Using Web-Data to Estimate Spatial Regression Models

Giuseppe Arbia¹ and Vincenzo Nardelli²

Abstract
Macro econometrics has been recently affected by the so-called ‘Google Econometrics’. Comparatively less attention has been paid to the subject by the regional and spatial sciences where the Big Data revolution is challenging the conventional econometric techniques with the availability of a variety of non-traditionally collected data (such as, e.g., crowdsourcing, web scraping, etc) which are almost invariably geo-coded. However, these unconventionally collected data represent only what in statistics is known as a “convenience sample” that does not allow any sound probabilistic inference. This paper aims at making aware researchers of the consequence of the unwise use of such data in the applied work and to propose a technique to minimize such the negative effects in the estimation of spatial regression. The method consists of manipulating the data prior their use in an inferential context.

Keywords
big data, crowdsourcing, spatial microeconometrics, spatial regression, web scraping

Introduction
Macro econometrics has been recently widely affected by the so-called ‘Google Econometrics’, which can be described as the use of Google search index (and, by extension, of internet derived data) instead of (or together with) fundamental macroeconomic indicators. These indicators have proved particularly useful for monitoring

¹Catholic University of the Sacred Heart Milan, Rome, Italy
²University of Milan-Bicocca, Milan, Italy

Corresponding Author:
Vincenzo Nardelli, Università degli Studi di Milano-Bicocca, Piazza dell’Ateneo Nuovo, Milano 20126, Italy.
Email: v.nardelli2@campus.unimib.it
demand and making forecasts in various markets. For instance, Fondeur and Karamé (2013) used Google data to predict unemployment. Different contributions can be found also in forecasting house market indices in US (Kulkarni et al., 2009) and China (Wei and Cao 2017) amongst the others.

A great impulse in this direction has obviously been received by the advent of the Big Data. Following the celebrated definition of Laney (2001), Big Data are characterized by three V’s, namely: Volume, Velocity and Variety (Arbia, 2021). Although the three characteristics are intrinsically connected, Volume specifically relates to the size of the dataset while Velocity refers to the speed with which data are generated and collected. Finally, Variety refers to the fact that many of the data currently available derive from different types and unconventional sources such as images collected though satellite sensors or texts extracted from social media or economic variables (e.g. prices) collected from web pages. While the characteristics of volume and velocity pose serious challenges to statistical techniques from a computational point of view when dataset are very large and collected in almost real-time, the third “V” (Variety) creates other kinds of problems, in terms of the data generating mechanisms, which require formal solutions prior the application of the standard estimation techniques.

Indeed, the explosion of Big Data has led to a greater availability of micro data and the contemporaneous availability of previously unconceivably rich datasets to empirically validate theoretical models. In particular, in the spatial sciences, until only a few years ago researchers were mostly interested in analyzing data and estimating models based on regional data. However, as it is well known, inference based on regional aggregates is intrinsically biased by the presence of the Modifiable Areal Unit Problem (or MAUP), in that results are influenced by both the geographical scale of analysis and the geographical grouping criterion adopted (see Gehlke and Bihel, 1934; Arbia, 1989 and Arbia and Petrarca, 2011). The increasing availability of individual geo-coded micro-data has shown a way to tackle the MAUP giving rise to the new branch of research termed spatial microeconometrics (Arbia et al., 2021).

It should be emphasized right at the beginning of this paper, that our interest is restricted to the distortions which derive from the unwise use of a large variety of unconventionally collected datasets in spatial analysis. Consequently, the examples we have in mind are not necessarily related to the computational issues that may arise from the use of very large (volume) and quickly growing (velocity) data.

When considering new unconventional data sources, two very common examples are represented by crowdsourcing (data voluntarily collected by individuals) and web scraping (data extracted from websites and reshaped in a structured dataset). These typologies of data with the addition of a spatial reference are also known as Volunteered Geographic Information (VGI) (Goodchild, 2007). Crowdsourced data are common in many situations. An example is represented by data collected through smart phones, to measure phenomena that are otherwise difficult to quantify precisely and timely. Among these, e. g., are the crowdsourced collection of food prices in developing countries (see, e. g. Arbia et al. 2018) and epidemiological data (Crequit et al., 2018). The practice of extracting data from the web and using them in statistical analyses is
also becoming more and more popular, e. g. to collect online prices in the real estate market (Beręsewicz, 2015; Boeing and Waddell 2017) or in consumer goods markets (Cavallo and Rigobon 2016). House prices have been extensively studied in the field of economics and urban planning for many years. Traditionally, researchers have used official government data sources to investigate trends and patterns in housing markets. However, in recent years, there has been an increasing interest in using non-conventional data sources to complement and expand traditional data sources. Loveridge and Paredes (2018) were able to compare the cost of living in rural and urban areas by using the Big Mac Index as a proxy for prices. Ayoubah et al. (2020) use web scraping techniques to collect data over a period of 2 years to examine whether there is a correlation between the number of Airbnb listings and the price of long-term rentals. The study demonstrates the potential of web data to shed light on important social and economic phenomena, such as the impact of digital platforms on traditional industries. Similarly, Wrede (2022) developed a spatial city model to analyze the relationship between quality, distance, and rents using web data.

A common characteristic of these two unconventional data collection sources is represented by the lack of any precise sample design. In crowdsourcing the participation is generally voluntary thus producing an evident self-selection of the collectors. Similar problems are encountered with web scraping. Indeed, most of the contents reported on the web are available for free. However, the different business models of the commercial web platforms affect the information ranking process, for instance in the order with which data are offered to the user. Furthermore, the advertisements and algorithms that optimize the time spent by the users inside the platform influence their navigation. Since many of the webpages are sponsored, different users on the same platform may end up accessing different contents depending on their profiling. This process has an impact on the data that can be retrieved from web and, as a consequence, webscraped data cannot be considered random because they are “filtered” by these mechanisms that induce self-selection bias and coverage problems. In the particular example reported in the empirical application in A Case Study: Estimating a Spatial Price Model With Web-Scraped Data in Milan (Italy) section, an obvious selection bias is constituted by the fact that we scraped only one real estate agency.

A similar problem emerges, for instance, when we extract from the web annual reports of small- and medium-sized companies to predict bankruptcy (Zoričák et al., 2020). This situation is described in statistics as “convenience sampling” (or more generally non-random sampling” (Hansen et al., 1953)) in the presence of which we cannot use the standard probabilistic inference methods. Indeed, while in a formal sample design the choice of observations is suggested by a precise mechanism which allows the calculation of the probabilities of inclusion of each unit (and, hence, sound probabilistic inference), on the contrary with a convenience collection no probability of inclusion can be calculated thus giving rise to over- under-representativeness of the sample units.

