Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 509–529, http://dx.doi.org/10.2478/JOS-2019-0022

Correlates of Representation Errors in Internet Data Sources for Real Estate Market

Maciej Beręsewicz[1]

New data sources, namely big data and the Internet, have become an important issue in statistics and for official statistics in particular. However, before these sources can be used for statistics, it is necessary to conduct a thorough analysis of sources of nonrepresentativeness.

In the article, we focus on detecting correlates of the selection mechanism that underlies Internet data sources for the secondary real estate market in Poland and results in representation errors (frame and selection errors). In order to identify characteristics of properties offered online we link data collected from the two largest advertisements services in Poland and the Register of Real Estate Prices and Values, which covers all transactions made in Poland. Quarterly data for 2016 were linked at a domain level defined by local administrative units (LAU1), the urban/rural distinction and usable floor area (UFA), categorized into four groups. To identify correlates of representation error we used a generalized additive mixed model based on almost 5,500 domains including quarters.

Results indicate that properties not advertised online differ significantly from those shown in the Internet in terms of UFA and location. A non-linear relationship with the average price per m^2 can be observed, which diminishes after accounting for LAU1 units.

Key words: Big data; non-ignorable missing data; representation error; self-selection error; INLA.

1. Introduction

Big data and the Internet as a data source have become an important issue in statistics, in particular in official statistics. There are number of multinational initiatives (e.g., ESSnet on Big Data) that focus on the quality and suitability of estimates based on new data sources to complement or supplement existing statistical information. Before these data can be used for official statistics, it is crucial to explore potential sources of nonrepresentativeness. In this context, Daas et al. (2015), Buelens et al. (2014), Beręsewicz (2016, 2017), and Citro (2014) discussed coverage, nonresponse and measurement errors. Japec et al. (2015), Pfeffermann (2015), and Beręsewicz et al. (2018)

[1] Poznań University of Economics and Business, Department of Statistics, al. Niepodległości 10, 61-875 Poznań, Poland. Email: maciej.beresewicz@ue.poznan.pl

Acknowledgments: This study was supported by the National Science Center, Preludium 7 grant no. 2014/13/N/HS4/02999. I would like to kindly thank (1) Grupa OLX sp. z o.o. owner of the Otodom service, (2) Polska Press sp. z o.o. owner of the Dom.Gratka.pl service for providing access to the data via API and historical data in an aggregated form and (3) The Trade and Service Department, Statistics Poland for preparing aggregated data from the Register of Real Estates Prices and Values in Poland. I thank Alicja Szabelska-Beręsewicz, Kamil Wilak, Tomasz Klimanek, Michal Szczudlak and Tomasz Hinc for helpful comments and discussions. Moreover, I sincerely thank three anonymous reviewers and the associate editor for a valuable discussion that led to significant improvements in the article.

elaborate on the coverage and nonresponse error, which can lead to significant bias in big data sources, in particular if missingess is nonignorable.

A number of empirical studies compare new data sources with official statistics. For instance, Daas et al. (2015) studied the consumer confidence index based on social media and survey data in the Netherlands; the Billion Price Project is aimed at calculating CPI based on web-scraped data and Cavallo (2013) provided an insight into discrepancies between official indicators in Argentina.

Research is also conducted on the use of Internet data sources for the real estate market. Notable examples include *the number and prices of houses for sale* indices published by Statistics Netherlands from 2013 to 2016. The indices were calculated based on properties offered for sale on the JAAL.nl website (Statistics Netherlands 2018). Hoekstra et al. (2012) discussed details regarding data collection which, according to the authors' knowledge, can be considered the first use of online real estate data to produce official statistics. Unfortunately, the indices have been discontinued as part of cost cutting measures, being non-compulsory statistics.

Other examples can be found in the literature on real estate. For instance, there are a number of studies focusing on asking and transaction prices and values (Ihlanfeldt and Martinez-Vazquez 1986; Kiel and Zabel 1999) but they were mainly based on household surveys and register data (cf. Fleishman and Gubman 2015). In this context, Lozano-Gracia and Anselin (2012) describe the use of advertising signs and newspaper ads to survey asking prices of properties and link with cadastral records between 2002 and 2007 in Bogota, Columbia; Anenberg and Laufer (2017) used online ads to create an up-to-date list price index as a proxy for the price index based on administrative sources, and Beręsewicz (2016) investigated sources of bias in estimates of the average asking price per m^2 for residential properties by comparing survey data and advertising services in Poland.

New sources contain both measurement and representation errors associated with variables and objects (Wallgren and Wallgren 2014; Zhang 2012; Reid et al. 2017). Zhang (2012) proposed a two-phase life cycle model for integrated statistical microdata, where the first phase is based on a single source and the second one – on integrated sources. Reid et al. (2017) extended this model by including a third phase devoted to the evaluation or estimation of the quality of the final outputs, taking into account all sources of error.

In the article we focus on the representation aspect, emphasizing that it is crucial to keep in mind differences in measurements when using these data sources for statistics. Representation or non-observation errors include frame, selection and missing/redundancy errors. All of these errors are discussed briefly in the case of a single source.

Frame errors are differences between the target population and the accessible set. In this context Reid et al. (2017) distinguish the following measures: lag in updating population changes, undercoverage, overcoverage and authenticity (incorrect or multiple identifiers). In the case of Internet data sources, reporting lags can be linked to differences between the moment when information is posted online and the actual event (e.g., a flat is offered for sale a couple of weeks before it is published online). Undercoverage refers to a situation where some units are not observed online (e.g., properties advertised in newspaper or between friends), while overcoverage error occurs when a given unit does not belong to the target population. The second situation is common in Internet data sources because online services rarely have tools to verify if a given unit is correctly classified (e.g., a house or flat) or whether it exists (e.g., a future investment or property that is already being built). Finally, information published online may not contain any identifiers (e.g., no property parcel number).

Selection errors arise when objects in the accessible set do not appear in the accessed set. Reid et al. (2017) propose four indicators: adherence to the reporting period, dynamics of births and deaths, readability and inconsistent objects/units. The reporting period may refer to a situation when a statistical agency has an agreement with a data provider, which states that the data should be delivered on fixed date(s). However, in most cases special tools are developed in order to scrape or access the databases directly to decrease the reporting burden on the data holder (Hoekstra et al. 2012). Readability is certainly an issue in Internet data sources, mainly owing to restrictions on publicly available data (e.g., 1% sample of public tweets) or limited access rights to data (e.g., query results from the browser and API may vary).

Finally, missing/redundancy errors arise from the misalignment between the accessed set and the observed set, which could be measured by unit nonresponse rate, share of duplicated records or share of units that have to be adjusted to create statistical units (Reid et al. 2017). In Internet data sources we only observe units that either use the Internet or place information online (e.g., property ads). Given unrestricted possibilities of creating multiple accounts and content, the number of duplicated records is significant (e.g., several advertisements for the same property). Finally, objects in Internet data sources may refer to multiple statistical units (e.g., advertisement for a property and a garage with separate mortgages).

