# A two-step procedure to measure representativeness of Internet data sources

Maciej Beręsewcz[1,2]

[1]Department of Statistics, Faculty of Informatics and Electronic Economy, Poznan University of Economics and Business, Al. Niepodleglosci 10, 61-875 Poznan, Poland. E-mail: maciej.beresewicz@ue.poznan.pl

[2]Centre for Small Area Estimation, Statistical Office in Poznan, Wojska Polskiego 27/29, 60-624 Poznan, Poland

**Abstract**

So far statistics has mainly relied on information collected from censuses and sample surveys, which are used to produce statistics about selected characteristics of the population. However, due to cost cuts and increasing non-response in sample surveys statisticians have started to search for new sources of information, such as registers, Internet data sources (IDSs, i.e. web portals) or big data. Administrative sources are already used for purposes of official statistics while the suitability of the latter two sources is currently being discussed in the literature. Unfortunately, only a few papers devoted to statistical theory point out methodological problems related to the use of IDSs, particularly in the context of survey methodology. The unknown generation mechanism and the complexity of such data are often neglected in view of their size. Hence, before IDSs can be used for statistical purposes, especially for official statistics, they need to be assessed in terms of such fundamental issues as representativeness, non-sampling errors or bias. The paper attempts to fill the first gap by proposing a two-step procedure to measure representativeness of IDSs. The procedure will be exemplified using data about the secondary real estate market in Poland.

*Key words:* representativeness, self-selection mechanism, non-probabilistic samples, Internet data sources, big data, real estate market

# 1 Introduction

Growing information needs at a low level of aggregation not only encourage the development of small area estimation but also stimulate the search for new data sources that could support or enhance existing sources (censuses, surveys or reporting). This trend has been continuing since 1970s when statisticians at National Statistical Institutes (NSIs) started using and adopting administrative records as part of their statistical systems (Wallgren & Wallgren, 2014). The use of registers not only significantly reduces survey costs but also usually offers a better coverage of the population and can provide more accurate and timely statistics (Wallgren & Wallgren, 2014). However, the statistical theory underlying the use of administrative registers is currently the subject of research and development (Zhang, 2011, 2012). Nonetheless, the interest in new data sources has brought about a change in thinking about statistical data sources. In the literature and during statistical conferences this process is often described as *a change of paradigm in official statistics*, which involves the adoption of existing data sources instead of creating new ones.

Although administrative records often provide unit-level data, their scope is usually limited to a specific field that is relevant for their administrators. Initially, registers were not created for statistical purposes, which means that they need to be transformed to suit statistical purposes. However, in the environment of electronic economy, characterized by the increasing use of the Internet (by households and companies alike) and the Internet of Things (e.g. mobile technologies, scanner data), surveys as well as administrative registers tend to lag behind the changing setting. Therefore, information gaps in certain fields are growing and new data sources should be examined to improve information coverage.

New data sources, mainly big data, have gained recognition as a potential source of statistical information (Daas et al., 2015; Citro, 2014; Japec et al., 2015; Pfeffermann, 2015; Shlomo & Goldstein, 2015). However, as Citro (2014) states *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 – for example, access to public opinion polls or to local property records.* In the paper we will focus on data sources mentioned by Citro (2014) that we collectively call Internet data sources and a formal definition will be given in the next section. In comparison to big data, these data are in most cases considerably smaller in terms of volume and velocity but contain a variety of data formats (e.g. json, xml, text files, images, natural language). These data are often identifiable, that is contain records that could be meaningfully associated with a unit at a given place and time, such as an individual, institution, product or geographical location (Shlomo & Goldstein, 2015).

Nonetheless, these sources are not created by statisticians or for statistical purposes, which makes them similar to register data. It should be noted that statistical information differs from other types of information; in particular, it is characterized by relevance, accuracy (of estimation), timeliness, accessibility and clarity, comparability (in time and space) and coherence. Moreover, representativeness and non-sampling errors are key factors that can introduce bias. Internet data sources are still not recognized and their suitability as statistical sources is often unknown. Therefore, there is a gap in statistical science, particularly in survey methodology.

## 2 Internet data sources

Internet data sources are present in the literature devoted to information systems and technology, e-commerce, sociology or political studies (cf. Ginsberg et al., 2008; Miller, 2011; Lazer et al., 2014; Couper, 2013). Unfortunately, only a few papers devoted to statistical theory point out methodological problems related to the use of IDSs as sources of statistical information, particularly in the context of estimation theory. Recently, most studies of IDSs (and big data) have come from statistical offices, mainly Statistics Netherlands (Hoekstra et al., 2012; Daas et al., 2012; Buelens et al., 2014; Eurostat, 2016). Japec et al. (2015) provides a review of the current approach to big data for public opinion research and Eurostat (2014) published a report on using the Internet for the collection of information on society and other statistics. Papers on this topic refer to various aspects of the use of new data sources and point to the possibility of extending the existing sources and the scope of research by replacing current methods of data collection and creating new statistics. Examples include the study of prices of products and services, primary and secondary property market research, job vacancies or the use of social media (Facebook, Twitter) and the possibility of constructing sentiment indices. It should also be noted that the use of Internet data sources may pose a potential competition for official statistics, i.e. *The Billion Price Project* (Cavallo, 2012, 2013).

However, to clarify the following discussion some basic definitions concerning the Internet and World Wide Web will be presented. According to the Oxford Dictionary and the glossary of Poland's Central Statistical Office, the *Internet* is a global and public system of interconnected computer networks that use the standardized Internet Protocol Suite (TCP/IP). It is a network of networks consisting of millions of local networks and individual computers from all other the world. E-mail, www, FTP and other services are available to use via the Internet. The *World Wide Web* (WWW/the Web) is part of the Internet; it is an information system on the Internet, which allows documents to be connected to other documents by hypertext links, enabling the user to search for information by moving from one document to another. To access the Web, users rely mainly on *web browsers*, which are programs used to navigate the Web by connecting to a web server, allowing the user to locate, access, and display hypertext documents (one particular example is a mobile browser that can be accessed on mobile devices, such as smartphones or tablets) or *mobile applications* that are computer programs designed to run on mobile devices, such as smartphones and tablets (mobile applications must first be downloaded and installed on mobile devices). Lately *the Internet of Things* (IoT) has gained recognition as a potential source of information. IoT is defined as the interconnection via the Internet of computing devices embedded in everyday objects, enabling them to send and receive data (independently of the user). Finally, a *web service* is a service offered by an electronic device to another electronic device, communicating with each other via the Web. In practice, a web service typically provides an object-oriented web-based interface to a database server, utilized, for example, by another web server, or by a mobile application, which provides a user interface to the end user.

In addition, Bethlehem & Biffignandi (2011) provide the following definitions of surveys that may be taken into account:

**Internet survey** A general term for various forms of data collection via the Internet. Examples are a web survey and an e-mail survey. Also included are forms of data collection that use the Internet just to transport questionnaires and collected data.

**Web survey** A form of data collection via the Internet, in which respondents complete questionnaires on the World Wide Web. The questionnaire is accessed by means of a link to a web page.

**Self-selection survey** A survey for which the sample has been recruited by means of self-selection. Users can decide whether or not to participate in the survey.

In the light of the above, the following definition of an IDS is proposed, which extends the definition proposed earlier in Beręsewicz (2015):

**Definition 2.1.** 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.

The definition emphasizes a number of aspects. First, despite its volume, an IDS should be treated as a *sample*. This is because an IDS does not contain all units from 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 the *non-probabilistic* character of IDSs. 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; official databases, such as Eurostat database, OECD statistics, Statistics Finland StatFin or the Polish Local Data Bank are excluded from the IDS category.

Finally, *IDS are created as a result of a new type of Internet surveys*, where data are collected directly from a given online service. This type of survey could be termed an *IDS survey* or an *IDS-based survey* (which resembles the term *register survey* proposed by Wallgren & Wallgren 2014). Recently, Diaz et al. (2016) described social media as an *imperfect continuous panel survey*, a description, which is also valid for an IDS since units are observed over time. IDSs are the result of interactions of individuals and enterprises with the Internet.

