formally, in terms of response probability (response propensity, propensity score) MAR can be defined as:

Definition 2. Response probability under the Missing at Random Mechanism is the expectation of the response indicator variable conditional on auxiliary variables, but not on the target variable itself.

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

Definition 3. Response probability under the Not Missing at Random Mechanism is the expectation of the response indicator variable conditional on auxiliary variables and the target variable itself.

Therefore, what needs to be determined is whether the self-selection mechanism for units observed online is associated with the MAR or MNAR pattern.

4. 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. In Poland there are several data sources about the real estate market i.e. the Land and Property register, the Mortgage Register or the Register of Real Estate Prices and Values. Since the study focuses on properties that are on sale, the obvious data source to use is the register of transactions. In Poland, each sold property is registered in the Register of Real Estate Prices and Values, which is maintained by local authorities at LAU1 level. It represents another population, which is related to the general population – the register population.

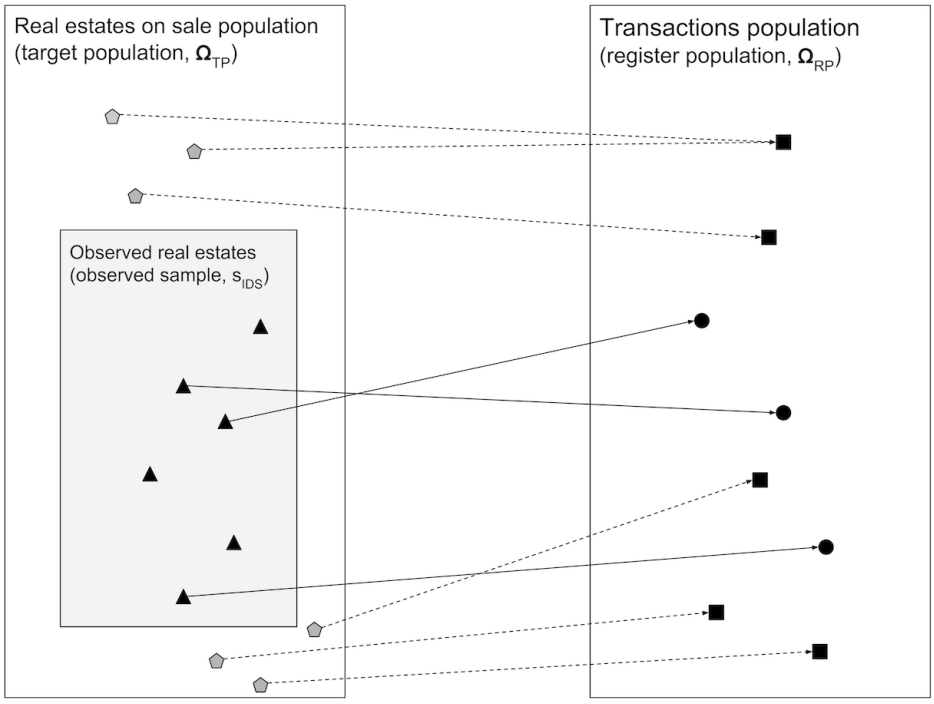

Fig 3 The relation between the register and the IDS population

Figure 3 shows the IDS population on the left and the register population on the right, with lines indicating transitions between these populations. The observed sample s_{IDSs} is obtained from IDSs and includes sale offers for units that will not be sold (not observed in Ω_{RP}).

Transactions denoted by black circles represent sold properties that were members of the IDS population, while black rectangles refer to those that were not observed online.

There may be different linkage scenarios for s_{IDSs} and Ω_{RP}. For one thing, the period between the placement of offers and transactions is often unknown. More importantly, though, in both data sources the same information should exist in order for them to be linked. However, it is desirable to use data that are already available. Linkage with the register population can provide insights into differences between properties offered online (links) and offline (non-links). All in all, identification of different populations in the real estate market will be crucial for detecting potential data sources. Linkage scenarios for these populations can provide information about selection mechanisms.

However, due to the lack of common identifiers between units observed in IDSs and registers it is not possible to link units by means of a deterministic method. Instead, probabilistic methods should be applied to determine the selection mechanism. The number of potential linking variables varies between data sources. An important variable that can verify the correctness of the linking procedure is the property location; however, it was not available in the IDS. Therefore, 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) and the transaction price (from the register). To make the linkage procedure more stable, two blocking variables were created based on floor area (10 groups) and number of rooms (10 groups). Within these blocks linked properties varied in terms of the relation between the offer and transaction price and floor area. The threshold of probability that two records refer to the same (or similar) unit was set at 60%.