To assess these errors, it is necessary to rely on external data sources. However, in the case of Internet data sources (e.g., advertisements services), this may be problematic. For instance, there are rarely official statistics on this topic, there is no sampling frame for such units or persons/institutions that publish information online, and research on why such services are used is scarce. Some information about sources of errors can be found, for instance, in *the Information and Communication Technology* surveys coordinated by Eurostat. Certainly, differences between the target and observed set result from the underlying selection mechanism that prompts persons/companies to use certain services. Thus, it may be difficult to disentangle error and bias in new sources into frame and selection error and bias, which is why we will use representation errors as a general term for these errors.

In situations where sources of representation errors are unknown, they can be detected by comparing new data sources with auxiliary sources already used in statistics that is, surveys or registers (Beręsewicz 2017; Pfeffermann 2015; Lohr and Raghunathan 2017). Data can be linked at a domain level to provide information about the selection mechanism and characteristics of units that are not present online. It is crucial to discover the underlying selectivity because it may be linked to the effects and methods of dealing with these errors (Brick 2015).

The study described below focuses on residential properties in the secondary real estate market in Poland. The population of interest consists of residential properties offered for sale in the secondary market. This population may be of interest to official statistics, particularly when it comes to estimating the asking-to-transaction-price ratio, price indices, measuring time-to-sale, or as an indicator of the situation in the real estate market. As this population is not dynamic, unlike, for instance, the population of mobile phone or social website users, the related data may not be considered as big data in terms of volume, but should be treated as such in terms of variety or complexity. One should keep in mind that, as Citro (2014) states, the Internet, (. . .), not only generates a great deal of today's "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.

Given the nature of the population, as well as limited research on real estate market brokers and owners, we cannot separate representation errors into frame and selection errors. To investigate possible correlates of these errors, we used an auxiliary data source, namely the Register of Real Estates Prices and Values, which covers all sold properties that have an established ownership. Using this independent source on a different but related population, we can obtain information about types of properties that are not advertised online but are sold. Since we obtained these data a year after the last transaction took place, errors due to lags in registration can be regarded as negligible.

In the absence of access to a national unit-level register, quarterly data for 2016 were linked at a domain level defined as an interaction between the urban/rural distinction, the category of usable floor area and Local Administrative Unit (LAU 1, 380 districts) in Poland. To account for the time-lag between the moment of publishing advertisements and actual transactions, we linked transactions from q with advertisements from $q-1$, where $q=\{2,3,4\}$ refers to the given quarter in 2016. This variable represents the time-to-sale, which is a lag between posting an advertisement online and the sale of the property. In total, 5,507 domains (including quarters) with a non-zero number of transactions for 376 districts were analysed. Bias was not examined because of the measurement error resulting from the difference in the definition of the target variables in the sources (asking vs transaction price).

The research questions that the article seeks to answer are as follows:

- what are correlates of non-observation errors?
- is the non-observation error non-ignorable?

The article has the following structure. Section 2 covers the data collection design and a thorough description of data sources used in the study. Section 3 defines the measure of selectivity, describes the modelling procedure and the model used to explore correlates of selectivity. Section 4 is devoted to exploratory data analysis and the presentation of modelling results. The article ends with conclusions and a discussion of the results.

2. Data

2.1. Internet Data Sources – Otodom.pl and Dom.Gratka.pl

The process of data acquisition was designed to minimize errors and interruptions. This is why web-scraping was not considered as a mode of data collection, as it is sensitive to changes in the structure of the webpage, the IP can be blocked and often not all data are present on the webpage.

Instead, we decided to approach owners of two leading Polish online real estate advertising services – Otodom.pl (https://www.otodom.pl, further on referred to as Otodom) and Dom.Gratka (http://dom.gratka.pl, later on referred to as Gratka) – to inquire about the possibility of accessing their databases via the Application Programming Interface (API). Acquiring data through the API is a more robust solution and results in structured and highly dimensional data.

We contacted the companies by sending a formal e-mail inquiry with a clear description of the purpose the data would be used for. Both companies were open to collaboration, interested in the results and shared their data. In the case of Otodom, access was free of charge, while data from Gratka were made available for a small fee. No special conditions were made by the companies, except for a request made by Gratka to prepare a short note for their blog about current trends in prices on their service. Finally, we were given special access tokens and passwords to connect to the databases through the Simple Object Access Protocol (SOAP) and Representational State Transfer (REST) APIs. In addition, we received monthly historical data in an aggregated form, which had been prepared according to our request.

The scope of access varied in each case. Gratka API only offered access to 26 variables, which described each advertised property, while the Otodom data set included 46 fields, which, in addition to property characteristics, contained anonymised information about the person/company that placed the ad, and ad characteristics (e.g., promoted, number of views). This corresponds to the distinction between 'accessible set' and 'accessed set' proposed by Zhang (2012) but with regard to the number of variables made available.

The data collection process started in Q4 2015 and still continues. Data were collected on a weekly basis. Each Saturday night a script was run to download all advertisements available for the whole of Poland. First, the raw data were stored in plain text files: the initial volume was about 100 GB for Gratka and 900 GB for Otodom. Then, the data were processed on a Linux server using `bash` and `jq` (data were stored in JavaScript Object Notation, JSON, format). Later on, the data were processed using R (R Core Team 2017) language with the help of the following packages: `data.table`, `tidyverse` and `stringi`. The initial number of advertisements from Otodom was over 20 million and from Gratka – 28 million. However, the data set contained duplicates as a result of the data collection process and the organization of the Polish property market.

The editing phase started by analyzing the data structure and metadata associated with the variables (e.g., definitions). Then, we removed objects that did not belong to the target population (e.g., not located in Poland, unfinished investments, primary market), contained erroneous or missing values in prices or usable floor area (e.g., properties with PLN 1 price or with UFA of over 30 000 m^2), did not belong to the reference period based on the date of the last modification or contained information regarding multiple properties. Then, some variables were harmonized (e.g., build year) in order to ensure consistency between partially standardized and non-standardized data (e.g., information provided directly by the owner/broker).

Properties in the Polish real estate market can be sold under closed or open agreements, which means that the same properties are advertised multiple times. Under an open agreement, multiple brokers can place an ad regarding the same property, often with different descriptions, and data holders cannot remove duplicates. This required additional attention during the data processing stage. As the study involved quarterly data, deduplication was conducted within quarters based on the following *naive* procedure: (1) the most recent advertisements were selected based on the ad identification number, and then (2) using combinations of variables referring to province, price, usable floor area, number of rooms, year of construction and street name. Certainly, not all duplicates were located because of slight differences in values in floor area or street name. This problem could be resolved by probabilistic methods (cf. for recent advances Steorts et al. 2016; Chen et al. 2018). However, as most cases of duplicate ads were found in large cities (regional capitals), which are already covered by advertising portals, they did not pose a problem in the analysis at the domain level.

Finally, properties that did not have any information regarding location or could not be geocoded in order to be classified as either rural or urban were removed. All advertisements were geolocated based on the district and location using Google Geocode API. In view of the goal of the study no imputation was applied.

After the cleaning process, the final Gratka data set consisted of 816,100 ads with 526,720 and for Otodom – 699,958 ads with 394,953 unique objects.