IDSs data can be obtained by web-scraping, specific web services via Application Programming Interface, API (e.g. representational state transfer, REST; Simple Object Access Protocol, SOAP) or directly from the owner. The access level is set by the data source holder and regulated by terms and conditions of use specified by a given portal. The possibility of automatic data collection should be discussed with the data holder and depends on the level of cooperation with statistical agencies. An IDS rarely consists of statistical units – data tend to be objects (non-statistical units), actions or aggregated data that should be transformed. These data sources might also consist of processed data published in reports or statistics (see Google Trends). In this situation, a detailed methodological framework is required to specify when and how representativeness of new data sources can be measured.

# 3 The concept of representativeness of Internet data sources

The starting point for measuring representativeness of IDSs is the explanation of the concept itself. There is, however, no straightforward definition of representativeness, which was already noted by Kruskal & Mosteller (1979a,b,c), who provides a list of denotations used 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.

However, for the purpose of clarity the following denotations will not be studied in detail: general acclaim for the data, typical or ideal case(s), a vague term to be made precise and representative sampling good enough for a particular purpose. The motivation is that these definitions do not directly refer to the representative sample in the context of survey methodology, but are rather general statements about how representativeness is perceived. Given the above concepts and definitions, let us discuss the possibility of applying them to IDSs.

### 3.1 Absence of selective forces

Recently, Schouten et al. (2009) has proposed two definitions of representativeness with respect to survey response – strong (given in Definition 3.1) and weak (given in Definition 3.2). This definition can be associated with the second denotation: *absence of selective forces*. The proposed approach assumes that a probabilistic sample has been drawn, but the final outcome is ultimately determined by the selection mechanism. Information about sampled units and their participation is required in order to assess the representativeness of the response.

**Definition 3.1.** (*strong*) A response subset is representative with respect to the sample if response propensities $\rho_i$ are the same for all units in the population:

$$\forall_i \; E(R_i) = \rho_i = P(R_i = 1 \mid I_i = 1) = \rho \tag{1}$$

and if the response of a unit is independent of the response of all other units (Schouten et al., 2009),

where $R_i$ denotes the response of unit $i$ and $I_i$ is an indicator showing whether a unit took part in the survey. Schouten et al. (2009) notes that strong representativeness corresponds to the MCAR pattern for every target variable $y$. It means that non-response does not cause estimators to be biased. Although the definition is appealing, its validity can never be tested in practice. To solve this problem a weaker definition of representativeness was introduced by Schouten et al. (2009).

**Definition 3.2.** (*weak*) A response subset is representative of categorical variable $\boldsymbol{x}$ with $H$ categories if the average response propensity over the categories is constant

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} \rho_{ih} = \rho, \text{for } h = 1, 2, ..., H, \tag{2}$$

where $N_h$ is the population size of category $h$, $\rho_{ih}$ is the response propensity of unit $i$ in class $h$ and summation is over all units in this category (Schouten et al., 2009).

Weak representativeness means that it is not possible to distinguish respondents from non-respondents just using information with respect to $\boldsymbol{x}$. Unlike in the strong definition, response propensities can be estimated within corresponding strata (based on $\boldsymbol{x}$) and therefore the assumption of weak representativeness can be checked in practice. This definition could be also identified the denotation *representative sampling as permitting good estimation.* Schouten et al. (2009) also introduced R-indicators for the evaluation of a representative response based on measures of variability in response propensities. In order to obtain these propensities the following issues should be considered:

- information about other data sources may be required,
- strong auxiliary variables $\boldsymbol{x}$ that explain $y$ should be available,
- other variables, e.g. from paradata ($\boldsymbol{v}$), that explain the selection and response mechanism should be available,
- the relationship between the selection / response mechanism and the target variable $y$ should be checked,
- population totals or means for auxiliary variables (or proxies, as shown in Reilly et al. 2001) should be known.

The above definition indicates that all units of the population are selected via probability sampling and no self-selection mechanism influences the probability of inclusion in the sample. In the case of IDSs, we do observe several self-selection mechanisms. Units of the target population denoted by $\Omega$ (independently) decide whether to use the Internet, register on a certain web portal and provide information on the target variable. The self-selection mechanism can be associated with observed ($\boldsymbol{x}$), non-observed ($\boldsymbol{z}$) or target variables ($y$). Hence, understanding representativeness in the context of *absence of selective forces* is the most suitable with respect to IDSs.

## 3.2 Mirror or miniature of the population

According to another denotation, a sample is viewed as a *miniature of the population.* This concept was formalized by Bethlehem (2009), who proposed two definitions of a representative survey sample.

**Definition 3.3.** A survey data set is defined to be *representative with respect to variable(s)* $\boldsymbol{x}$ if the distribution of $\boldsymbol{x}$ in the data set is equal to the distribution of this variable in the population ($\Omega$).

$$f_s(\boldsymbol{x}, I_i = 1) = f_\Omega(\boldsymbol{x}), \tag{3}$$

where $f$ denotes the probability density function (PDF) and $I_i$ is defined as previously. A vector of auxiliary variables $\boldsymbol{x}$ refers to characteristics of the population, particularly demographic variables, such as sex, age, education or marital status. In the case of the real estate market, $\boldsymbol{x}$ may refer to property type, floor area or the number of rooms.

The second definition formulated by Bethlehem (2009) refers to a weighted sample adjusted to the known population marginal distribution of $\boldsymbol{x}$.

**Definition 3.4.** A survey data set is defined to be *representative with respect to auxiliary variables* $\boldsymbol{x}$ if the distribution of $\boldsymbol{x}$ is adjusted to the known distribution marginal distribution of $\boldsymbol{x}$ in population $\Omega$.

$$f_s(\boldsymbol{x}|w_i, I_i = 1) = f_\Omega(\boldsymbol{x}), \tag{4}$$

where $w_i$ denotes adjusted weights for each unit $i$. It could be either the inverse of the probability of inclusion in the sample $w_i = d_i = 1/\pi_i$ or corrected weights (e.g. through calibration, post-stratification $w_i = \lambda_i d_i$). In other words, a sample should have the same characteristics as the target population. Nonetheless, Bethlehem's definitions refer to the case when a weighting scheme must be applied in order to make sample and population distributions equal under probabilistic sample selection.

In the theory of survey methodology reference distributions are often known in advance from i.e. a sampling frame, which may be a list of addresses or a population or business register. However, in the case of IDSs it may happen that data about the target population are unavailable, which makes them similar to i.e. households surveys. For example, Facebook or Twitter data may be used to identify the target population as people living in a given country. Therefore, we may have reference marginal distributions that come from this population. On the other hand, we may be interested in studying a population for which we have limited or no information about the distribution of auxiliary variables. This is precisely the case with the secondary real estate market: while information about sold properties is available, there are only limited, survey-based, data about apartments put up for sale.

Moreover, sample characteristics observed in IDSs are often unknown a priori and no aggregate data are available for assessment. In fact, these data sources are sometimes 'living organisms', which means that the distribution of $\boldsymbol{x}$ is likely to vary over time (cf. Diaz et al., 2016). For example, Google Trends provide information about the popularity of certain keywords in time and space but do not provide demographic background information on those who made the searches. Facebook, in contrast, publishes demographic information about users and provides limited access to these data via APIs. IDSs about the real estate market in Poland publish only a limited number of characteristics in their reports, including the mean price per square meter classified by the flat size or the number of rooms, the fraction of flats by type and categorized according to floor area or the number of rooms. The scope of data provided by different IDS owners may vary and may not be consistent with available official statistics (different categories, definitions). Furthermore, information provided on websites may contain errors, which is why access to individual data is crucial for assessing whether distributions are consistent with population data. In the case of real estate market, this access is often possible via manual collection, web-scrapping or an API.

## 3.3 Coverage of the population

Another denotation of representativeness mentioned by Kruskal & Mosteller (1979a,b,c) is *coverage of the target population*. Administrative sources, i.e. population registers, like sampling frames, are considered to fully cover the target population, mainly owing to their obligatory character. However, there is evidence suggesting this may not always be the case (Zhang, 2012; Coleman, 2013; Baffour et al., 2013). For instance, Zhang (2015) discusses the modelling of coverage errors in administrative records, Blackwell et al. (2015) presents patterns of under- and over-coverage in the National Health Service Patient Register. IDSs may also not fully cover the target population.