In the applied work this is a serious problem in that no inferential result is still valid leading to biases and inefficiencies in the estimators and unreliable confidence intervals and hypothesis testing procedures.
Given the relevance of the problem theoretical solutions are needed providing applied researchers with the necessary tools to tackle it.

Although several methods have been proposed in the literature to minimize the biases and inefficiencies deriving from non-probability sampling (see, e. g., Chen et al., 2020; Kim & Wang, 2019; Elliott & Valliant, 2017; Kim & Tam, 2020 and Yang et al., 2020), they do not deal specifically with the case of a spatial sample where we can exploit spatial trends and interactions to gather useful contextual information.

When we have access to this kind of data, Arbia et al. (2018) proposed a technique (termed post-sampling) which can be seen as a form of post-stratification (Holt and Smith, 1979; Little, 1993) and also falls within the general category of calibration methods (Deville & Särndal, 1992). In its original proposal the method consists of cancelling some of the data prior to their use in an inferential context to estimate a population mean. In this paper we aim to generalize such an approach to the estimation of spatial regression models. It is important to remark that the approach presented here is not just a data-mining strategy which eliminates problematic points or outliers to improve the model fit, but it is instead, a theoretically grounded approach that aims at transforming a set of chaotically collected data into a formal sample design that obeys to specific rules.

The layout of the paper is the following. In A General Correction Procedure for Data Gathered Through Convenience Sampling section we introduce our data-correction procedure in the general context of parameter estimation when we avail data gathered through convenience sampling. A Monte Carlo Assessment of the Effects of the Correction Mechanism in the Estimation of Spatial Regression Parameters section is devoted to show the implications of using convenience samples to estimate spatial regression models. In this section, we show that, by using data collected without a proper design, the parameter estimates in a spatial regression model (Arbia, 2014) are biased, and that such a bias can be reduced by using our correction mechanism although at the cost of increasing the estimator’s variance. In An Operational MSE-Correction Procedure section we present an operational procedure leading to an MSE-reduction mechanism which is then used in A Case Study: Estimating a Spatial Price Model With Web-Scraped Data in Milan (Italy) section to estimate a hedonic price model using data that are scraped from the web. In this case study, we show that it is possible to fine tune our strategy to find a compromise between the bias reduction and the increase in the variance of the estimators. The paper concludes with some final comments contained in Summary, Open Problems and Future Work section and with an Appendix reporting some formal derivations.

A General Correction Procedure for Data Gathered Through Convenience Sampling

The idea that if we aim at a satisfactory generalization of the sample results the sample experiment needs to be rigorously programmed, can be traced back to the early contributions of Sir Ronald Fisher (Fisher, 1935). Indeed, lacking a formal sample
design, data must be considered *convenience* (or *judgment*) sampling with which no sound probabilistic inference is possible (Hansen et al., 1953). This problem emerges dramatically in the Big Data era when we increasingly avail huge quantities of data which, almost invariably, do not satisfy the necessary conditions for probabilistic inference. In recent years researchers have become increasingly aware of this problem and suggested solutions to reduce the distorting effects inherent to non-probabilistic designs (see e. g. Fricker and Schonlau, 2002). Chen et al. (2020) develop a rigorous procedure for estimating the propensity scores for units in a non-probability sample and suggest a doubly robust technique to estimate a population mean. A doubly robust strategy is also suggested by Yang et al. (2020) to combine probability and non-probability samples. To tackle the problem of the selection bias often encountered when using big data Kim & Wang (2019) and Kim & Tam (2020) suggest to use auxiliary information (adopting an “inverse sampling” strategy) and data derived from independently observed probability samples. Finally, Elliott & Valliant (2017) discuss the relative merits of two possible approaches: the *quasi-randomization*, where pseudo inclusion probabilities are estimated, and the *superpopulation modeling* framework where a model is used in order to predict the values of the unsampled units. Although extremely useful in practical cases when estimating the basic population parameters (such as a mean or a percentage), all these methods do not contribute specifically to the case of estimating a spatial regression, a task which is, conversely, the aim of the present paper.

Generally speaking, one possible strategy that has been suggested is to discard some of the redundant data and to concentrate only on smaller datasets obtained as a rigorously programmed sample. In particular, Arbia et al. (2018) suggested to transform crowdsourced datasets by discarding observations in such a way that the reduced dataset resembles a formal sample design. This procedure has been termed *post-sampling*. The rationale behind this strategy can be traced back to the ideas expressed by the philosopher of science Jules-Henri Poincare’ long before the advent of Big Data: “If the scientist had at his disposal infinite time, it would only be necessary to say to him: ‘Look and notice well’; but, as there is not time to see everything, and as it is better not to see than to see wrongly, it is necessary for him to make a choice” (Poincare’, 1908).

Arbia et al. (2018) discussed three possible forms of post-sampling. The first procedure (termed *hard-core* post-sampling) consists in drawing a random subset of the units from the original dataset so that the surviving observations can be considered a random sample. This strategy has been used by Arbia et al. (2018) to estimate the food price index in Nigeria using crowdsourced data. Adopting this strategy, however, while correcting for the bias, we are doomed to produce an increase in the variance of the estimators due to the diminished sample size.

For this reason, a second procedure (termed *flexible* post sampling) can be adopted allowing only a partial elimination of the units thus compromising between the reduction of bias and the increase in the estimator’s variance. To this aim, we eliminate only a share of the units that are necessary to consider the dataset a random sample, controlling this share with a parameter, say \( \zeta \) such that \( 0 \leq \zeta \leq 1 \). When \( \zeta = 0 \) no
correction is introduced. Conversely, when $\zeta = 1$ we have the hard-core post-sampling previously described which minimizes the bias, but increases the estimation variance. Varying the value of the parameter $\zeta$ in the range between 0 and 1 it is possible to create different datasets compromising between bias and the increase in the estimation variance (Nardelli, 2019).

Indeed, both the hard-core and the flexible approach can be nested into a third procedure (termed weighted post-sampling) in which. All units are maintained in the sample, but they are assigned different weights.

Indeed, in the hard-core and in the flexible case some units are discarded randomly so that we can describe the procedure as the random assignment of dichotomous weights that can only assume the value of 0 or 1. In contrast, in the weighted procedure, the weighting scheme allows deviations from the proportionality of the ideal sampling design.