2.2. *The Register of Real Estate Prices and Values*

The Register of Real Estate Prices and Values, later on referred to as the register, is a public register maintained by the district governor, which contains information about real estate prices included in notarial deeds and real estate values provided by real estate appraisers in appraisal reports, whose abridged versions are included in the register of land and buildings. It is worth noting that there is no single national register but each district (LAU1) in Poland maintains its own register (380 units). The data are reported to the Central Statistical Office in Poland five times a year – two months after each quarter and by April of the following year for the whole previous year. The register contains information on the population of sold properties: flats, buildings, built-up and land properties. The data are used to prepare an annual report entitled *Real Estate Sales*. Since 2015, these statistics have been broken down by market type – primary and secondary.

The Central Statistical Office divides residential properties into those sold in the primary and the secondary market. A primary market sale is defined as a transaction made in the free market, where the selling party is a legal person and the average price per 1 m^2 of usable floor area is at least 2,000 PLN. Transactions in the secondary market include other market transactions carried out in the free market and auction sales.

Aggregated quarterly data for 2015 and 2016 from the Register were obtained from the Trade and Service Department of the Central Statistical Office. The data contain information for the following variables: (1) district identifier and name, (2) market type (total, primary, secondary), (3) location (total, rural/urban, town with district status, town with district status with population of under 200,00, town with a district status with a population of over 200,000, unknown location), (4) categorized usable floor area in m^2 (total, under 40, 40–60, 60–80, and over 80 m^2), and variables (1) number of transactions, (2) value of transactions, (3) sum of floor area and (4) median price per m^2.

For the purpose of the study we selected domains defined as interactions between three quarters (2016 Q2-Q4), LAU 1, rural/urban distinction and four categories of usable floor

Table 1. Distribution of number of transactions within studied domains between Q2 and Q4 2016.

Total	No of domains	Min	1st Quantile	Median	Mean	3rd Quantile	Max
74 588	5 507	1	2	4	13.54	10	1 450

area. The analysis involved a total of 5,507 domains with a non-zero number of transactions in the secondary market for 2016. In that year transactions in the secondary market were reported in 376 out of 380 districts.

Table 1 presents the distribution of the number of transactions across the domains for 2016. The median number of transactions is 5 and the mean is 13.5, which indicates right skewness. In total, over 73,000 transactions were made in the secondary real estate market.

Table 3 and Figure 7 in the Appendix (Section 6) present a comparison between the register, Otodom and Gratka at domain levels. Note that there is a time-lag: online data refer to the Q1-Q3 2016 period and transactions to the Q2-Q4 2016 period. Pearson's correlation coefficient between the log-transformed number of transactions and the log-transformed number of ads in Otodom and Gratka is equal to 0.76 and 0.73 respectively, and between Otodom and Gratka – 0.88. Points over the black line indicate domains where the number of transactions is larger than that of ads.

2.3. Other Auxiliary Information

To account for the Internet coverage error, we used the broadband penetration ratio calculated as the number of buildings with access to broadband Internet (i.e., buildings for which Internet providers are able to provide broadband services) to all buildings in a given domain. This measure is calculated by the Office of Electronic Communications for all cities in Poland on a yearly basis. We calculated this indicator for urban/rural areas within LAUs using data for 31 December 2015.

3. Methods

3.1. The Approach

The following approach was adopted in the study. Otodom and Gratka were linked with register data at domain level. Because only domains with a non-zero number of transactions were selected, representation error was measured with reference to domains containing sold residential properties. Further, we defined the target variable, denoting non-observation error, by equation (1).

$$y_d^{(q)} = \begin{cases} 1, & \text{when } n_d^{(q)} > m_{d,otodom}^{(q-1)} \text{ and } n_d > m_{d,gratka}^{(q-1)}, \\ 0, & \text{else,} \end{cases} \tag{1}$$

Where $n_d^{(q)}$, $m_{d,otodom}^{(q-1)}$, $m_{d,gratka}^{(q-1)}$, denote the number of transactions, advertisements on Otodom, advertisements on Gratka in domain d and q-th quarter respectively. Domains were created by interaction between LAU1 units, urban/rural area and four floor area categories for each quarter. The target variable refers to transactions that did not involve online advertisements. Moreover, as we are dealing with two sources, we are interested in domains that are not represented in any of the advertisement sources. In total, 1,533 out of 5,507 domains were not represented (27.8%).

To detect the correlates of representation errors, and to avoid measurement error, we constructed a generalized additive mixed model which only contains variables from the register

- usable floor area categorized into four groups (`floor_area`),
- urban or rural location (`urban_rural`),
- average price per m^2 (in 1,000 PLN) at domain level, (`average_pricem2`, centered at overall mean equal to 2,554 PLN), and
- broadband Internet coverage in urban and rural areas at LAU1 level (`net_coverage`, centered at overall mean equal to 78.5%).

In order to verify which variables are correlated with the target variable we built the following four models, each serving a different purpose:

- `Model 1` – a generalized linear model with `net_coverage`, `floor_area`, `urban_rural` and interaction `floor_area` and `urban_rural` – this model is used as a baseline model to verify the relationship with the dependent variable.
- `Model 2` – we extended `Model 1` by adding the `average_pricem2` variable, assuming a non-linear relationship, using smoothing spline and thus obtaining a generalized additive (mixed) model – this model is used to verify if representation error is non-ignorable.
- `Model 3` – we extended `Model 1` by adding the `LAU 1` (i.i.d.) random effect and thus obtaining generalized a mixed model – this model is used to account simultaneously for clustering and for the characteristics of the local market at `LAU 1` level.
- `Model 4` – we combined `Model 2` and `Model 3` to obtain a generalized additive mixed model – this final model is used to verify if the errors are non-ignorable by accounting for both the average price and characteristics of the local market.

3.2. The Model

We used Integrated Nested Laplace approximation (INLA), which is a new approach to Bayesian inference for latent Gaussian Markov random field (GMRF) models proposed by Rue et al. (2009). The basic idea behind INLA involves using a deterministic approach to approximate Bayesian inference for latent Gaussian models (i.e., GMRF), which, in most cases, makes INLA faster (i.e., a matter of seconds rather than minutes) and more accurate than MCMC alternatives to GMRF. It provides a number of likelihoods and latent models, including spatial random effects, but it is not as flexible as standard Bayesian approaches (see Chen et al. 2014 and Mercer et al. 2014 for an application of INLA for small area estimation with sampling weights and for an introductory book – Faraway et al. 2018). INLA is implemented in C++ but it can be applied by using `R-INLA` package (Lindgren and Rue 2015).