Be that as it may, we should consider two categories of under-coverage. First, the Internet coverage and its usage play a crucial role in assessing the suitability of IDSs for statistics. Internet coverage depends mainly on location while its usage is highly correlated with age. For example, the Internet usage is high in big cities and among young people while it is low in small towns and among older people. Continuing the example of the real estate market, the use of the Internet may vary depending on market participants: sellers – brokers, owners and buyers – legal and natural persons.

Secondly, IDSs vary in their coverage of the target population and each data source should be investigated separately. We can suspect that its level is related to the popularity of a given IDS. With respect to the real estate market example, in Poland local government agencies are obligated to publish information on a government website, while owners or brokers are free to make their own decisions. Brokers often use customer relationship management (CRM) applications to place ads on more than one IDS. These systems automatize this procedure through multiple listing services (MLS).

We should also consider the problem of over-coverage which arises from the presence in the frame of units not belonging to the target population and of units belonging to the target population that appear in the frame more than once. We can suspect that classified services consist of ads listed multiple times (e.g. due to MLS on real estate) while data holders may control classification of these ads into correct groups (e.g. properties for sale or to rent) or may remove of objects that do not meet criteria (e.g. ads referring to non-existing properties). However, this error may be higher in case of social media as accounts may be associated with persons, companies or bots and this relation may be many-to-many. It could be a serious problem if over-coverage is a result of self-selection. Nonetheless, this error may be corrected to some level when individual data are available.

The coverage of a given IDS is related to both the Internet usage and self-selection from the Internet population of market participants. Bethlehem (2010) show that self-selection can introduce substantial bias in comparison to the bias due to the coverage error. Each IDS should be treated as an *imperfect frame* and IDSs as *imperfect overlapping frames*.

## 3.4 Representative sampling as a specific sampling method

Representativeness is also understood as *a specific sampling method*, particularly when it is based on probabilistic assumptions. According to sample survey theory, valid inferences about the target population can only be made if a probabilistic sample is selected (all units have a known non-zero probability of inclusion $\pi_i > 0$). However, probabilistic sample surveys tend to have non-sampling errors (e.g. non-response) and the final sample consists of units that have decided to participate. For instance, in 2014 non-response in the Polish Household Budget Survey was over 54%, in the Polish EU-SILC nearly 26% and in the Labour Force Survey over 30%.

In the case of IDSs, certain units may have $\pi_i = 0$ and, therefore, may not be included in the sample. The problem can be substantial when units with $\pi_i = 0$ differ from units for which $\pi_i > 0$. For this reason these units may never be present in the IDS sample. However, units with $\pi_i = 0$ should be thoroughly investigated. In the case of the real estate market, there is a lack of research on this matter. Moreover, given the absence of population frames, inclusion probabilities $\pi_i$ are unknown in advance. One possible way to obtain $\pi_i$ is to reweight the sample to known population totals following the calibration approach to obtain weights $w_i$. A similar solution is used to obtain register-based statistics when register totals are not equal to population totals. It should be noted that these weights will be constant in subgroups, so it is assumed that each unit has the same probability of inclusion. Another way of obtaining weights for units in an IDS is to apply indirect sampling (Lavallée, 2007) when we sample a population related to the target population, for instance real estate brokers.

## 3.5 Representative sampling as permitting good estimation

Pfeffermann (2011) discusses complex surveys and the Missing Not at Random (MNAR) mechanism in the context of the modelling process. Pfeffermann (2011) does not invoke the term *representativeness* directly, but emphasizes that under informative non-response the conditional distribution of target variable $y$ in the sample and the population is not equal. Hence, after rewriting the concepts provided by Pfeffermann (2011), a *representative model* can be defined as:

**Definition 3.5.** A model is *representative* only when the conditional distribution of $y$ given $\boldsymbol{x}$ is equal in the sample and in the population. That is, $f_s(y_i|\boldsymbol{x}_i) = f_\Omega(y_i|\boldsymbol{x}_i)$ only if

$$Pr(R_i = 1 \mid \boldsymbol{x}_i, y_i, I_i = 1) = Pr(R_i = 1 \mid \boldsymbol{x}_i, I_i = 1), \tag{5}$$

because the conditional distribution of $y$ given $\boldsymbol{x}$ can be expressed as:

$$f_s(y_i|\boldsymbol{x}_i) = f(y_i|\boldsymbol{x}_i, I_i = 1, R_i = 1) = \frac{Pr(R_i = 1 \mid \boldsymbol{x}_i, y_i, I_i = 1) f_\Omega(y_i|\boldsymbol{x}_i)}{Pr(R_i = 1|\boldsymbol{x}_i, I_i = 1)}. \tag{6}$$

where $f$ denotes the probability density function (PDF), $f_s(y_i|\boldsymbol{x}_i)$ refers to sample conditional PDF, $f_\Omega(y_i|\boldsymbol{x}_i)$ and $I_i, R_i$ are defined as previously. In view of the above, the definition based on Pfeffermann (2011) can be matched to *representative sampling as permitting good estimation* but also to *absence of selective forces*. Pfeffermann (2011) mention the use of response propensity $\rho$, which can account for the self-selection mechanism.

To sum up, self-selection mechanism observed in IDSs results in an under-coverage of the target population, discrepancies in the distribution of auxiliary variables, and finally, may substantially affect estimation. Identifying the this mechanism is a crucial part for the measurement of representativeness of Internet data sources.

# 4 A proposed two-step procedure to measure representativeness of Internet data sources

In the light of the above I propose a procedure to measure representativeness consisting of two steps. Figure 1 illustrates the general idea of the procedure. It starts with a more general question: *Step I: Is the Internet useful for providing statistics?*. This question refers to the overall suitability of the Internet as a data source for statistics. The first step can lead on to Step II or the procedure ends with *END: Search for other source than the Internet.*

`Figure 1 around here.`

The second step *Step II: Internet data source(s)* focuses on a given data source and its representativeness with respect to existing statistical and non-statistical data sources. It starts with one IDS (or multiple integrated IDSs). The second step ends with the measurement of representativeness of IDSs in box *END: Measure representativeness* or with a suggestion regarding an existing data source *END: conduct/modify a survey or search for admin data source*. This alternative outcome suggests that in order to assess representativeness: (1) an additional survey should be conducted, (2) additional questions concerning IDSs should be added to existing sources or (3) other administrative sources should be found (e.g. it is not available at a given time). The second step of the procedure was partially inspired by the work done by Buelens et al. (2014). Buelens et al. (2014) provide a flow diagram to assess the selectivity of big data sources, which was adapted and modified to take into account several aspects that were not originally included.

Figure 1 consists of different shapes to distinguish the steps of the procedure. The squares denote the start and the three possible outcomes of the measurement process. The diamonds denote questions and provide directions to further steps, the letters A and B refer to different levels of the reference data source and the numbers denote the order of steps. The diamonds marked with the letter A refer to statistical sources (census, sample surveys, reporting, statistical registers) and administrative sources (non-statistical data sources). These sources should be used directly or transformed to be used for statistical purposes (by NSIs). It is assumed that registers contain both statistical units (e.g. persons) and non-statistical units (e.g. transactions, legal units), which can be transformed into statistical units. Diamonds marked with the letter B refer to statistical and non-statistical sources that are only available as aggregate data at the domain level. The solid lines with arrows denote the answer *Yes* and the dashed lines with arrows the answer *No* to questions stated in Figure 1.

The term *reference data* will be used with respect to statistical or non-statistical data sources used for statistics to underline that these sources can be used as a reference for comparisons with IDSs. The term *domain* will be used to describe the available level of reference data. The term *Individual data* refers to object-level (non-statistical unit) or unit-level (statistical) data. For simplicity, the term object-level is used to describe both actions (e.g. transactions) and objects (e.g. advertisement, account). Data processing and cleaning is not included in Figure 1. However, it should be noted that this is a crucial stage of the procedure that affects the derivation of units or estimates based on new data sources (Daas et al., 2015).