Although ideally the weighted procedure is preferable, the identification of the weights is problematic because the theory has not been developed ye. For this reason in the present context we will employ the flexible approach which constitute a reasonable compromise.

The flexible post-sampling adopted in this paper extends the proposal of Arbia et al. (2018) to the case of spatial regression. In the remained of the paper we will show how the suggested mechanism improves the quality of the estimates with respect to the standard estimation methods which do not consider any correction.

### A Monte Carlo Assessment of the Effects of the Correction Mechanism in the Estimation of Spatial Regression Parameters

#### Simulation Setup

In this section we report the results of a set of Monte Carlo experiments designed to simulate a situation where we have to tackle the distortions induced by Variety on spatial modelling. In particular, we want to investigate the use of spatially data collected following a convenience criterion (such as web-scraped or crowdsourced data) to estimate the parameters of a spatial regression model and to verify the effects of the proposed correction mechanism described in A General Correction Procedure for Data Gathered Through Convenience Sampling section. The case study which we will consider in A Case Study: Estimating a Spatial Price Model With Web-Scraped Data in Milan (Italy) section refers to hedonic house prices where both the presence of geographical externalities (Xiao, 2017) and the likely presence of spatial correlation among prices (Can, 1992) suggest the use of a spatial econometric formulation (Anselin and Lozano-Grazia, 2009; Anselin and LeGallo 2006; Anselin et al., 2010). This motivates, our choice of the following Spatial Lag Model (SLM, see Arbia, 2014) in our simulations
In equation (1) \( y \) is an \( n \)-by-1 vector of observations of the dependent variable \( y \), \( x \) is an \( n \)-by-1 vector of observations of the independent variable (for the sake of simplicity we consider only one predictor in our model), \( \varepsilon \sim i.i.d. N(0, 1) \) are the independent innovations and \( \beta, \rho \) and \( \sigma^2 \) are scalar parameters. Furthermore, in equation (1) \( W \) is an exogenously specified weight matrix taking care of the links of proximity between the \( n \) units. In particular, the spatial weights are derived using the distance-threshold neighbouring criterion (Arbia et al., 2021), by selecting a threshold that guarantees the absence of isolated points.

In our Monte Carlo experiment, we consider two different population sizes, say, \( N = 5600 \) points (Simulation 1) and \( N = 11200 \) (Simulation 2), and we generate the points in such a way that they are randomly distributed in the two-dimensional space with different densities in the four quadrants of a unitary square (Diggle and Milne 1983). We further assume that, in both instances, only \( n < N \) of such observations are available following a convenience sampling. In all our experiments we set the sample size \( n = 270 \) (See Table 1 and Figure 1 for the case of \( N = 5600 \) ) thus considering two cases of different sample proportion corresponding to 4.8% (simulation 1) and 2.4% (Simulation 2) respectively.

In each replication of the Monte Carlo experiments, the vector \( y \) is then generated through the following expression

\[
y = (I - \rho W)^{-1}\beta x + \sigma^2 (I - \rho W)^{-1}\varepsilon
\]

where we simulate the vector of observations of the exogenous regressor from a Gaussian distribution with expected value equal to 10 and a unitary variance. This vector is kept constant in each simulation run, whereas, at each run, the error term \( \varepsilon \) is generated from a standardized normal distribution. Furthermore, the parameter \( \rho \) can assume the following values: \( \rho = [0; 0.2; 0.4; 0.6; 0.8] \) and the parameters \( \beta \) and \( \sigma^2 \) are kept constant. In a similar fashion, we simulated the effects of the proposed methodology also with the Spatial Error Model (SEM, see Arbia, 2014) defined as

| Sample proportion (simulation 1) | Q1  | Q2  | Q3  | Q4  | Total |
|----------------------------------|-----|-----|-----|-----|-------|
| N (simulation 1)                 | 2000| 200 | 1000| 2400| 5600  |
| N (simulation 2)                 | 4000| 400 | 2000| 4800| 11,200|
| n (Sim 1–2)                     | 70  | 20  | 150 | 30  | 270   |

Table 1. Population size (\( N \)), sample size (\( n \)) and sample proportion in the four quadrants reported in Figure 1 in the two simulations.
\[ y = \beta x + \lambda W u + \epsilon \]  

where \( y \) is an \( n \)-by-1 vector of observations of the dependent variable \( y \), \( x \) is an \( n \)-by-1 vector of observations of the independent variable, \( \epsilon \sim i.i.d. N(0,1) \) are the independent innovations and \( \beta \) and \( \lambda \) are scalar parameters.

**Effects of the Correction Mechanism**

In our Monte Carlo experiment, in each replication a subset of the dataset is selected randomly. Given the results of the previous section we will limit the analysis to the Spatial Lag Model reported in equation (1).
The problem of simulating a convenience sampling is not simple given the intrinsic lack of regularity which is typical of this data collection mechanism. To mimic a process of convenience sampling related to the Variety of data, we considered a marked departure from a scheme of geographically random stratified sampling, and we assumed the data to be observed less frequently in the more densely populated quadrants and comparatively more frequently in the low-density quadrants. More specifically, in Simulation 1, we considered the following sampling proportions in the four quadrats (see Figure 1): Q1 = 3.5% of units, Q2 = 10% units, Q3 = 15% and Q4 = 1.25%. In Simulation 2, the proportions are modified as follows: Q1 = 1.75% of units, Q2 5% of units, Q3 = 7.5% of units and Q4 = 0.625% of units. Our process mimics an imitation effect among the voluntary collectors in a crowdsourcing experience, as it is observed, e. g., in Arbia et al. (2018). Alternatively, it can be seen as a sample selection bias linked to some geographical feature inherent to the web-scraping process.

The aim of our Monte Carlo experiment is to monitor the sensitivity of the performances of the various estimators to the proportion of deleted observations described by the correction parameter $\zeta$ (see A General Correction Procedure for Data Gathered Through Convenience Sampling section) both in terms of bias and of the estimation variance. To this aim, first, we estimated Model (1) through Maximum Likelihood using all the 270 random sample observations without imposing any correction ($\zeta = 0$). We then introduced progressively a correction mechanism using various levels of the parameter $\zeta = (0.2; 0.4; 0.6; 0.8; 1)$ progressively deleting data so as to let the sample look like a random stratified design. As already stated, the case of $\zeta = 1$ corresponds to the hard-core post-sampling when we discard all necessary data to obtain a random sample. In this case the sample size reduces to $n = 70$. Table 2 reports the number of observations that survive after we apply the correction mechanism at each different level of the parameter $\zeta$.