In the empirical study, we modelled the $\mathbf{y}$ variable defined in (1) (we drop q for simplicity), and therefore we assume that it has a binomial distribution given by

(2), where ρ_d denotes the propensity of representation error modelled by means of logistic regression:

$$(y_d|\rho_d) \sim Binomial(n_d, \rho_d), \quad d = 1, \ldots, D.$$
$$\rho_d(\eta_d) = \frac{\exp(\eta_d)}{1 + \exp(\eta_d)}. \tag{2}$$

We considered four models for $\rho_d(\eta_d)$:

- `Model 1`

$$\rho_d(\eta_d) = \frac{\exp(\eta_d)}{1 + \exp(\eta_d)} = \mathbf{x}_d^T\beta, \tag{3}$$

where β are fixed effects parameters and $\mathbf{x}_d$ are independent variables at domain level (i.e., `net_coverage`, `floor_area`, `urban_rural` and `floor_area` × `urban_rural`),

- `Model 2`

$$\rho_d(\eta_d) = \frac{\exp(\eta_d)}{1 + \exp(\eta_d)} = \mathbf{x}_d^T\beta + \nu_j, \tag{4}$$

where ν_j refers to smoothing spline for the price per m^2. In INLA it is modelled by random walk of order 2 (RW2) and implemented as a random effect. For the Gaussian vector $\nu = (\nu_1, \ldots, \nu_n)$ smoothing spline is constructed assuming independent second-order increments:

$$\Delta^2\nu_j = \nu_j - 2\nu_{j-1} + \nu_{j-2} \sim N\big(0, \tau_\nu^{-1}\big). \tag{5}$$

The density π of ν is derived from its $n - 2$ second-order increments as:

$$\pi(\nu|\tau_\nu) \propto \tau_\nu^{(n-2)/2}\exp\left\{-\frac{\tau_\nu}{2}\sum(\Delta^2\nu_j)^2\right\} = \tau_\nu^{(n-2)/2}\exp\left\{-\frac{1}{2}\nu^\tau\mathbf{Q}\nu\right\}, \tag{6}$$

where $\mathbf{Q} = \tau\mathbf{R}$ and $\mathbf{R}$ is the structure matrix reflecting the neighborhood structure of the model given by:

$$\mathbf{R} = \tau_\nu \begin{pmatrix} 1 & -1 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}. \tag{7}$$

We considered $n \in \{10, 15, 20, 25\}$ and, based on information criteria described in the next section, we selected $n = 20$.

- `Model 3`

$$\rho_d(\eta_d) = \frac{\exp(\eta_d)}{1 + \exp(\eta_d)} = \mathbf{x}_d^T\beta + \psi_i, \tag{8}$$

where $\psi_i \sim N\left(0, \tau_{\psi}^{-1}\right)$ refers to i.i.d. random effect for LAU 1 units, indexed by $i = 1, \ldots, 376$.

- `Model 4`

$$\rho_d(\eta_d) = \frac{\exp(\eta_d)}{1 + \exp(\eta_d)} = \mathbf{x}_d^T \beta + \nu_j + \psi_i, \tag{9}$$

where all parameters are defined as previously.

For both random effects (ν_j, ψ_i) we used the same penalized complexity (PC) prior suggested by Simpson et al. (2017). Under this new framework, a PC prior for the standard deviation $\sigma = 1/\sqrt{\tau}$ of a latent effect is set by defining parameters (u, α), so the interpretation is

$$P(\sigma > u) = \alpha, \quad u > 0, 0 < \alpha < 1. \tag{10}$$

Hence, PC priors provide a different way to propose priors on the model hyperparameters. In this study, we believe that the probability of the standard deviation being higher than 1 is quite small, so we set $u = 1$ and $\alpha = 0.01$ in the following prior for τ

$$\pi(\tau) = \frac{\lambda}{2} \tau^{-3/2} \exp(-\lambda \tau^{-1/2}), \tau > 0, \tag{11}$$

for $\lambda > 0$ where

$$\lambda = -\frac{\text{In}(\alpha)}{u}, \tag{12}$$

and (u, α) are the parameters of this prior.

3.3. Model Selection

Further, in order to select the most suitable model we used deviance information criterion (DIC; Spiegelhalter et al. 2002), Watanabe-Akaike information criterion (WAIC; Watanabe 2010) and the sum of the log of the conditional predictive ordinate (CPO; Held et al. 2010) values.

The DIC statistic is based on the deviance measure and the number of effective parameters. As in the case of AIC and BIC, models with smaller DIC are better supported by the data.

The WAIC statistic is a more fully Bayesian approach for estimating the out-of-sample expectation starting with the computed log point-wise posterior predictive density and then adding a correction for the effective number of parameters to adjust for overfitting (Gelman et al. 2014, Subsection 3.4).

However, as DIC may underpenalize complex models with many random effects, CPO statistic is often calculated. The CPO is based on the leave-one-out cross-validation procedure, which checks without re-running the model for each observation in turn. For more detail, see Held et al. (2010).

4. Results

4.1. Exploratory Data Analysis

Figure 1 presents the share of domains observed and not observed online for three categorical variables: four categories of usable floor area (`floor_area`), urban/rural area

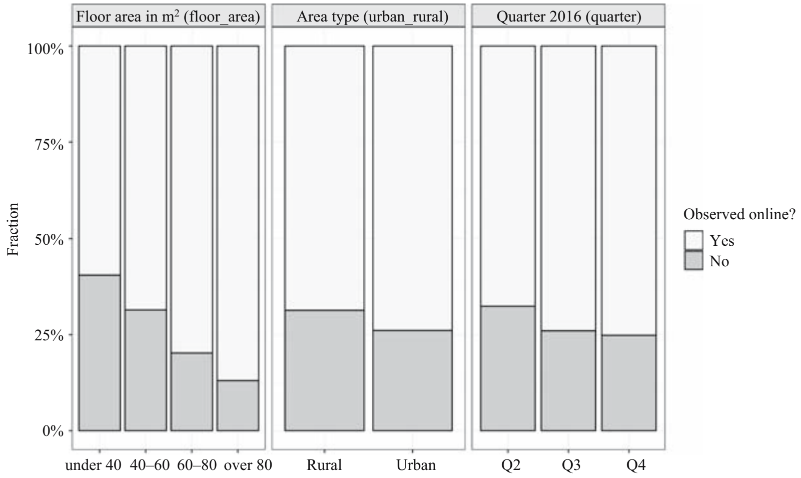

Fig. 1. Share of domains observed and not observed online for three categorical variables: usable floor area, location and quarters in 2016.

(urban_rural) and three quarters of 2016. The number of domains not observed online in the first quarter of 2016 is higher than in other quarters.

As expected, domains located in rural areas are less frequently observed online than those located in urban areas. This can be due to the lower broadband Internet coverage, as shown in Figure 2. Median Internet coverage for domains represented in the Internet sources was 83.5% and for those not represented online 75.6%. Another possible explanation is the difference in the use of the Internet by rural and urban dwellers and the fact that online advertising may not be equally necessary in small communities.

There is a linear relation between the categories of usable floor area (UFA) and the fact of being advertised online. Residential properties with UFA under 40 m^2 are more likely to be

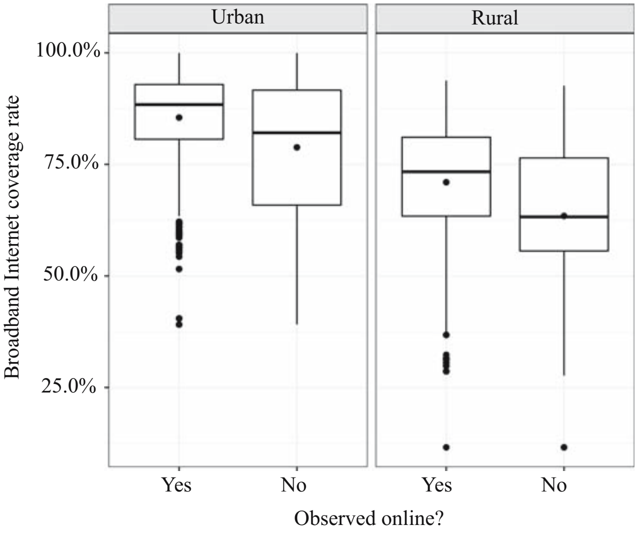

Fig. 2. Distribution of the broadband Internet coverage ratio in rural/urban areas and being observed online.