The proposed procedure takes into account that the Internet may not be suitable or necessary to derive statistics (answer *no* to the question *Is the Internet useful for providing statistics?*). The proposed diagram also takes into account the fact that representativeness can be measured with respect to different populations. For instance, IDSs for the real estate market may contain information on properties as well as brokers. Thus, representativeness may be measured with respect to brokers (what fraction of brokers can be observed in IDSs), properties (what fraction of properties is offered online) or both (what fraction of properties is offered by brokers online).

## 4.1 Step I: Usefulness of the Internet for statistics

The first question *Is the Internet useful for providing statistics?* is used to determine whether the Internet can be used for statistics, particularly for official statistics. To answer this question, the following topics should be examined:

- What problems are associated with existing statistical and non-statistical data sources?
- What kind of information do users of official statistics need and is it possible to obtain this information online?
- What is the Internet coverage and usage in the target population? Is the Internet a relevant source of information about the target population?
- For what purposes is the Internet used and is it important for the target population?

First, one should identify problems with existing data sources. As far as surveys are concerned, the key issues include increasing unit non-response, panel attrition and high respondent burden (Groves, 2006; Brick, 2013). These problems should be further studied to find out whether the use of available online data can improve surveys. Another problem is whether the Internet can be used to obtain similar or new information. For instance, real estate market statistics in Poland are limited to quarterly and annual information with a very narrow scope. Advertisement services may be used to provide more timely and detailed information, but the quality of these data is unknown a priori.

The most essential parts of the first step are the last two questions. Information about the Internet coverage and usage is crucial for determining whether IDSs can be considered a relevant data source. These questions may limit further work on the assessment of IDSs. For instance, data from countries with a low or moderate Internet coverage or those where only specific units use the Internet are likely to be unsuitable for this approach.

In order to answer the question stated in the first step of the procedure data sources should be identified. Most NSIs conduct surveys or maintain administrative data sources which can be analysed to answer the questions. However, it should also be remembered that IDSs may contain data that are not yet available in official statistics. In such situations, data sources with proxy variables should be considered.

### 4.1.1 Data sources that can be used in the first step

NSIs of EU countries are obligated to conduct the Information and Communication Technologies (ICT) survey. The survey is standardized and coordinated by Eurostat to produce coherent EU statistics. Results of the survey are published as part of statistics on Information Society and cover e-business, postal services, ICT in enterprises and households (Eurostat, 2015a). Other NSIs, such as the US Census Bureau, the Australian Bureau of Statistics or Statistics Canada, conduct similar surveys in this field. Because the Central Statistical Office in Poland (CSO) is obliged to conduct the ICT survey as part of the Eurostat ICT survey, the main focus will be placed on this survey.

The ICT survey is conducted annually in all Member States, as well as in two countries of the European Free Trade Association (EFTA), and acceding and candidate countries aspiring to join the EU. The data collection is based on Regulation (EC) 808/2004 of the European Parliament and the Council. The transmission of micro data to Eurostat was voluntary until the reference year 2010 and has been mandatory since 2011 (Eurostat, 2015a). The ICT survey in different countries is slightly different and a detailed description is provided in the Methodological Manual (Eurostat, 2015b).

The ICT survey gathers information on the access to and use of ICT by means of two separate questionnaires: one for enterprises and another for households and its members. The ICT survey covers households with at least one member aged between 16 and 74 and individuals aged between 16 and 74. The survey collects data on the use of information and communication technologies, the Internet (e.g. social networks), e-government and electronic skills (e.g. security, e-banking, ordering or buying goods or services for private use) in households and by individuals. In addition, a number of background variables are collected including household composition, income and location as well as the age, gender, educational attainment and employment situation of persons.

As for enterprises, the target group includes companies employing at least 10 persons. The activity coverage is restricted to those enterprises whose principal activity is within NACE Rev. 2 Sections C through N, excluding Section K and Division 75 but including Group 95.1 (see `http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Community_survey_on_ICT_usage_in_enterprises`). The population of enterprises includes section L related to real estate activities, which consist of activities involving the operation of own or leased property and conducted on a fee or contract basis. The focus is on data about the use of information and communication technology (e.g. mobile connection, cloud computing), the Internet (e.g. usage of social media), e-government, e-business (e.g. using customer relationship management applications) and e-commerce (e-sales) in enterprises. The key variables on characteristics of enterprises include total turnover, main economic activity or the average number of employees.

Finally, analysis of selected data sources should provide the basis for answering the question asked in the first step – *Is the Internet useful for providing statistics?*. If the answer to the question is *yes*, then the procedure moves on to the second step, which focuses on particular data source(s). This step assumes that access to IDSs is possible and data from these sources are available. If the answer to the question is *no*, then the procedure ends with the suggestion that sources other than the Internet should be considered.

## 4.2 Step II: Representativeness of Internet data sources

The main question to be answered in the second step is: *are specific IDSs representative of the target population?*. In addition, this is related to a practical question: *given the available data, how to measure the representativeness of IDSs?* The diagram is separated into two parts that are distinguished by the question: *are individual IDS data available?*. The diamonds with the letter A in Figure 1 refer to individual data that are made available for the statistician, while the diamonds with the letter B represent aggregate data that are available or are shared with the statistician.

### 4.2.1 Case 1: only domain-level data are available

The case when individual data are not available is selected when the answer to the question *are individual IDS data available?* is *no*. The answer leads to question B1, which verifies whether reference data are available at the domain level (*B1: Are reference domain-level data available?*). To answer question B1 the following conditions should be taken into account. Data can be made available to the statistician (1) at the level of aggregation predefined by the statistician/NSI; (2) at the level defined by the data source owner.

The first case refers to the situation when the statistician/NSI collaborates with the data owner. Access to individual data is not possible due to privacy or practical aspects (e.g. volume of the data). For instance, the author of this article established collaboration with three data owners operating in the secondary real estate market. Advertisement services provided historical data in the form of predefined aggregates. These levels were more detailed than currently available official statistics. Table 1 presents sample rows and columns from the data file that was made available by one of the data owners.

`Table 1 around here.`

Another example of such aggregates are publicly available indices, for instance accessible through special services. The most known service is Google Trends, which provides information on the popularity of words and terms searched on Google. Google Trends classify queries into groups using algorithms developed by Google and provide indices at levels defined by Google (e.g. provinces in Poland, or classification of searches in Google). Continuing with the example of the real estate market, in Poland several online advertising services provide price statistics and indices based on ads (For instance, see `https://ceny.szybko.pl/ceny-nieruchomosci` or `http://www.morizon.pl/ceny`). It should be noted that domain-level data do not allow the statistician to fully assess data quality. Calculations are made by the data holder, who does not always provide information about data processing or the cleaning process or the methodology applied.

Therefore, the answer *yes* to the question *B1: Are reference domain-level data available?* represents two possible cases. Reference data are available at the same level of aggregation as in the IDS or both data sources should be harmonized before they can be compared. In this case the procedure ends with *END: Measure representativeness*. Otherwise, the question *B1: Are proxy data available?* should be answered. Proxy data can refer to the same population but contain a variable with a different definition or a variable with a similar definition but for a different population. Examples of proxy data for IDSs include transactions in the real estate market or primary market characteristics (when the secondary market is considered). In the case when proxy domain-level data are available, the answer *yes* to question B1 leads to *END: Measure representativeness*.

### 4.2.2 Case 1: the measurement of representativeness

Available measures of representativeness such as R-indicators cannot be applied. Instead, evaluation should depend on IDS and reference data based estimates of the target variable and auxiliary variables. This could include estimation of bias in the target variable or tabular comparisons of the distribution of auxiliary variables. Such an approach is common when register data are verified using sample data (eg. LFS status). Another approach that could be useful was proposed by Zhang et al. (2013). This method included evaluation of bias of binary $y$ under MAR and NMAR assumption when only aggregated data are available. Zhang et al. (2013) approach included the estimation of the ratio of direct and post-stratified sample mean ($\bar{y}$) based on the observed sample alone. When time series are available for both data sources, decomposition to trend and seasonality could be considered. This step would make it possible to investigate whether the structure of time series is similar or a co-integration between IDSs and reference data is present.