Figure 2 clearly reveals that, without any correction (case of $\zeta = 0$), the ML estimates of the parameter $\beta$ in equation (1) present a marked upward bias which appears to be independent of the level of the spatial correlation parameter $\rho$ (only with a peak at $\rho = 0$ in Simulation 1). In this situation the estimation process requires a correction to restore the condition of unbiasedness that went lost with the convenience sampling. Figure 2 also shows that the (squared) bias decreases monotonically to 0 when $\zeta$.

| $\zeta$ | Sample Size |
|--------|-------------|
| 0      | 270         |
| 0.2    | 222         |
| 0.4    | 174         |
| 0.6    | 126         |
| 0.8    | 89          |
| 1      | 70          |

Table 2. Sample Size as a Function of Parameter $\zeta$ in the Monte Carlo Experiment.
increases towards 1. This behavior shows that the proposed procedure is effective in achieving the desired bias-correction mechanism. However, Figure 2 also shows that the procedure produces progressively a higher variance of the estimators which is intuitively due to the reduced number of observations available. Furthermore, Figure 2 shows that, in each simulation case, the MSE criterion identifies an optimal level of $\zeta$ which compromises between bias and estimation variance. In our experiments this value is at a level $\zeta = 0.6$ (we only observe lower levels of $\zeta$ when $\rho$ is very high). Finally, Figure 2 also allows a comparison between the performances of the estimation procedure at different sample sizes (Simulation 1 and 2) corresponding to two different sample proportions. Indeed, by comparing the results of Simulation 1 with those of Simulation 2, we observe that the correction procedure seems to be unaffected by the sample proportion. The only differences we observe in the graphs are when $\zeta = 0$, that is when no correction is considered.

To verify the impact of various specifications of the W matrix on the properties of the estimators, we repeated our Monte Carlo experiments using different definitions of the spatial weight matrices. In particular, we compared the results obtained with the threshold distance (TD) definition with two alternative definitions of neighbours, namely: a) a 4-nearest neighbours (KNN) and b) an inverse distance (ID) definition. With this experiment we also aim to investigate if the density of the W matrix can affect...
the results. In fact, the 4 nearest neighbours definition provides an example of a low-density W matrix (neglecting border effects the density in this case is equal to $4/n$), while the inverse distance definition is associated with a very high density matrix (the density in this case is equal to $(n-1)/n$). The definition based on the threshold distance can be used as an intermediate case of density.

The results obtained in this second Monte Carlo experiment are summarized in Figure 3. Our experiments show that the variance of the estimators is almost unaffected by the choice of the W matrix with only a slightly larger estimation variance observed when we employ an inverse distance definition and when the selection procedure is introduced with a low $\zeta$ ($\zeta < 0.4$). In contrast, the bias of the estimator is strongly affected by the choice of W. Indeed, again only for low values of $\zeta$ ($\zeta = 0.2$), we observe a higher bias for the KNN definition and a lower bias for the ID definition with the threshold-distance case in an intermediate position. In this respect there seems to be a relationship between bias and the sparsity of W with a higher bias associated to sparser configurations. When the correction mechanism is stronger, however ($\zeta \geq 0.4$), all three definitions lead approximately to the same results both in terms of the bias and of the variance of the estimators. Figure 3 also reveals that the minimum MSE is achieved at $\zeta$

![Figure 3](image_url)

**Figure 3.** Comparison of bias, estimation variance and MSE of the parameter $\beta$ for the SLM with three different W matrix specifications: TD = threshold distance; KNN = k-nearest neighbours ($k = 4$); ID = inverse-distance. In all cases $\rho = 0.2$. 
≥0.6 using the TD or the KNN criterion, whereas it remains constant in the interval between 0.2 and 0.6 if we employ the ID specification of W.

In summary, the inverse-distance definition produces the highest increase in the estimation variance, but also an immediate reduction of the bias as soon as the correction mechanism is introduced. In terms of the MSE, inverse-distance connectivity matrices must be preferred at low values of $\zeta < 0.6$, while, for higher values, the other two specifications guarantee a lower MSE. Figure 4 reports the result of the simulation for the SEM model. The simulation results obtained with this model are consistent with those previously presented, providing further confirmation of the validity and reliability of the findings.

It should be noted that, in comparing the results obtained with different W matrices, a problem of miss-specification may emerge. Indeed, when data are discarded in the correction process, the W matrix needs to be respecified and it could happen that data generated by a certain matrix (say $W_1$) are eventually modelled by a different matrix (say $W_2$) simply because the reduction of the number of observations is too large. Under this point of view the KNN specification seems to guarantee more robust results for high values of $\zeta$, because its structure remains unchanged (both in terms of its density and of its single entries) regardless the number of observations considered.

**An Operational MSE-Correction Procedure**

Although in the previous section we considered quite general simulation cases, in principle, we might expect that in different empirical situations the level of $\zeta$ which optimally compromises between bias and estimation variance could be different and this should be evaluated prior the estimation process.

To make the proposed procedure operational and to identify such an optimal value in practical cases we can adopt the following strategy.

![Figure 4](image.png)

**Figure 4.** Comparison of bias, estimation variance and MSE of the parameter $\beta$ for the SEM with three different W matrix specifications: TD = threshold distance; KNN = k-nearest neighbours ($k = 4$); ID = inverse-distance. In all cases $\rho = 0.2$. 

Arbia and Nardelli 215
First, let us call $\hat{\beta}_\zeta$ the maximum likelihood (ML) estimator of the parameter $\beta$ obtained using the data correction mechanism at a given level of the parameter $\zeta$. Secondly, from the simulations reported in the previous section, we have noticed the tendency of the estimation bias towards 0 when $\zeta$ goes to 1 (see Figures 2 and 3). For this reason, we can assume that the bias is minimized when $\zeta = 1$ and that an estimation of the absolute bias of the parameter $\beta$ at different levels of $\zeta$ can be expressed as $B(\hat{\beta}, \zeta) = |\hat{\beta}_\zeta - \hat{\beta}_1|$.