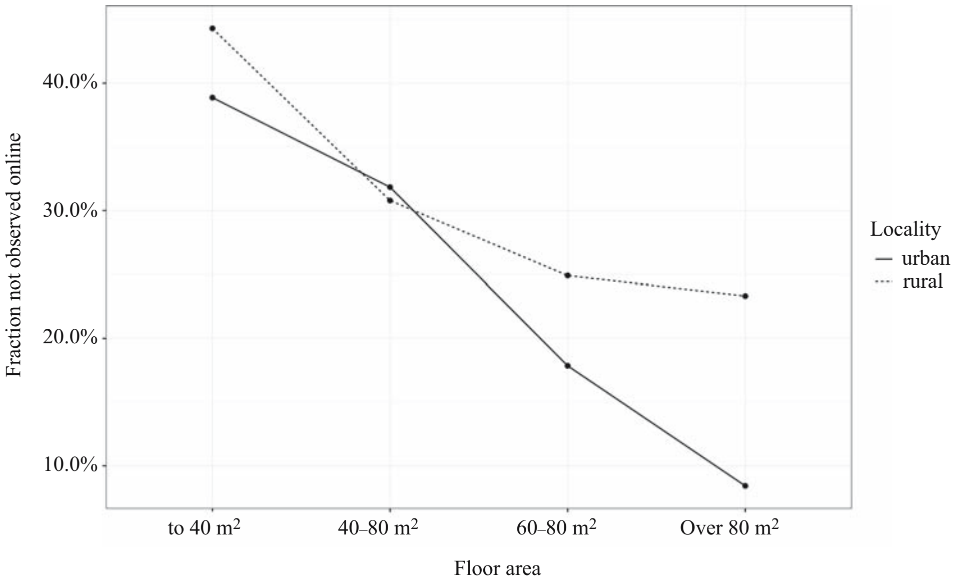

Fig. 3. Share of domains not observed online by locality.

sold without being advertised on two leading Polish portals. This indicates that the left tail of the UFA distribution is underrepresented on the Internet. Moreover, Figure 3 indicates the presence of an interaction between locality and UFA. Properties with floor area over 60 m^2 are more likely to be observed online in urban areas rather than in rural ones.

Figure 4 presents discrepancies in the distribution of `average_pricem2` between domains observed and not observed online. For clarity, `average_pricem2` is presented on a natural log scale and complemented by a rug plot under the density plots to visualize domain observations. The observed shift in the density plots suggests that Internet sources truncate the left tail of the price distribution.

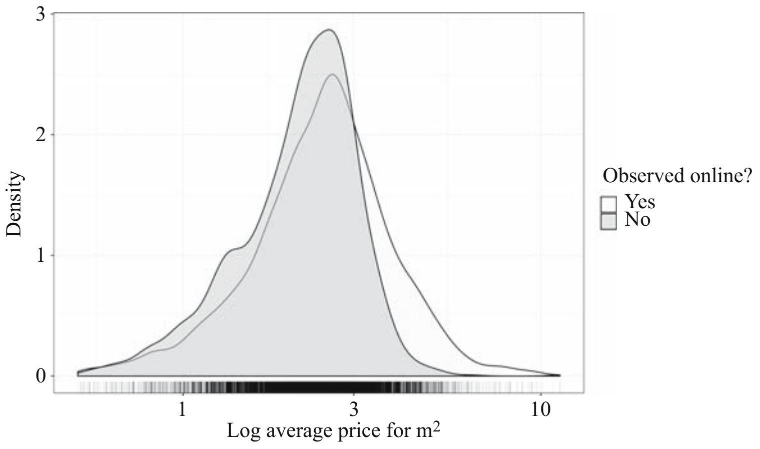

Fig. 4. Distribution of log-transformed average price m^2 depending on whether the domain was observed online or not (selectivity indicator) in 2016.

The results presented in the exploratory data analysis suggest the possibility of a non-ignorable selection mechanism at work in the secondary property market in Poland. In order to verify this hypothesis, we built a model that takes into account multiple covariates to detect the underlying data-creation mechanism.

4.2. Modeling Results

Table 2 consists of three parts and presents the summary of four estimated models described in Subsection 3.1. The top part presents estimates of odds ratios and standard errors for fixed effects. The middle part presents standard deviations and standard error for random effects. The bottom part contains three model selection measures.

If an estimate of a fixed effects parameter is larger than one, this means its odds of not being included in Internet data sources are high; if it is less than one, it means that domains with these characteristics are more likely to be observed online. More results regarding the model are presented in the Appendix (Section 6). Figure 8 in the Appendix presents posterior densities of the fixed effects estimated from Model 3 to facilitate the visual analysis of whether the model parameters differ from 0.

The results are in line with the analysis presented in Subsection 4.1. Bigger properties (over 40 m^2) are more often present online in comparison to those up to 40 m^2. The interaction between `urban_rural` and `floor_area` reveals differences between urban and rural areas. Bigger residential properties (over 60 m^2) in rural areas are more frequently absent from the Internet compared to urban areas. Only properties with UFA of 60–80 m^2 in urban areas seem to be equally represented in the online sources and the real estate register. As can be expected, the wider the Internet coverage, the smaller the non-observation propensity.

The parameters for `urban_rural` and `net_coverage` change slightly when random effect for LAU1 is introduced (Model 3 and 4). This is to be expected as these variables are characteristics of LAU1 units.

The random effects component accounts for the informativeness of selectivity measured is by adding `average_pricem2` to Model 1. WAIC and DIC statistics for Model 2 indicate that the average price per m^2 is a non-linear term because it improves Model 1. WAIC drops from 5905.1466 to 5676.7058 and DIC 5905.0123 to 5676.3313. This result suggests that selectivity might be non-ignorable given other characteristics of the real estate market.

However, if we introduce random effect for LAU1 unit rather than for `average_pricem2`, the drop in WAIC and DIC is significantly higher. WAIC drops by 1583.73 (in comparison to Model 1) and DIC decreases by 1559.917 (in comparison to Model 1). This suggests that Model 3 is better than Model 2. The variance component for LAU1 is almost twice as high as for `average_pricem2`.

Further, information criteria for Model 4 indicate that the model with both `average_pricem2` and LAU1 unit performs slightly less well than Model 3. This indicates that the LAU1 effect may account for prices within these units. This hypothesis is supported by Figures 5 and 6.

Figure 5 presents the relationship between the average price per m^2 and the non-observation propensity for Models 2 and 4. For both models we observe a non-linear relationship with the average price, and less expensive properties are more likely not to be observed online in comparison to more expensive properties. However, for Model 4 we

Table 2. Summary of fixed and random parameters and information criteria for the estimated models. For fixed effects only odds ratios are reported.