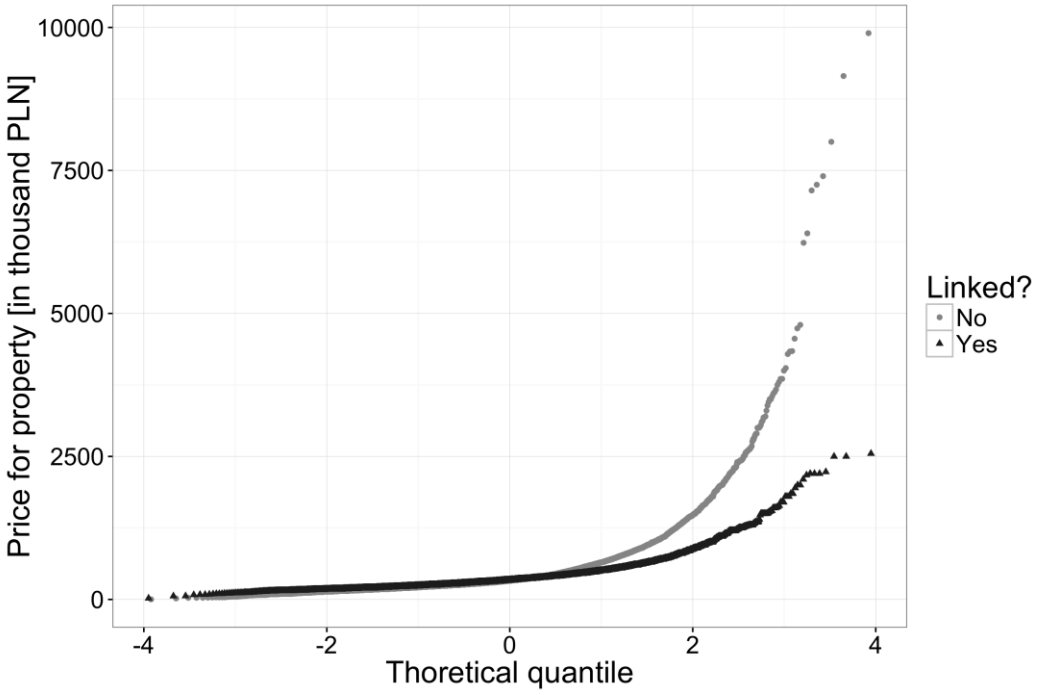

Fig 4 A comparison of the distributions of property prices for linked and non-lined units in the secondary real estate market in Warsaw between 2012 and 2014.

Figure 4 shows a of the distributions of property prices for linked and non-lined units from the register of transactions for Warsaw. Substantial discrepancies between distributions are visible in the right tail, which increase as the price of property grows. Figure 5 presents a comparison of price distributions for Poznań. A similar pattern in the right tail can be observed as in Warsaw, but the differences between linked and not linked units are smaller. Moreover, there is a small difference in the left tail, where properties offered online are more expensive than those online. Differences between Poznań and Warsaw can be explained by market characteristics. Properties in Warsaw are significantly more expensive than in other Polish cities. For instance, the maximum value of a sold property was close to 10 mln PLN, while in Poznań was over 3 mln PLN.

Although the plots indicate differences in distributions, in most cases we are interested in measuring the relation between statistics for properties that were linked (observed online) and not linked (not observed online).