The measurement of representativeness at domain level is not straightforward. For instance, if a dataset was prepared by the IDS owner, the statistician/NSI has no control over the data cleaning process or calculations. This lack of control may result in unit or measurement errors. Moreover, it may happen that time series data for different periods should be compared. For instance, NSIs provide information on a monthly, quarterly or annual basis, while IDS-based statistics are available on a daily or intraday basis. This poses another problem of how to aggregate IDS data. There may also be a time shift between official data and IDSs. Another issue is that owing to the character of official data (census, survey, reporting, or statistical registers) their aggregation level may be limited to the country, regional or city level. Such a limit can reduce the variance of estimates and will not capture differences between regions or cities. Due to the character of reference data (differences in concepts or populations) there may be discrepancies between IDSs and official statistics. Direct comparison of estimates is based on the assumption that reference data provide unbiased estimates of a given characteristic of a target variable, while this might not be true for surveys that are subject to non-response. Finally, and most importantly, aggregate data may be insufficient to capture the selection process and can only provide an approximation of this mechanism.

### 4.2.3 Case 2: individual data are available

This section focuses on individual data that are available for IDS and reference data. The question: *Are objects statistical units?* attempts to verify whether IDS contain statistical units (e.g. persons, properties, establishments) or non-statistical units (e.g. advertisements, transactions). The answer to this question opens two possibilities. If the answer is *no* the next question is *Can objects be transformed into statistical units?*. If the answer is *yes*, then the path with the diamonds containing the letter A is selected.

In the case of IDSs, the answer to the question: *Are objects statistical units?* is more likely to be *no*. For instance, classified services contain ads that refer to statistical units or composite units (composition of statistical units); accounts on social media might refer to individual persons, groups of people or bots. The following question: *Can objects be transformed into statistical units?* is intended to find out whether it is possible to match objects to statistical units. This part is the most challenging aspect of IDSs. For one thing, multiple records can refer to the same statistical unit. For example, in the real estate market it is common for one flat to be put up for sale a number of times. Moreover, information provided or collected by statisticians can be limited to protect privacy and is not sufficient to identify statistical units or data holders do not collect or have these data at all (Shlomo & Goldstein, 2015). It can be argued that the following statistical units present in IDSs could be identified quite easily: (1) products (e.g. groceries, electronics) or services, (2) vehicles (e.g. using VIN number), (3) job offers, (4) properties (but it depends on the market and country) or (5) establishments/enterprises. On the other hand, identification based on the following services can prove challenging: social media (e.g. for Facebook or LinkedIn it may be easier than for Twitter) or query data (e.g. who has made a given query).

Transforming objects into statistical units should also be done for registers. For instance, the natural choice in the case of the real estate market in Poland is the Register of Real Estate Prices and Values. The register contains data about transactions in the primary and secondary market. Table 2 contains selected variables from the Register of Transactions. Note that transactions can refer to one or several properties. For instance, one transaction can refer to a property with a garage. The Register of Transactions and other registers about the real estate market in Poland will also be discussed later as a potential source for linkage with IDSs.

Table 2 around here.

It should be taken into account that transforming objects into statistical units may not be possible. In such situation, the measurement of representativeness can be done only at the aggregated level based on non-statistical units. This leads to the answer *no* to the question about the possibility of transformation and the circle *Aggregate data to domain level* in Figure 1. The aggregation process should be followed by determining possible levels of comparison with official statistics and data cleaning in order to derive characteristics of statistical units. Possible approaches to deriving statistical units from non-statistical units are presented in Zhang (2011) and Wallgren & Wallgren (2014, ch. 7). Another interesting approach is profiling which is based on paradata associated with a given object (Flekova & Gurevych, 2013; Daas et al., 2015). After data aggregation, the process continues along the path for domain-level data presented in Section 4.2.1.

The answer *yes* to the question *Can objects be transformed into statistical unit?* leads to the same path as the positive answer to the question *Are objects statistical units?*. In this path three methods of linking IDS with reference data are considered: deterministic record linkage, probabilistic record linkage and statistical matching/data fusion.

**Deterministic record linkage** The question *A1: Can deterministic record linkage be applied?* is designed to determine whether data can be linked using common identifiers. This type of linkage considers the case when IDSs and reference data contain the same units and identifiers are present in both sources. For instance, when a legal or natural person registers business activity in Poland, they receive a REGON identifier created by CSO (ang. *National Official Business Register*). Thus, hypothetically, it should be possible to link units from IDSs and the REGON register using the REGON ID. However, in the case of properties, no common identifiers that can be used for linkage purposes are available in the statistical system. Advertisement services use different IDs that can be used for deterministic linkage.

The positive answer to question A1 leads to the measurement of representativeness. Nonetheless, owing to privacy restrictions or the policy of IDS owners this type of linkage is rarely possible in practice, which results in the more likely answer *no* to question A1 and is followed by the question *A2: Can probabilistic record linkage be applied?*.

**Probabilistic record linkage** Question A2 refers to probabilistic methods for linking records from IDSs and reference data. This type of linkage does not require common identifiers but a set of common variables that are present in both data sources. The theoretical foundations of the method are presented in Fellegi & Sunter (1969) and are still being developed (cf. Harron et al., 2015). The general idea consists of assigning a linkage weight that refers to the probability that two units from dataset $\boldsymbol{A}$ (e.g. an IDS) and dataset $\boldsymbol{B}$ (e.g. a census) refer to the same unit. Probabilistic record linkage is used in official statistics. For instance, when the system of registers is used for statistical purposes or census and survey data are linked (Chambers, 2009). However, the underlying theory for methods based on probabilistically linked data is currently the subject of ongoing research (cf. Samart, 2011).

In the case of the real estate market, probabilistic record linkage can be used as follows: NBP/CSO conduct a survey of brokers, which covers both the primary and secondary market. Brokers report information about properties offered for sale, but the questionnaire does not include a question about which properties are offered online. Assuming that brokers provide full information about their offers it may be possible to link IDS data with (1) brokers to check which brokers use the Internet to publish information about properties, or (2) properties with those reported by brokers.

Moreover, in Poland there are several registers that could be used for purposes of data linkage. There is already the Integrated Cadastral Register (ICR), which consists of the Register of Real Estate Prices and Values, Mortgage Registers and the Land and Buildings register. In addition, the following registers may be considered: Tax registers (e.g. VAT or NIP) or business registers (CEIDG). On $13^{th}$ June 2013 several professions (including brokers) were deregulated (The Real Estate Management Act, 1997). Under the new regulations no professional license is required to conduct brokerage activity. Thus, since 2013 there has been no official register of brokers in Poland. In this situation the statistical register REGON could be used instead. REGON covers: legal persons, organizational units without the status of a legal person, natural persons conducting economic activity. The scope of the REGON register makes it possible to link brokers and real estate agencies that use advertisement services. This can be done by using the company or person's name and address.

Taking into account the above mentioned registers, the Register of Real Estate Prices and Values seems to be the most suitable one. The register records transactions made both in the primary and secondary market as well as those made by local governments at auctions. There are several reasons for choosing the Register of Transactions: (1) classified services devoted to the real estate market contain information about properties put up for sale; (2) the register covers all transactions involving fully-owned properties (properties owned by housing co-operatives are excluded) and linkage may help to identify units that are not observed in IDSs; (3) information in the register can be used to identify objects; (4) data stored in the register are unbiased and free of measurement error (information is taken from the Mortgage Register); and (5) most importantly, due to the new legislation that came into effect on $12^{th}$ of July 2014 researchers form Polish Universities can acquire register data for research and teaching purposes free of charge (The Geodetic and Cartographic Act, 1989, art. 40a par. 2 pt 2). Table 2 presents possible variables that could be used for linkage. To sum up, the Register of Real Estate Prices and Values is the most suitable data source for linkage with IDSs related to the property market.

Finally, if assumptions of probabilistic linkage are too strict, the question *A2: Can probabilistic record linkage be applied?* is answered negatively, which leads to the question *A3: Can statistical matching be applied?*.