	Model 1		Model 2		Model 3		Model 4	
Parameters	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Fixed effects (odds ratios)								
Intercept	0.7572	0.0504	0.7311	0.0544	0.4908	0.0737	0.4842	0.0722
`net_coverage`	0.0168	0.0041	0.0227	0.0057	0.3504	0.1786	0.3321	0.1683
`urban_rural: Rural`	0.6741	0.0839	0.5607	0.0762	1.1612	0.1999	1.1134	0.2025
`floor_area: 40-60`	0.7112	0.0672	0.6782	0.0658	0.5526	0.0666	0.5428	0.0657
`floor_area: 60-80`	0.3267	0.0353	0.2983	0.0330	0.1700	0.0237	0.1660	0.0232
`floor_area: over 80`	0.1517	0.0242	0.1254	0.0203	0.0888	0.0178	0.0851	0.0172
`urban_rural x floor_area` interaction								
Rural and 40–60	0.7628	0.1265	0.7393	0.1250	0.6576	0.1348	0.6510	0.1343
Rural and 60–80	1.2024	0.2286	1.2666	0.2357	1.3543	0.3045	1.3560	0.3051
Rural and over 80	2.8316	0.6792	3.2472	0.7900	4.2850	1.2461	4.4505	1.3065
Random effects (σ)								
LAU1	–	–	–	–	2.2368	0.1287	2.1150	0.1254
`average_pricem2`	–	–	1.2704	0.3437	–	–	1.0382	0.3691
Model selection								
WAIC	5905.1464	–	5677.3259	–	4095.8715	–	4134.2167	–
DIC	5905.0121	–	5676.9841	–	4118.6910	–	4152.6810	–
$\sum \log(\mathrm{CPO}_d)$	−2952.5730	–	−2838.6670	–	−2051.4940	–	−2071.2050	–

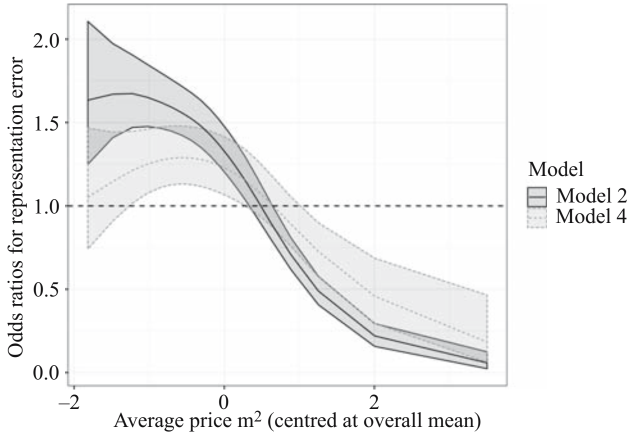

Fig. 5. Point estimates of odds ratios and 95% credible intervals estimated from random walk of order 2 for the average price per m^2 based on Model 2 and 4.

observe a diminishing effect for cheap properties, while for expensive ones, it remains more or less at the same level.

Figure 6 shows the distribution of the LAU 1 random effect for Model 4 in relation to the average price for m^2. This price was calculated as an average price for all domains within each LAU 1 unit. That is why the range of prices is different from that presented in Figure 5. There are several areas where properties, despite their average price, are always observed online (points below the dashed line) but the majority of LAU 1 units are above or close to the overall mean. 120 LAU 1 units (31%) have a credible interval over the dashed line denoting zero, which suggests that some domains within these LAU 1 units are not observed online. This, however, depends largely on the number of categories of floor area, and the time granularity considered.

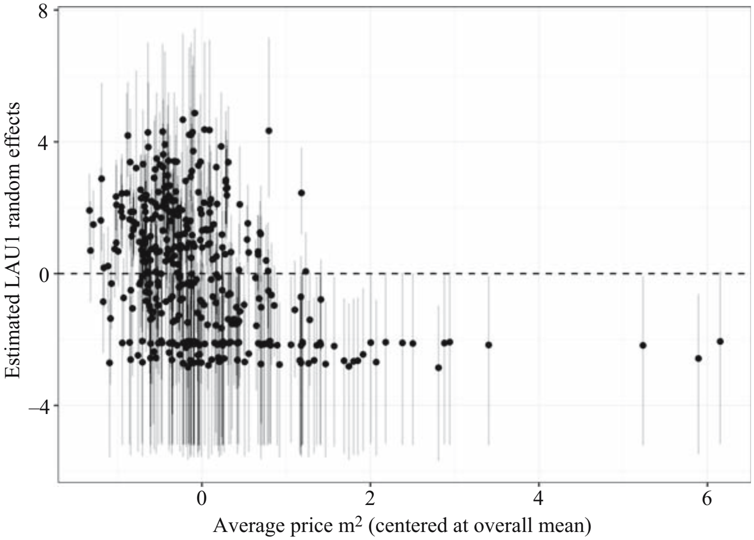

Fig. 6. Point estimates and 95% credible intervals of LAU 1 random effects and average price m^2.

5. Conclusion and Discussion

In the article we studied representation errors in Internet data sources for residential properties in the secondary market in Poland. We used the two biggest online advertising platforms that list real estate offers and one administrative source, which covers all transactions in this market. The auxiliary data source was used to detect correlates of representation errors and determine whether its missingness is non-ignorable.

The results suggest that representation errors are strongly correlated with usable floor area and Internet coverage. As expected, the selection mechanism is connected with the low level of aggregation (LAU 1 level), which is the dominant factor in random effects in the proposed models.

However, results of the estimated models are ambiguous. Based solely on information criteria, errors could be regarded as ignorable; however, when analyzing the relationship between the price and the fact of not being present online, a clear non-linearity is visible. This might also be connected with smaller properties (in terms of UFA) that are also characterized by lower prices.

A number of explanations can be proposed to explain such results. First, despite their size, these portals are mainly used by real estate brokers. Only 5% of offers listed on Otodom are placed by individual customers; data obtained from Gratka API do not contain such information and results in undercoverage. It is likely that brokers are the target group of premium customers because they place ads for more expensive properties. One potential way to overcome this problem is to use other services, which are targeted at different groups of people. In Poland, the OLX classified service can be a good example, as it also lists properties but, according to the OLX group, which owns Otodom and OLX, OLX users mainly include owners and people from rural areas. Second, properties in Poland do not have to be sold online, nor are they officially registered. Transactions involving properties not listed online can take place between family members or specific, small groups of customers.

Even though results are promising and support the research questions stated in the introduction, one should take into account that the study was conducted at domain level, which may have influenced the results. If units listed online could be linked with those included in the register, the analysis of correlates of self-selection error could be more accurate.

The problem of overcoverage regarding (1) duplicated entries, (2) outdated entries, (3) no longer for sale, and (4) false advertisements was not addressed in the data cleaning procedure. This issue cannot be easily tackled and requires additional attention. Therefore, to some (yet unknown) extent, results presented in the article may underestimate effects of the correlates of selectivity.