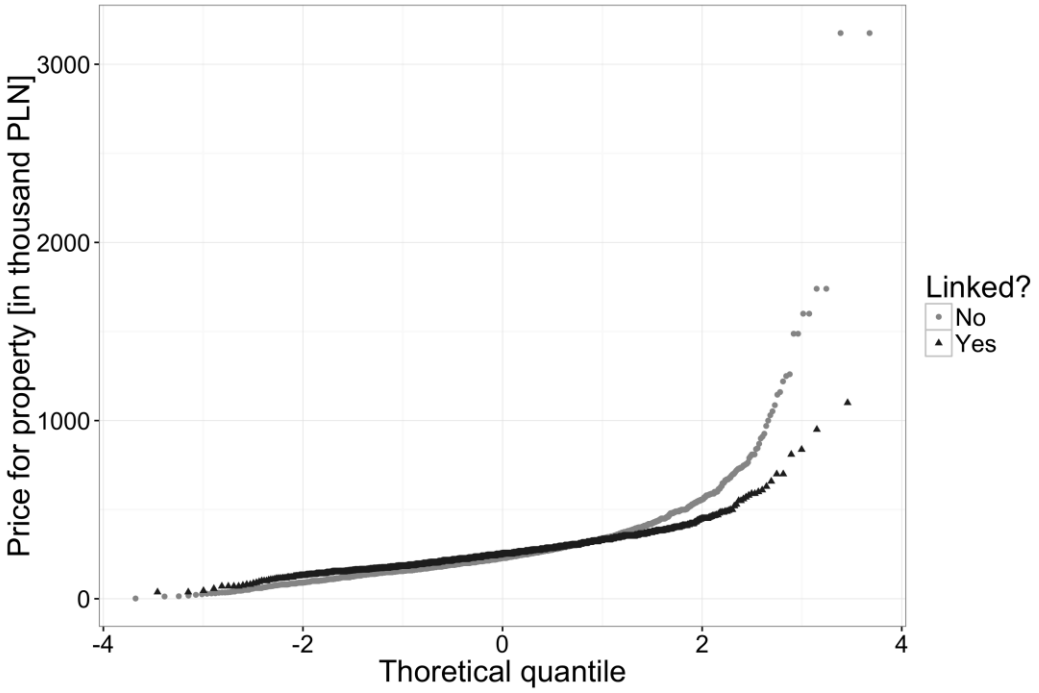

Fig 5 A comparison of the distributions of property prices for linked and non-linked units in the secondary real estate market in Poznań between 2012 and 2014.

Table 1 presents point estimates of mean and median price and the average price for m2. The biggest difference between the mean price of properties offered online and offline can be observed in Warsaw. However, this is due to many outliers that are present in the market, because the differences between median prices are lower and close to 10.000 PLN. In Poznań there is an opposite relation: properties offered online are more expensive than those offered offline.

Table 1. Descriptive statistics for Warsaw and Poznań

City	Linked	Mean Price [PLN]	Median Price [PLN]	Average Price m2 [PLN / 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*

To verify the differences between these two markets, propensity scores were calcuated based on logistic regression and using covariates that were used in the process of record linkage. Figure 6 presents a comparison of distributions of propensity scores for Poznań and Warsaw. There are substantial differences between these distributions. For Poznań, propensities are right skewed, while in Warsaw they have a slight left skew. Average propensity score for Poznań is 0.30 (sd = 0.16) and for Warsaw 0.55 (sd = 0.19). The relation between the distributions indicates that properties from Warsaw are more likely to be presented online (in particular on Nieruchomosci-online.pl) than properties from Poznań. Estimates of propensity scores can be further used to calculate R-indicators and to compare different domains and cities (Schouten et al. 2009).

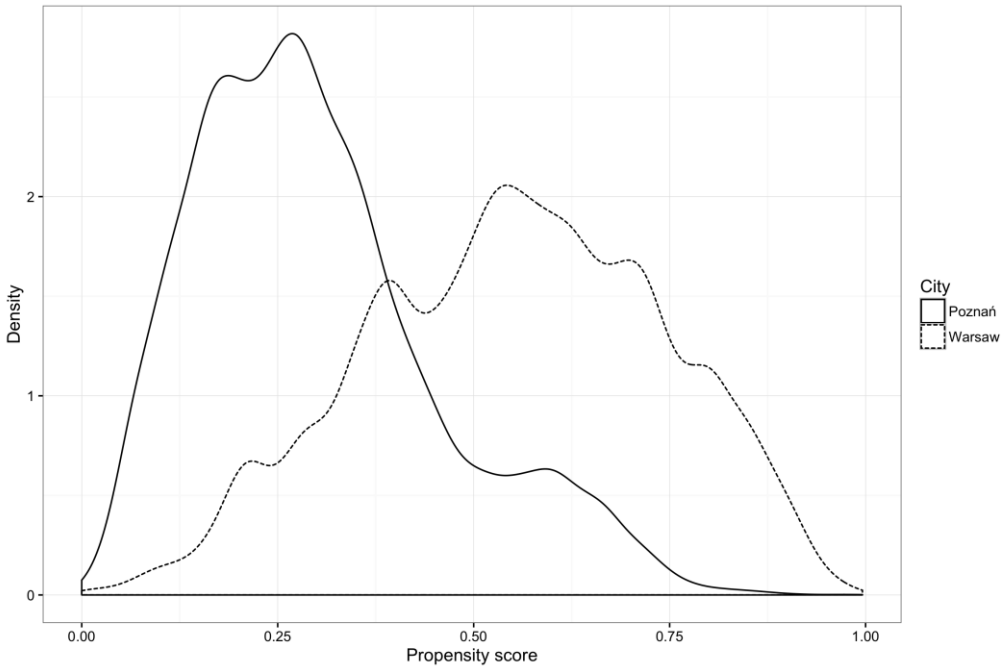

Fig. 6 A comparison of the distributions of propensity scores for Poznań and Warsaw.

5. Conclusion and discussion

This paper addresses key issues concerning new data sources, in particular the Internet. Despite the advantage of their large size, they are still non-probability, 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. Results presented in the paper show differences between properties observed online in Warsaw and Poznań in comparison to the register of transactions. Despite differences in distributions, the bias of the average price for m2 is small in both cases (does not exceed 5%). Further studies focusing on other cities and using other data sources should be conducted to determine the self-selection process operating in real estate markets in Poland.

6. 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., & 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., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. Survey Methodology, 35(1), 101-113