**Statistical matching/data fusion** Statistical matching or data fusion (Rubin, 1986; D'Orazio et al., 2006; Rässler, 2012) assumes that there are two (or more) data sources containing statistical units from the same target population. Following the definitions provided by D'Orazio et al. (2006), there are two data sources denoted by $\boldsymbol{A}$ and $\boldsymbol{B}$, which share a set of variables $\boldsymbol{x}$, and variable $\boldsymbol{y}$ is available only in $\boldsymbol{A}$ and variable $\boldsymbol{z}$ is only present in $\boldsymbol{B}$. Variables $\boldsymbol{x}$ are shared by both data sources, while variables $\boldsymbol{y}$ and $\boldsymbol{z}$ are not observed jointly.

Statistical matching (SM) consists in investigating the relationship between $\boldsymbol{z}$ and $\boldsymbol{y}$ at *micro* or *macro* level. At micro level, the aim of SM is to create a *synthetic* data source in which all the variables, $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$, are available (usually $\boldsymbol{A} \cup \boldsymbol{B}$ with all the missing values filled in or simply $\boldsymbol{A}$ filled in with the values of $\boldsymbol{z}$). At macro level, the data sources are integrated to derive an estimate of the parameter of interest, e.g. the correlation coefficient between $\boldsymbol{y}$ and $\boldsymbol{z}$ or the contingency table $\boldsymbol{y} \times \boldsymbol{z}$. SM methods assume *conditional independence* of $\boldsymbol{y}$ and $\boldsymbol{z}$ given $\boldsymbol{x}$, which, as D'Orazio et al. (2006) notes, is strong and seldom holds in practice.

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = f(\boldsymbol{y} \mid \boldsymbol{x}) \times f(\boldsymbol{z} \mid \boldsymbol{y}) \times f(\boldsymbol{x}). \tag{7}$$

As a result of the matching procedure, correlation and contingency tables can be used to compare distributions between $\boldsymbol{y}$ and $\boldsymbol{z}$ when jointly observed. The basic statistical matching/data fusion technique aims at obtaining a complete dataset based on the input datasets.

However, for purposes of measuring representativeness, the idea of using statistical matching should be considered in two situations: (1) create a complete dataset to provide information about the differences between $\boldsymbol{y}$ presented online (IDSs) and $\boldsymbol{z}$ that comes from reference data; (2) link units between sources using non-parametric measures and evaluate linkage quality (e.g. matching by propensity score, (Rässler, 2012, chap. 2.3-2.4), using distance functions in the KNN approach, (D'Orazio et al., 2006, chap. 2-3)). To link units one can consider dissimilarity measures such as Mahalanobis, maximum or Gower (1971) distance. The latter might be considered the most suitable because it takes into account variables of different type and missing data. If the data contain strings (e.g. street names) approximate matching and string distance measures such as Jaro-Winkler or Levenshtein can be considered.

The final Gower's dissimilarity between the $i$-th and $j$-th unit is obtained as a weighted sum of dissimilarities for each variable and is given by the following equation:

$$d(i,j) = \sum_k (\delta_{ijk} \times d_{ijk}) / \sum_k \delta_{ijk}, \tag{8}$$

where $d_{ijk}$ represents the distance between the $i$-th and $j$-th unit computed considering the $k$-th variable and $\delta_{ijk} = 0$ when $x_{ik} = NA$ or the variable is asymmetrically binary and $\delta_{ijk} = 1$ otherwise. The measure depends on the nature of the variable – logical, categorical, ordered or numeric. Then, for each unit from the statistical dataset and IDSs we obtain a matrix of distances and compute the overall distance given by equation (8) from which a minimum value for each units should be taken. This measure will be investigated at last stage of the proposed approach. When the use of deterministic or probabilistic record linkage or statistical matching/data fusion is not possible (negative answers to A1, A2 and A3, then aggregation to domain level is the recommended solution.

### 4.2.4 Case 2: the measurement of representativeness

The final outcome of the proposed procedure to the measurement of representativeness depends on the linkage method applied. If it is possible to link units observed in IDS to a register that covers (almost) entire population using deterministic record linkage further steps are straightforward. It can include a comparison of linked and non-linked units to determine if bias is present. For instance, units that cannot be linked to the Register of Transactions can provide information about the characteristics of properties that are not presented online. This step can also include weighting adjustments, such as calibration, to investigate if it reduces bias. Moreover, parametric and non-parametric models to explain the self-selection mechanism should be considered, where the dependent variable would be defined as 1 = linked, 0 otherwise. Finally, this can be followed by calculation of R-indicators.

A similar approach can be applied for a probabilistic methods of linkage. However, for methods based on an inexact identification of the same units, a correction for linkage errors should be applied. Chambers (2009) showed that regression parameters are substantially biased if this error is neglected. This is a crucial issue at the stage of determining the selection mechanism, which is the basis for propensity weighting or R-indicators.

The statistical matching approach is somewhat different because it does not assume the same units are linked. One solution to measure representativeness is to carefully study the distance between units observed in the IDS and reference data (e.g. surveys) with respect to auxiliary variables and the target (or proxy) variable. Another approach may involve building a model to describe the distance variable to detect characteristics of units with a high level of dissimilarity measures.

However, all cases presented above assume that the reference dataset contains all possible subgroups from the target population and can be used to detect the selection/response mechanism (including MNAR), the coverage of the target population, comparison of distributions and estimation of bias in the target variable.

# 5 Implications for statistics

The previous section dealt with the theoretical basis of Internet data sources, focusing on the concept of representativeness and its measurement. For this purpose, basic notation was introduced and selected definitions based on the literature were presented and discussed in the context of IDSs. Furthermore, the concept of representativeness was examined based on statistical literature with regard to IDSs. Based on the theoretical concepts a two-step representativeness measurement procedure was proposed. The proposed method is a general approach to quality assessment of IDSs. The idea takes into account different data sources (e.g. surveys, register data) and aggregation levels (unit and domain).

The second step of the procedure has two possible outcomes – the measurement of the representativeness of IDSs or the need for new data to assess IDSs. The second case can occur in the absence of reference data for comparison at a given time. A possible solution is to conduct a new survey or extend existing surveys by adding new questions that focus on IDSs. Table 3 contains a summary of advantages and disadvantages of using individual and domain data for the measurement of representativeness.

`Table 3 around here.`

NSIs are aware of the significance of new data sources. For instance, the ICT survey covers the use of social media, ownership of websites in enterprises and households (Central Statistical Office, 2015). The Household Budget Survey (Pol. *Badanie Budżetów Gospodarstw Domowych*, BBGD) collects data from households which record every expense made and should indicate whether a given product was bought via the Internet or in the traditional way (Central Statistical Office, 2014). Taking this into account, the NBP/CSO survey of brokers (National Bank of Poland, 2014b,a) could be extended by adding two questions: (1) is a given property being advertised online?; (2) where are advertisements published (own web page, advertisement service)? The first question would be used to determine the fraction of properties presented online, the second would show the use of online services. In addition, the questionnaire about properties could be extended by adding a column for the ID assigned by the broker.

Another possible solution presented in Figure 1 is to search for new administrative data sources that were not available at a given time. However, in the Polish context, when the Register of Real Estate Prices and Values is already available, there is no need to look for additional information.

There are also different ways of using IDSs for statistics, which were studied by Beręsewicz (2016) and include (1) the use of IDSs as a single source for production of certain statistic of the target variable, (2) the use of IDSs as a source of auxiliary information for model-assisted or model-based estimators either at unit or domain level and (3) the use of IDSs to change the current collection method for certain subpopulations.

To assess the representativeness of new data sources (including IDSs) time series of historical data should be considered. A comparison over time may reveal a similarity (in level or trend) to existing data sources. For instance, Daas et al. (2015, p. 255) conducted a monthly comparison between social media sentiments and Dutch consumer confidence between 2010 and 2012, Cavallo (2013) provided a comparison of online CPI with the official CPI calculated by the Census Bureau and Beręsewicz (2016) investigated the self-selection mechanism in the Polish property market. The lack of high quality reference data, such as registers or samples, limits the possibilities of measuring representativeness of IDSs.