Once the bias is estimated, we need to estimate the mean squared error (MSE) so as to identify the optimal value of $\zeta$ which minimizes it. In general, we can express the MSE of $\hat{\beta}$ as

$$MSE(\hat{\beta}) = B(\hat{\beta})^2 + \text{Var}(\hat{\beta})$$ (4)

Thus, to estimate $MSE(\hat{\beta})$ it is necessary to have an estimation of the variance of $\hat{\beta}$. In this respect, if we consider the ML estimator of the parameters of the Spatial Lag Model reported in equation (1), we can rely on an asymptotic result (see the Appendix). Indeed, for model (1) the Hessian associated to the likelihood function has been formally derived (see Anselin, 1988; Lee, 2004) so, in any specific empirical situation, we can evaluate the Fisher information matrix and, through this, the expected asymptotic variance of the estimator of $\beta$ for any given value of $\zeta$, calculated distinctively for both the model estimated on the original dataset and for the model based on the data after the correction mechanism. Let us call these two asymptotical variances $AVar(\hat{\beta})$ and $AVar(\hat{\beta})$, respectively, using an obvious symbolism.

In summary, making use of the results reported in the Appendix, we can identify the optimal level of the correction mechanism employing the following operational steps.

**STEP 1.** First, we estimate the regression parameter $\beta$ of the Spatial Lag Model in equation (1), using MLE with different values of $\zeta$. Let us refer to this set of estimates as $\hat{\beta}_\zeta, \zeta \in [0; 1]$.

**STEP 2.** Secondly, we estimate the absolute bias associated to each level of the parameter $\zeta$ as $B(\hat{\beta}, \zeta) = |\hat{\beta}_\zeta - \hat{\beta}_1|$ assuming that for $\zeta = 1$ the upward bias due to the convenience collection is minimized.

**STEP 3.** Thirdly, for each level of $\zeta$, we estimate the asymptotic variance of $\hat{\beta}$ using equation (13) in the Appendix. We call it $AVar(\hat{\beta})$.

**STEP 4.** Fourthly, we estimate the MSE of the estimator of $\beta$ for each level of the parameter $\zeta$ by using the expression:

$$MSE(\hat{\beta}, \zeta) = (\hat{\beta}_\zeta - \hat{\beta}_1)^2 + AVar(\hat{\beta})$$ (5)

using the results obtained in STEP 2 and STEP 3.

**STEP 5.** We then identify the optimal level of the parameter $\zeta$ (say $\tilde{\zeta}$) as:
\[ \zeta = \arg\min_{\zeta \in [0,1]} \{MSE(\beta, \zeta)\} \] (6)

**STEP 6.** Finally, we pick the optimal \( \zeta \) as our final estimate and we estimate the parameter \( \beta \) using the corresponding MLE, say \( \hat{\beta}_\zeta \).

The practical use of the procedure described in this section will be illustrated in the next section making use of some real house price data scraped from the web.

**A Case Study: Estimating a Spatial Price Model With Web-Scraped Data in Milan (Italy)**

In the Monte Carlo study reported in *A Monte Carlo Assessment of the Effects of the Correction Mechanism in the Estimation of Spatial Regression Parameters* section, the optimal value of \( \zeta \) to be used in the correction mechanism, is obtained with a computational search of the value that minimizes the MSE of the estimators. In this section we aim at showing how this procedure can be made operational in the case of real data.

In particular, we will employ a flexible post-sampling strategy to estimate a house price spatial lag model. In our case study, we will estimate our model using real data that were scraped from the web by visiting the real estate company advertisements appearing on the web in the city of Milan (Italy) in May 2019. House prices represent good examples of spatial models because houses that are close in the geographic space are likely to display similar attributes. Indeed, the house market price is given by a series of characteristics linked to the residential structure (such as its style, lot size, and the number of rooms) and to the externalities associated with the geographic location (Xiao, 2017). Furthermore, the existence of geographical submarkets leads to the definition of spatial models which can account for the spatial dependence among the observations. One further reason why house prices may be spatially autocorrelated is that property values in the same neighborhood capitalize shared location amenities. All of these effects motivate the application of a spatial econometrics model.

To gather our working dataset, we performed a systematic web-scraping of one real estate company in Milan for the whole month of May 2019. The operation of extracting the relevant information from the internet is not an easy task because most of the data available on the websites are presented in an unstructured format. The complexity is also related to factors that depend on how the website was designed, on how it works and on the particular technologies used. In particular, to analyze the supply side prices in the real estate market in Milan, we used a Python web-scraping algorithm to save the source code of the web page into the HTML format, to extract the useful information parsed through the XPath language and, finally, to store the structured dataset thus obtained after each round of scrape in a relational database management system (RDMS). The rest of the analysis is performed using the commands included in the R packages spdep and spatialreg. After the systematic web-scraping of one real estate company in Milan for the whole month of May 2019, our final dataset is composed by 1000 data points. The dataset refers to individual houses advertised for which we avail
information about the geographical coordinates, the requested selling price as it appears from the advertisements and a series of other house characteristics such as their size, the number of rooms and bathrooms, the typology of the building and the presence of amenities in the neighbourhood. Figure 5 shows the geographical location of the individual points of observation and clearly displays a marked center-periphery trend of the prices with the more expansive houses concentrated in the historical area located in the city center. This empirical evidence further supports the choice of a Spatial Lag Model.

As stated in the introduction, however, this dataset suffers from the lack of a statistical sample design in the data collection process. Indeed, the dataset is just constituted by the convenience collection of the last 1000 listing available on the real estate companies’ websites without any consideration of their spatial distribution nor of the fulfilment of the requirements of a formal sample design.

To reduce the distortions that may derive from the convenience sampling, a geographically stratified sample scheme is superimposed on the map, with the stratification units represented by the 88 neighborhoods (termed Nuclei di Identità Locali, NIL, see Figure 3) and with the number of families used as a proxy of the variable of interest and employed as a stratification variable. This auxiliary information is an open datum made available through the parcellation of the cadastral maps

Figure 5. Map of the coordinates of the houses scraped from the real estate company websites in Milan (Italy) in May 2019. The partition refers to the 88 neighborhood of Milan (Nuclei di Identità Locali, NIL).
publicly available by Milan municipality. Following the procedure described in the previous *An Operational MSE-Correction Procedure* section, a SLM is the estimated via MLE using various levels of the parameter $\zeta = [0; 1]$.

In our empirical case, we postulate a simple spatial regression model which predicts the price of a house as a function of only one independent variable (namely the size expressed in square meters) with an additional spatial lag component that is intended to capture the locational effects.

The estimation results based on the six steps procedure described in *An Operational MSE-Correction Procedure* section are reported in Table 3. In all cases reported, both the estimation of $\beta$ and $\rho$ were found significant at the 5% level.