Keeping that in mind, the methods presented in the article can be used to select an appropriate method of correcting the selection bias. For instance, probabilities estimated on the basis of models described above could be used for propensity score weighting and then applied to online data. Another possible use involves the application of the model-based approach under the missing not at random (MNAR) mechanism to estimate asking prices for domains not covered by the online services. Other possible applications can be found in (Riddles et al. 2016; Sverchkov and Pfeffermann 2018; Sikov 2018; Heckman 1979; Marra et al. 2017).

Finally, the approach presented in the article could be applied to other sources given the availability of auxiliary variables (including proxies), both in these sources and in independent data (e.g., administrative records, sample surveys). Without access to such covariates, it will not be possible to detect errors or reduce bias. In other words, researchers interested in big data for official statistics should focus on variables that are highly correlated with the target variable.

6. Appendix

Table 3. Distribution of number of transactions (Register) and advertisements (Gratka, Otodom).

Quarter	Source	Total	Min	Q1	Median	Mean	Q3	Max
2016 Q1	Gratka	224 719	0	0	3	119	19	26 079
	Otodom	165 374	0	1	5	87	22	15 344
	Register	–	–	–	–	–	–	–
2016 Q2	Gratka	205 443	0	0	3	112	18	25 937
	Otodom	180 724	0	2	7	98	31	15 247
	Register	27 436	1	2	4	14	11	1 276
2016 Q3	Gratka	193 786	0	0	3	109	18	23 456
	Otodom	174 418	0	2	8	98	31	14 314
	Register	25 428	1	2	4	14	10	1 450
2016 Q4	Gratka	–	–	–	–	–	–	–
	Otodom	–	–	–	–	–	–	–
	Register	21 724	1	2	4	12	9	1 140

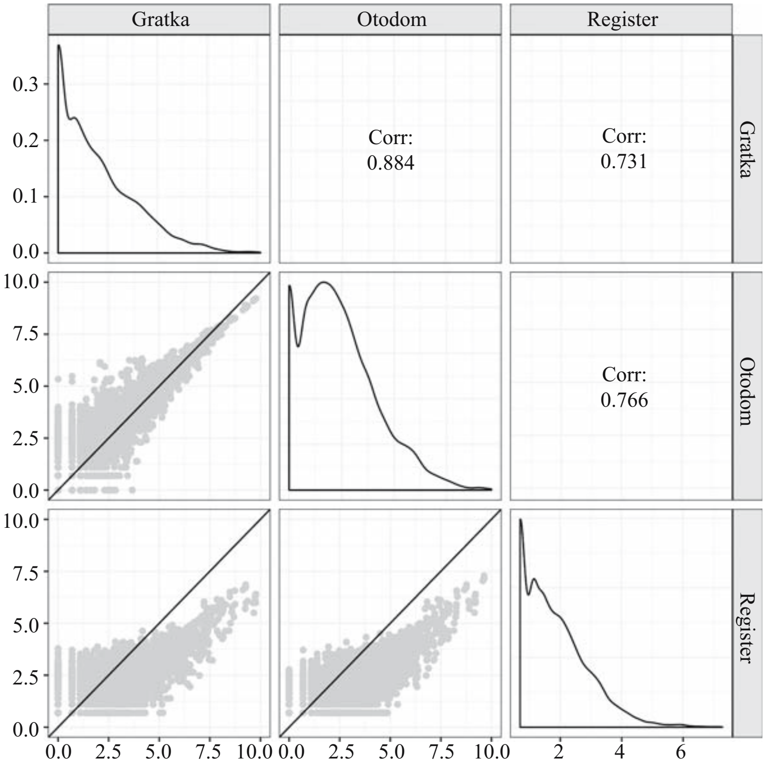

Fig. 7. Correlation of log number of transactions (Register) and advertisements (Gratka, Otodom) at domain level.

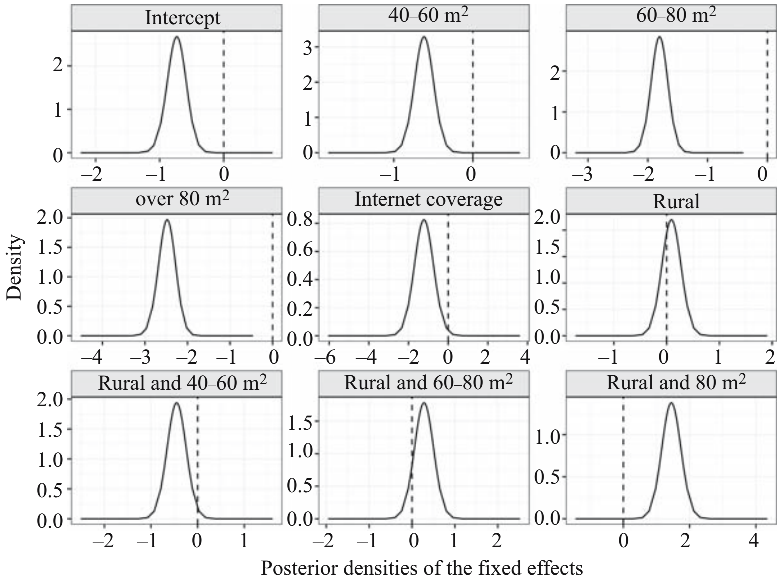

Fig. 8. Posterior densities of the fixed effects estimated under Model 3.

7. References

Anenberg, E. and S. Laufer. 2017. “A More Timely House Price Index.” *Review of Economics and Statistics* 99(4): 722–734. Doi: https://doi.org/10.1162/REST_a_00634.

Beręsewicz, M. 2016. *Internet Data Sources for Real Estate Market Statistics*. PhD diss., Poznań University of Economics and Business. Available at: http://www.wbc.poznan.pl/dlibra/docmetadata?id=393454 (accessed February 2019).

Beręsewicz, M. 2017. “A Two-Step Procedure to Measure Representativeness of Internet Data Sources.” *International Statistical Review* 85(3): 473–493. Doi: https://doi.org/10.1111/insr.12217.

Beręsewicz, M., R. Lehtonen, F. Reis, L. Di Consiglio, and M. Karlberg. 2018. *An Overview of Methods for Treating Selectivity in Big Data Sources*. Statistical Working Papers. Eurostat. Doi: https://doi.org./10.2785/312232.

Brick, J.M. 2015. “Unit Nonresponse and Weighting Adjustments: A Critical Review.” *Journal of Official Statistics* 29(3): 329–353. Doi: https://doi.org/10.2478/jos-2013-0026.

Buelens, B., P. Daas, J. Burger, M. Puts, and J. van den Brakel. 2014. *Selectivity of Big Data*. Discussion paper 201411. Statistics Netherlands, The Hague/Heerlen, The Netherlands. Available at: http://pietdaas.nl/beta/pubs/pubs/Selectivity_Buelens.pdf (accessed February 2019).

Cavallo, A. 2013. “Online and Official Price Indexes: Measuring Argentina’s Inflation.” *Journal of Monetary Economics* 60(2): 152–165. Doi: https://doi.org/10.1016/j.jmoneco.2012.10.002.

Chen, B., A. Shrivastava, and R.C. Steorts. 2018. "Unique entity estimation with application to the Syrian conflict." *The Annals of Applied Statistics* 12(2): 1039–1067. Doi: https://doi.org/10.1214/18-AOAS1163.