The procedure is not limited to IDSs but could be extended to other types of new data sources. However, owing to the complexity and differences between big data sources with respect to the identification of statistical units and the level of data availability, a thorough profiling study should be undertaken before applying the proposed approach.

# Acknowledgements

This research has been supported by the National Science Center, Preludium 7 grant no. 2014/13/N/HS4/02999. I would like to thank Editor, Associate Editor, and referees for their comments which led to considerable improvements in the paper. In addition, I would like to thank Prof. Elżbieta Gołata and dr Marcin Szymkowiak for their valuable comments on an early draft of the present paper.

# References

Baffour, B., King, T., & Valente, P. (2013). The Modern Census: Evolution, Examples and Evaluation. *Internat. Statist. Rev*, 81, **3**, 407–425.

Beręsewicz, M. E. (2015). On representativeness of internet data sources for real estate market in Poland. *Aust. J. of Stat.*, 44, **2**, 45–57.

Beręsewicz, M. E. (2016). *Internet data sources for real estate market statistics.* PhD Thesis. Available at: `https://berenz.github.io/assets/phd/Beresewicz_Maciej_dissertation.pdf`. Accessed January 2017.

Bethlehem, J. (2009). *Applied survey methods: A statistical perspective.* New York: Wiley.

Bethlehem, J. (2010). Selection Bias in Web Surveys. *Internat. Statist. Rev*, 78, **2**, 161–188.

Bethlehem, J. & Biffignandi, S. (2011). *Handbook of web surveys.* New York: Wiley.

Blackwell, L. and Charlesworth, A. & Rogers, N. J. (2016). Linkage of Census and Administrative Data to Quality Assure the 2011 Census for England and Wale. *J. Official Statist.*, 31, **3**, 453–473.

Brick, J. M. (2013). Unit Nonresponse and Weighting Adjustments : A Critical Review. *J. Official Statist.*, 29, **3**, 329–353.

Buelens, B., Daas, P. J. H., Burger, J., Puts, M., and van den Brakel, J. (2014). Selectivity of Big Data. Available at: `https://www.cbs.nl/nl-nl/achtergrond/2014/14/selectivity-of-big-data`. Accessed January 2016.

Cavallo, A. (2012). Scraped data and sticky prices (Working paper no. 21490). Retrieved from National Bureau of Economic Research website: `http://www.nber.org/papers/w21490`.

Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. *J. Monetary Econ.*, 60, **2**, 152–165.

Chambers, R. (2009). Regression Analysis of Probability-Linked Data. *Official Statistics Research Series*, **4**. Statistics New Zealand.

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

Coleman, D. (2013). The Twilight of the Census. *Popul. Dev. Rev*, 38, **1**, 334–351.

Couper, M. P. (2013). Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Surv. Res. Meth.*, 7, **3**, 145–156.

Central Statistical Office (2014). The household budget survey. Available at: http://stat.gov.pl/en/topics/living-conditions/living-conditions/household-budget-survey-in-2014,2,9.html. Accessed January 2016.

Central Statistical Office (2015). Information society in Poland in 2010-2015. Available at: http://stat.gov.pl/en/topics/science-and-technology/information-society/. Accessed January 2016.

Daas, P., Puts, M., Buelens, B., & Van den Hurk, P. (2015). Big data as a source for official statistics. *J. Official Statist.*, 2, **31**, 249-262

Daas, P., Roos, M., van de Ven, M., & Neroni, J. (2012). Twitter as a potential data source for statistics. Available at: https://www.cbs.nl/nl-nl/achtergrond/2012/41/twitter-as-a-potential-data-source-for-statistics. Accessed January 2016.

Diaz, F., Gamon, M., Hofman, J. M., Kiciman, E., & Rothschild, D. (2016). Online and Social Media Data As an Imperfect Continuous Panel Survey Fernando. *PLoS ONE*, 11, **1**, 1–21.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice.* New York: Wiley.

Eurostat (2014). Analysis of methodologies for using the Internet for the collection of information society and other statistics – Final Report. Available at: http://ec.europa.eu/eurostat/cros/system/files/D0.4_Finaltechnicalreport_20140430_1.pdf. Accessed January 2017.

Eurostat (2015a). Information society. Available at: http://ec.europa.eu/eurostat/web/information-society. Accessed January 2016.

Eurostat (2015b). Methodological manual for statistics on the information society. Available at: http://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-BG-06-004. Accessed January 2016.

Eurostat (2016). ESSnet on Big Data. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page. Accessed January 2017.

Fellegi, I. P. & Sunter, A. B. (1969). A theory for record linkage. *J. Am. Stat. Assoc.*, 64, **328**, 1183–1210.

Flekova, L. & Gurevych, I. (2013). Can we hide in the web? large scale simultaneous age and gender author profiling in social media. Paper for the evaluation lab on uncovering plagiarism, authorship, and social software misuse at Conference and Labs Evaluation Forum 2013, September 23–26, Valencia, Spain.

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**, 1012–1014.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27, **4**, 857-871.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opin. Quart.*, 70, **5**, 646–675.

Harron, K., Goldstein, H., & Dibben, C. (2015). *Methodological Developments in Data Linkage.* New York: Wiley.

Hoekstra, R., ten Bosch, O., and Harteveld, F. (2012). Automated data collection from web sources for official statistics: First experiences. *Stat. J. Int. Assoc. Offcial Stat.*, 28, **3**, 99–111.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., and Usher, A. (2015). Big data in survey research AAPOR task force report. *Public Opin. Quart.*, 79, **4**, 839–880.

Kruskal, W. & Mosteller, F. (1979a). Representative sampling I: Non-scientific literature. *Internat. Statist. Rev*, 47, **1**, 13–24.

Kruskal, W. & Mosteller, F. (1979b). Representative sampling II: Scientific literature excluding statistics. *Internat. Statist. Rev*, 47, **2**, 111–123.

Kruskal, W. & Mosteller, F. (1979c). Representative sampling III: The current statistical literature. *Internat. Statist. Rev*, 47, **3**, 245–265.

Lavallée, P. (2007). *Indirect Sampling.* Springer.

Lazer, D. M., Kennedy, R., King, G., & Vespignani, A. (2014). *The parable of Google Flu: traps in big data analysis. Science*, **343**, 1203–1205.

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

National Bank of Poland (2014a). *The real estate market - Information Quarterly.* Available at: `http://www.nbp.pl/homen.aspx?f=/en/publikacje/inne/real\_estate\_market\_pre.html`. Accessed January 2016.

National Bank of Poland (2014b). *Report on the situation on the markets of residential and commercial property in Poland in 2013.* Available at: `http://www.nbp.pl/homen.aspx?f=/en/publikacje/inne/real\_estate\_market\_pre.html`. Accessed January 2016.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? why is it a problem? how can we approach it. *Surv. Methodol.*, 37, **2**, 115–136.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics. *J. Surv Stat. Methodol.*, 3, **4**, 425–483.

Rässler, S. (2012). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*, volume 168. New York: Springer-Verlag.

Reilly, C., Gelman, A., & Katz, J. (2001). Poststratification Without Population Level Information on the Poststratifying Variable With Application to Political Polling. *J. Am. Stat. Assoc.*, 96, **453**, 1–11.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econ. Stat.*, 4, **1**, 87–94.

Samart, K. (2011). *Analysis of probabilistically linked data.* PhD thesis. Available at `http://ro.uow.edu.au/theses/3513/`. Accessed January 2016.

Shlomo, N., & Goldstein, H. (2015). Editorial: Big data in social research. *J. R. Stat. Soc. Ser. A*, 178, **4**, 787–790.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Surv. Methodol.*, 35, **1**, 101–113.

The Geodetic and Cartographic Act (1989). Available at `http://isap.sejm.gov.pl/DetailsServlet?id=WDU19890300163`. Accessed January 2016.

The Real Estate Management Act (1997). Available at `http://isap.sejm.gov.pl/DetailsServlet?id=WDU19971150741`. Accessed January 2016.