In the absence of the correction mechanism the slope is 4230.29 which is an estimation of the unitary price per squared meters expressed in Euros. This value raises up to 4406.36 in the case of full application of the method when the bias is reduced to a minimum although at the expenses of reduced efficiency. Table 3 shows that the trade-off between bias and estimation variance in estimating the regression slope $\beta$ is obtained at the value $\zeta = 0.4$ which minimizes the MSE of the estimator. At this MSE-optimal level of the correction mechanism the relative bias goes down to 1.9% starting from a level of 3.9% when no correction is considered, and all 1000 observations are included in the estimation procedure. Finally, Table 3 also displays a trend in the estimation of the parameter $\rho$ when we increase the level of correction, with values that decrease from 0.96 to 0.85 when $\zeta$ goes from 0 to 1. Finally, we compute the spatial impacts associated to our model at different values of the $\zeta$ parameter (Table 4). This is a common practice in the literature to provide a correct interpretation of the parameters of a SLM model. For example, Anselin and LeGallo (2006) evaluate the impacts for investigating the sensitivity of hedonic models of house prices to the spatial interpolation of measures of air quality. The theory on spatial impact is discussed e. g. in LeSage and Pace (2009).

It is noteworthy that, while the direct impact follows the same moderate increasing trend of the estimation of $\beta$ when the parameter $\zeta$ converges to 1, the indirect impact follows exactly the inverse trend with a sharp decrease when $\zeta$ moves from 0.2 to 1. The same effect is observed in the total impact where the increase in the direct effect is more than compensated by the decrease of the indirect effect. At the optimal MSE level

| $\zeta$ | Sample Size | $\hat{\rho}$ | $\hat{\beta}$ | Relative Bias | $\text{MSE}(\hat{\beta}, \zeta)$ |
|--------|-------------|--------------|--------------|---------------|-------------------------------|
| 0      | 1000        | 0.96         | 4230.29      | 4.0%          | 39,723.64                    |
| 0.2    | 810         | 0.96         | 4236.00      | 3.9%          | 39,682.80                    |
| 0.4    | 610         | 0.92         | 4321.10      | 1.9%          | 22,713.80                    |
| 0.6    | 410         | 0.72         | 4383.30      | 0.5%          | 24,126.22                    |
| 0.8    | 250         | 0.77         | 4368.91      | 0.8%          | 38,836.56                    |
| 1      | 220         | 0.85         | 4406.36      | 0.0%          | 38,395.37                    |
(when $\zeta = 0.4$) both indirect and total impact are halved with respect to the estimation without post-sampling correction. This effect can be explained by the fact that the off-diagonal elements of the impact matrix are affected by both the value of $\beta$ and $\rho$ (LeSage and Pace, 2009). Our case study generally confirms the findings of the Monte Carlo experiments and suggests that, while conducting empirical research making use of data collected without following a proper sample design, substantive conclusions could be biased. Following our suggested strategy, we can at least minimize bias and inefficiencies of the estimates.

**Summary, Open Problems and Future Work**

In the empirical practice, in an increasing number of cases geo-coded data, potentially very useful for estimating spatial regression models, are collected from the web without following any formal sampling scheme and based only on pure convenience. Popular examples are represented by the process of crowdsourcing and web scraping. However, due to the lack of a formal sample design, in these cases parameter’s estimators are unreliable and they lose their optimality properties (see, e.g., Hansen et al., 1953).

While this phenomenon affects any statistical analysis, in this paper we focused specifically on the estimation of a spatial regression models. Examining the results of a series of Monte Carlo experiments, we have shown that if we base our empirical analysis on these web-generated sources, the ML estimator of the regression slope is upward biased. To overcome this problem, by exploiting the idea introduced in Arbia et al. (2018), we proposed to manipulate the available data prior their use in a statistical inferential context, and cancel randomly some observation so that the final dataset resembles a formal sample design. Our Monte Carlo experiments showed that the proposed procedure, while having the positive effect of reducing the upward bias of the ML estimator of the regression slope, increases the estimation variance due to the diminished degrees of freedom. An obvious solution to these contrasting results is constituted by the use of the MSE criterion so as to compromise between decreasing bias and increasing error variance. Employing a Maximum Likelihood approach, we showed how it is possible to monitor the increase in the estimation variance

| $\zeta$ | Direct Impact | Indirect Impact | Total Impact |
|---------|---------------|----------------|--------------|
| 0       | 4552.68       | 102,522.69     | 107,075.37   |
| 0.2     | 4656.45       | 111,042.87     | 115,699.32   |
| 0.4     | 4674.06       | 51,473.28      | 56,147.34    |
| 0.6     | 4774.97       | 10,940.91      | 15,715.87    |
| 0.8     | 4959.15       | 14,069.17      | 19,028.31    |
| 1       | 4832.37       | 23,931.82      | 28,764.19    |
to the correction mechanism at different levels of the correction parameter so as to identify its optimal level.

The case study reported in *A Case Study: Estimating a Spatial Price Model With Web-Scraped Data in Milan (Italy)* section focused on the estimation of a price model using data that are web-harvested from real estate company web advertisements in Milan (Italy). Through it we showed how it is possible to exploit the idea of data correction to reduce the bias in the spatial regression parameter estimates while moderating the loss in efficiency.

The search for solutions to the problems raised by non-probabilistic data collections (such as web-harvested or crowdsourced datasets) is, undoubtedly, still in its infancy especially regarding their use in spatial econometrics. In this paper we aimed at shedding light on the problem for the benefit of applied researchers, and at providing some preliminary solution being aware that the method proposed is subject to a series of limitations that will need to be addressed in the future.

An apparent limitation is constituted by the fact that, in principle, our proposed methodology requires the knowledge of a stratification variable. This hypothesis is admittedly rarely realized in practice where the use of web-generated data is exactly motivated by the lack of information about the population together with the ease of collection. However, in this paper we suggest the use of an auxiliary variable (such as, in our empirical example, the number of households) at an aggregated geographical level.

Although in this paper we suggest an operational procedure to identify a compromise between bias and estimation variance in empirical cases, it is clear that the increase in variance, which is connatural to the approach suggested in this paper, represents a second major drawback in the application of the proposed method and that solutions should be devised in order to reduce it, if not eliminate it fully. One possible way out in this respect could be the use of a weighted version of the correction mechanism that we discussed briefly in *A General Correction Procedure for Data Gathered Through Convenience Sampling* section. We believe that incorporating a multivariate model specification would be a valuable extension of our proposed calibration technique. However, at this stage, it is beyond the scope of our current work. Our methodology relies on minimizing the Mean Squared Error (MSE) of the estimated coefficients for a univariate model. Extending this to a multivariate case would require calculating the MSE for several estimated coefficients simultaneously, which is not feasible with our current formulation.