Chen, C., J. Wakefield, and T. Lumely. 2014. "The Use of Sampling Weights in Bayesian Hierarchical Models for Small Area Estimation." *Spatial and Spatio-Temporal Epidemiology* 11: 33–43. Doi: https://doi.org/10.1016/j.sste.2014.07.002.

Citro, C.F. 2014. "From Multiple Modes for Surveys to Multiple Data Sources for Estimates." *Survey Methodology* 40(2): 137–161.

Daas, P.J., M.J. Puts, B. Buelens, and P.A. van den Hurk. 2015. "Big Data as a Source for Official Statistics." *Journal of Official Statistics* 31(2): 249–262. Doi: https://doi.org/10.1515/jos-2015-0016.

ESSnet Big Data. 2018. "ESSnet Big Data." Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data (accessed February 2018).

Faraway, J.J., X. Wang, and Y.Y. Ryan. 2018. *Bayesian Regression Modeling with INLA*. Chapman/Hall/CRC.

Fleishman, L. and Y. Gubman. 2015. "Mass Appraisal at the Census Level: Israeli Case." *Statistical Journal of the IAOS* 31(4): 597–612. Doi: https://doi.org/10.3233/SJI-150939.

Gelman, A., J. Hwang, and A. Vehtari. 2014. "Understanding Predictive Information Criteria for Bayesian Models." *Statistics and Computing* 24(6): 997–1016. Doi: https://doi.org/10.1007/s11222-013-9416-2.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47: 153–161. Doi: https://www.jstor.org/stable/1912352.

Held, L., B. Schrödle, and H. Rue. 2010. "Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA." In *Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir*, edited by T. Kneib and G. Tutz, 91–110. Heidelberg: Physica-Verlag HD. Doi: https://doi.org/10.1007/978-3-7908-2413-1_6.

Hoekstra, R., O. ten Bosch, and F. Harteveld. 2012. "Automated Data Collection from Web Sources for Official Statistics: First Experiences." *Statistical Journal of the IAOS* 28(3, 4): 99–111. Doi: https://doi.org/10.3233/SJI-2012-0750.

Ihlanfeldt, K.R. and J. Martinez-Vazquez. 1986. "Alternative Value Estimates of Owner-occupied Housing: Evidence on Sample Selection Bias and Systematic Errors." *Journal of Urban Economics* 20(3): 356–369. Doi: https://doi.org/10.1016/0094-1190(86)90025-2.

Japec, L., F. Kreuter, M. Berg, P. Biemer, P. Decker, C. Lampe, J. Lane, C. O'Neil, and A. Usher. 2015. "Big Data in Survey ResearchAAPOR Task Force Report." *Public Opinion Quarterly* 79(4): 839–880. Doi: https://dx.doi.org/10.1093/poq/nfv039.

Kiel, K.A. and J.E. Zabel. 1999. "The Accuracy of Owner-provided House Values: The 1978–1991 American Housing Survey." *Real Estate Economics* 27(2): 263–298. Doi: https://doi.org/10.1111/1540-6229.00774.

Lindgren, F. and H. Rue. 2015. "Bayesian Spatial Modelling with R-INLA." *Journal of Statistical Software* 63(19): 1–25. Doi: https://doi.org/10.18637/jss.v063.i19.

Lohr, S.L. and T.E. Raghunathan. 2017. "Combining Survey Data with Other Data Sources." *Statist. Sci.* 32(2) (May): 293–312. Doi: https://doi.org/10.1214/16-STS584.

Lozano-Gracia, N. and L. Anselin. 2012. “Is the Price Right?: Assessing Estimates of Cadastral Values for Bogotá, Colombia.” *Regional Science Policy & Practice* 4(4): 495–508. Doi: https://doi.org/10.1111/j.1757-7802.2012.01062.x.

Marra, G., R. Radice, T. Bärnighausen, S.N. Wood, and M.E. McGovern. 2017. “A Simultaneous Equation Approach to Estimating Hiv Prevalence with Nonignorable Missing Responses.” *Journal of the American Statistical Association* 112(518): 484–496. Doi: https://doi.org/10.1080/01621459.2016.1224713.

Mercer, L., J. Wakefield, C. Chen, and T. Lumley. 2014. “A Comparison of Spatial Smoothing Methods for Small Area Estimation with Sampling Weights.” *Spatial Statistics* 8: 69–85. Doi: https://10.1016/j.spasta.2013.12.001.

Pfeffermann, D. 2015. “Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture.” *Journal of Survey Statistics and Methodology* 3(4): 425–483. Doi: https://dx.doi.org/10.1093/jssam/smv035.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: www.R-project.org/ (accessed February 2019).

Reid, G., F. Zabala, and A. Holmberg. 2017. “Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ.” *Journal of Official Statistics* 33(2): 477–511. Doi: https://doi.org/10.1515/jos-2017-0023.

Riddles, M.K., J.K. Kim, and J. Im. 2016. “A Propensity-score-adjustment Method for Non-ignorable Nonresponse.” *Journal of Survey Statistics and Methodology* 4(2): 215–245. Doi: https://doi.org/10.1093/jssam/smv047.

Rue, H., S. Martino, and N. Chopin. 2009. “Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion).” *Journal of the Royal Statistical Society B* 71: 319–392. Doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x.

Sikov, A. 2018. “A Brief Review of Approaches to Non-ignorable Non-response.” *International Statistical Review* 86(3): 415–441. Doi: https://doi.org/10.1111/insr.12264.

Simpson, D., H. Rue, A. Riebler, T.G. Martins, S.H. Sørbye, et al. 2017. “Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors.” *Statistical Science* 32(1): 1–28. Doi: https://doi.org/10.1214/16-STS576.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A. Van Der Linde. 2002. “Bayesian Measures of Model Complexity and Fit.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4): 583–639. Doi: https://doi.org/10.1111/1467-9868.00353.

Statistics Netherlands. 2018. *Indicatoren bestaande woningen in verkoop*. Available at: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksb eschrijvingen/indicatoren-bestaande-woningen-in-verkoop (accessed November 2018).

Steorts, R.C., R. Hall, and S.E. Fienberg. 2016. “A Bayesian Approach to Graphical Record Linkage and Deduplication.” *Journal of the American Statistical Association* 111(516): 1660–1672. Doi: https://doi.org/10.1080/01621459.2015.1105807.

Sverchkov, M. and D. Pfeffermann. 2018. “Small Area Estimation Under Informative Sampling and Not Missing At Random Non-response.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181(4): 981–1008. Doi: https://doi.org/10.1111/rssa.12362.

Wallgren, A. and B. Wallgren. 2014. *Register-based Statistics: Statistical Methods for Administrative Data*. New York: Wiley.

Watanabe, S. 2010. “Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory.” *Journal of Machine Learning Research* 11(Dec): 3571–3594. Available at: http://www.jmlr.org/papers/v11/watanabe10a.html (accessed February 2019).

Zhang, L.-C. 2012. “Topics of Statistical Theory for Register-based Statistics and Data Integration.” *Statistica Neerlandica* 66(1): 41–63. Doi: https://doi.org/10.1111/j.1467-9574.2011.00508.x.

Received March 2018
Revised March 2019
Accepted April 2019