Wallgren, A. & Wallgren, B. (2014). *Register-based Statistics.* 2th ed. New York: Wiley.

Zhang, L.-C. (2011). A Unit-Error Theory for Register-Based Household Statistics. *J. Official Statist.*, 27, **3**, 415–432.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Stat. Neer.*, 66, **1**, 41–63.

Zhang, L. C., Thomsen, I., & Kleven, Øyvin. (2013). On the Use of Auxiliary and Paradata for Dealing With Non-sampling Errors in Household Surveys. *Internat. Statist. Rev*, 81, **2**, 270–288.

Zhang, L.-C. (2015). On Modelling Register Coverage Errors. *J. Official Statist.*, 31, **3**, 381—396.

Table 1: Sample rows and columns from the dataset prepared by one of the data owners

| City | Year and Month | Floor area [m$^2$] | Rooms | Count | Average price per $m^2$ |
|---|---|---|---|---|---|
| Poznań | 2011-01 | to 20 | 1 | 5 | 4363.2 |
| Poznań | 2011-01 | 20-30 | 1 | 285 | 6399.3 |
| Poznań | 2011-01 | 30-40 | 1 | 577 | 6370.3 |
| Poznań | 2011-01 | 40-50 | 1 | 56 | 5695.5 |
| Poznań | 2011-01 | 50-60 | 1 | 12 | 4210.3 |
| ... | ... | ... | ... | ... | ... |

Compiled using data from one of the data owners. City – the name of the city, Year and Month – the reference month with the year, Floor area – floor area divided into intervals, Rooms – the number of rooms, Count – the number of objects that meet the aggregation criterion, Average price per $m^2$ – average price per $m^2$ in a given aggregate .

Table 2: Selected variables available in the Register of Real estate Prices and Values

| Variable | Description |
|---|---|
| Date of transaction | Exact date of transaction |
| Transaction ID | Transaction ID which may refer to several properties (e.g. flat and garage) |
| Object ID | Objects that were included in the transaction |
| Total Price | Total transaction price (for all objects) |
| Object Price | Prices of objects included in the transaction |
| Use of Property | Intended use of the property (e.g. residential, non-residential) |
| Mortgage Number | The document number in the mortgage register |
| City | Name of the city where the property is located |
| Address | Precise information about location with street name, floor number, etc |
| Floor area | Floor area of a given property measured in square meters |
| Number of Rooms | The number of rooms in a given property |
| Transaction market | Primary, secondary market or auction |
| Seller | Information about the seller: legal or natural person |
| Buyer | Information about the buyer: legal or natural person |

Table 3: Advantages and disadvantages of measuring representativeness using individual and aggregate data

| Aggregation level | Advantages | Disadvantages |
|---|---|---|
| Individual data | 1. Measuring representativeness at any level | 1. Limited access to individual data |
| | 2. Detection of the selection mechanism | 2. Time consuming data cleaning process |
| | 3. Control over data processing and cleaning | 3. Linkage uncertainty |
| | 4. Linkage with units or objects from statistical and non-statistical data sources | 4. Linkage may be legally prohibited or impossible |
| | 5. Assessment of uncertainty of estimates | |
| Domain data | 1. Alternative when individual data are not available | 1. Limited possibilities of measuring representativeness and the selection mechanism |
| | 2. Overall information about consistency with official and non-official data | 2. Requires historical time series data for comparison |
| | 3. May provide a general overview of the data without time consuming data cleaning process | 3. Requires harmonization with available domain-level data |
| | 4. May indicate whether the use of such data is possible for official statistics | 4. Lack of a measure uncertainty of estimates (also for comparison) |

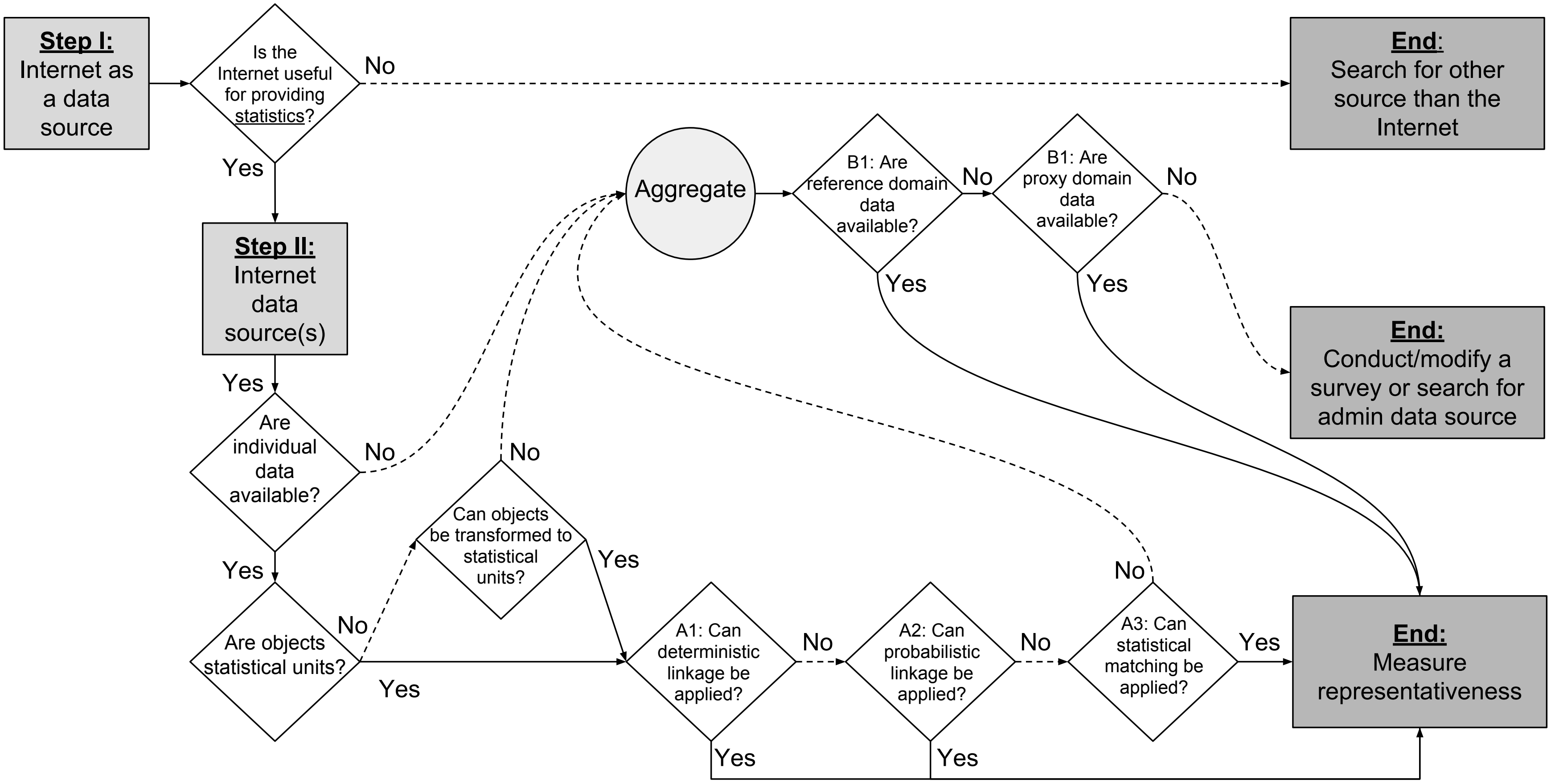
Step I:
Internet as a data source
Is the Internet useful for providing statistics?
No
End: Search for other source than the Internet
Yes
Step II: Internet data source(s)
Aggregate
B1: Are reference domain data available?
No
B1: Are proxy domain data available?
No
Yes
Yes
End: Conduct/modify a survey or search for admin data source
Yes
Are individual data available?
No
No
Can objects be transformed to statistical units?
Yes
Yes
Are objects statistical units?
No
Yes
No
A1: Can deterministic linkage be applied?
No
A2: Can probabilistic linkage be applied?
No
A3: Can statistical matching be applied?
Yes
End: Measure representativeness
Yes
Yes

Figure 1: The two-step procedure to measure representativeness