One strategy for using this methodology in the multivariate case is to minimize the MSE calculated individually for each variable. It is possible that in some cases, different MSEs derived from different variables may return similar zeta values, which could facilitate the choice of post-sampling parameter. However, it is also possible that the optimum zeta calculated for different variables could differ, making it challenging to apply our methodology to all variables simultaneously. This is an important limitation of our proposal and will be the subject of further future research work.
It is interesting to observe a parallel between the choice of the observations to drop, which is essential to the method proposed here, and the problem of designing optimal spatial samples. In this paper we suggest dropping observations using a stratified random sample, but the procedure can obviously be generalized by selecting the units following any probabilistic sampling scheme. In particular, spatial sample designs enable to take into account the similarity between observations in space (see e.g. Grafström and Tillé, 2012; Muller, 2007). The idea is that, if data are characterized by positive spatial autocorrelation, we have a preference to eliminate observations that are close in space to minimize the information loss (Arbia, 1993). Indeed, spatial sampling depends essentially on the pattern of the observed points process and on spatial autocorrelation. These effects are not considered here and are material for further work. However, some insights in this direction can be found in Arbia et al. (2022).

We also did not study the effects of our suggested procedure on the size and the power of the tests of hypothesis associated to the model. To this aim, in the absence of exact results, one possible way out could be employing a cross-validation strategy (see, e.g. Pohjankukka et al., 2017). All these issues are still open, and they constitute material for further refinements of the procedure to be developed in some future work.

Finally, while this paper is strongly focused on the third V of the Big Data (that is variety. Laney, 2001), it would be useful to test the computational feasibility of the suggested procedures also in the presence of volume of data that are a large and fast growing thus making reference also to the other two V’s that is Volume and Velocity.

Appendix: Variance MLE Estimators of the Parameter in a Spatial Lag Model Using the Correction Mechanism

In the procedure illustrated in An Operational MSE-Correction Procedure section, to identify the optimal level of the parameter ζ it is necessary to have an explicit expression for the estimation variances. In this respect we need to derive, first of all, the Hessian associated to the likelihood function of generic element $H(\theta) = \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta^T}$, $L()$ being the model’s likelihood function and $\theta$ the parameters’ vector. In the case of the Spatial Lag model considered in equation (1) we have $\theta = [\beta, \rho, \sigma^2]$. Anselin (1988) provides the exact expressions of the Hessian of a generic heteroskedastic SARAR model while Lee (2004) reports the same expressions for the homoscedastic Spatial Lag Model. From these results we have

$$H_{\beta\beta} = n\sigma^{-2}X^TX$$  \hspace{1cm} (7)

$$H_{\beta\rho} = n\sigma^{-2}X^TW(I - \rho W)^{-1}X\beta$$  \hspace{1cm} (8)

$$H_{\beta\sigma^2} = 0$$  \hspace{1cm} (9)
\[
H_{\beta\beta} = (n^{-1} \text{tr}(W(I - \rho W)^{-1})^2 + n^{-1} \text{tr}(W(I - \rho W)^{-1})^T (W(I - \rho W)^{-1})
\]
\[
+ n^{-1} \sigma^{-2}(W(I - \rho W)^{-1} X \beta)^T (W(I - \rho W)^{-1} X \beta)
\]

(10)

\[
H_{\sigma^2 \sigma^2} = n^{-1} \sigma^{-2} \text{tr}(W(I - \rho W)^{-1})
\]

(11)

\[
H_{\sigma^2 \sigma^2} = \frac{1}{2} \text{tr}(\Omega^{-2}) = \frac{1}{2 \sigma^4}
\]

(12)

These values can be ordered in the Hessian matrix

\[
H(\theta) = \begin{bmatrix}
H_{\beta\beta} & H_{\beta\rho} & 0 \\
H_{\beta\rho} & H_{\rho^2} & H_{\rho \sigma^2} \\
0 & H_{\rho \sigma^2} & H_{\sigma^2 \sigma^2}
\end{bmatrix}
\]

(13)

In principle the Fisher information matrix can be calculated as \(I(\theta) = -E[H(\theta)]\) and the parameters’ covariance matrix as

\[
I(\theta)^{-1} = -E[H(\theta)]^{-1}
\]

(14)

However, since the above expected value cannot be calculated in closed form in most practical cases because the elements of \(H(\theta)\) are complicated non-linear functions of the data, we can estimate the elements of the covariance matrix in each empirical case by substituting in equations (6)–(11) the vector of the ML estimators \(\hat{\theta} = [\hat{\beta}, \hat{\rho}, \hat{\sigma}^2]\) (see Greene, 2018), thus obtaining the estimated Fisher information matrix

\[
\hat{H}(\theta) = \begin{bmatrix}
H_{\beta\beta} & H_{\beta\rho} & 0 \\
H_{\beta\rho} & H_{\rho^2} & H_{\rho \sigma^2} \\
0 & H_{\rho \sigma^2} & H_{\sigma^2 \sigma^2}
\end{bmatrix}
\]

(15)

We can then calculate the empirical counterpart of the estimators’ asymptotic covariance matrix as

\[
\hat{I}(\theta)^{-1} = -[\hat{H}(\theta)]^{-1}
\]

(16)

Consequently, the asymptotic variances of the various parameters’ estimators, say \(\text{AVar}(\hat{\theta})\), are the diagonal elements of the matrix \(\hat{I}(\theta)^{-1}\).

In particular, we can use equation (15), to calculate the asymptotic variance of \(\hat{\beta}\) using the subset of data with correction parameter, say \(\text{AVar}(\hat{\beta})\). This expression can then be used in equation (3) to identify the optimal level of \(\zeta\) which minimizes the MSE.
Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Vincenzo Nardelli https://orcid.org/0000-0002-7215-7934

Note
1. R codes are available at https://github.com/vincnardelli/spatial-web.

References
Anselin, L. 1988. Spatial Econometric. Kluwer Academic Publisher.
Anselin, L., and J. LeGallo. 2006. “Interpolation of Air Quality Measures in Hedonic House Price Models: Spatial Aspects.” Spatial Economic Analysis 1 (1): 31-52.
Anselin, L., and N. Lozano-Grazia. 2009. “Spatial Hedonic Models” Palgrave Handbook of Econometrics, 1213-1250. https://doi.org/10.1057/9780230244405_26
Anselin, L., N. Lozano-Grazia, U Deichmann, and S. Lall. 2010. “Valuing Access to Water—A Spatial Hedonic Approach, with an Application to Bangalore, India.” Spatial Economic Analysis 5, 2: 161-179. https://doi.org/10.1080/17421771003730703
Arbia, G. 1989. Spatial data configuration in the statistical analysis of regional economic and related problems. Kluwer Academic Publisher.
Arbia, G. (1993) The use of GIS in spatial surveys, International Statistical Review, 61 (2): 339-359.
Arbia, G. 2014. A primer for spatial econometrics. Palgrave-MacMillen.
Arbia, G. 2021. Statistics, new empiricism and society in the era of Big Data. Springerbrief, Springer-Verlag.
Arbia, G., and F. Petrarca. 2011. “Effects of MAUP on spatial econometrics models (with Francesca Petrарca).” Letters in Spatial and Resource Sciences 4: 173-185.
Arbia, G., G. Espa, and D. Giuliani. 2021. Spatial microeconometrics. Routledge.
Arbia, G., G. Solano-Hermosilla, F. Micale, V. Nardelli, and G. Genovese (2018) Post- sampling crowdsourced data to allow reliable statistical inference: the case of food price indices in Nigeria. paper presented at the LX conference of the Italian Statistical Society, Palermo, June, 2018.
Arbia, G., Y Matsuda, and J. Wu. 2022. “Estimating spatial regression models with sample data-points: A Gibbs sampler solution.” Spatial Statistics 47: 100568. online first.
Ayouba, K., M.-L. Breuille, C. Grivault, and J. Le Gallo. 2020. “Does Airbnb disrupt the private rental market? An empirical analysis for French cities.” *International Regional Science Review* 43 (1–2): 76-104.

Bergelewicz, M. 2015. “On the Representativeness of Internet Data Sources for the Real Estate Market in Poland.” *Austrian Journal of Statistics* 44 (2): 45-57.

Boeing, G., and P. Waddell. 2017. “New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings.” *Journal of Planning Education and Research* 37 (4): 457-476.

Can, A. 1992. “Specification and Estimation of Hedonic Housing Price Models.” *Regional Science and Urban Economics* 22 (3): 453-474.

Cavallo, A., and R. Rigobon. 2016. “The Billion Prices Project: Using Online Prices for Measurement and Research.” *Journal of Economic Perspectives* 30 (2): 151-78.

Chen, Y., P. Li, and C. Wu. 2020. “Doubly Robust Inference With Nonprobability Survey Samples.” *Journal of the American Statistical Association* 115 (532): 2011-2021. https://doi.org/10.1080/01621459.2019.1677241

Crequit, P.P., G. Mansouri, M. Benchoufi, A. Vivot, and P. Ravaud. 2018. “Mapping of Crowdsourcing in Health: Systematic Review.” *Journal of Medical Internet Research* 20: 5. https://doi.org/10.2196/jmir.9330

Deville, J. C., and C. E. Särndal. 1992. “Calibration estimators in survey sampling.” *Journal of the American Statistical Association* 87 (418): 376-382. https://doi.org/10.1080/17421770600661337

Diggle, Peter J., and Robin K. Milne. 1983. “Bivariate Cox processes: some models for bivariate spatial point patterns.” *Journal of the Royal Statistical Society: Series B (Methodological)* 45.1: 11-21.

Elliott, M. R., and R. Valliant. 2017. “Inference for Nonprobability Samples.” *Statistical Science* 32 (2): 249-264. https://doi.org/10.1214/16-STS598

Fisher, R. 1935. *The Design of Experiments*. New York: Hafner Publishing Company.

Fondeur, Y., and F. Karamé. 2013. “Can Google Data Help Predict French Youth Unemployment?” *Economic Modelling* 30: 117-125.

Fricker, R. D., and M. Schonlau. 2002. “Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature.” *Field Methods* 14 (4): 347-367.

Gehlke, C. E., and K. Bihel. 1934. “Certain Effects of Grouping Upon the Size of Correlation Coefficient in Census Tracks Material.” *Journal of the American Statistical Association* 29: 169-170.

Goodchild, M. 2007. “Citizens as Sensors: The World of Volunteered Geography.” *GeoJournal* 69 (4): 211-221.

Grafström, A., and Y. Tillé. 2012. “Doubly Balanced Spatial Sampling With Spreading and Restitution of Auxiliary totals.” *Environmetrics* 24 (2): 120-131.

Greene, W. 2018. *Econometric analysis*. 8th edition. Pearson.

Hansen, M. H., W. N. Hurwitz, and W. G. Madow. 1953. *Sample Survey Methods and Theory*. New York: Wiley.

Holt, D., and T. M. F. Smith. 1979. “Post Stratification.” *Post Stratification, Series A* 142 (1): 33-46.
Kim, J. K., and S. M. Tam. 2020. “Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference.” *International Statistical Review* 89: 382-401.

Kim, J. K., and Z. Wang. 2019. “Sampling Techniques for Big Data Analysis.” *International Statistical Review* 87 (S1): S177-S191. https://doi.org/10.1111/insr.12290

Kulkarni, R., K. E. Haynes, R. R. Stough, and J. H. P. Paelinck. 2009. Forecasting Housing Prices with Google Econometrics. GMU School of Public Policy Research Paper No. 2009-10.

Laney, D. 2001. “3-D data Management: Controlling Data Volume, Velocity and Variety.” *Application Delivery Strategies by META Group Inc*, 949. (2001, February 6). Downloadable at the website. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lee, L.-F. 2004. “Asymptotic Distribution of Quasi Maximum Likelihood Estimators for Spatial Autoregressive Models.” *Econometrica* 72 (6): 1899-1925.

LeSage, J., and R. K. Pace. 2009. *Introduction to spatial econometrics*. Chapman and Hall/CRC.

Little, R. J. A. 1993. “Post-Stratification: A Modeler’s Perspective.” *Journal of the American Statistical Association* 88 (423): 1001-1012. https://doi.org/10.1080/01621459.1993.10476368

Loveridge, S., and D. Paredes. 2018. “Are Rural Costs of Living Lower? Evidence from a Big Mac Index Approach.” *International Regional Science Review* 41 (3): 364-382.

Muller, W. G. 2007. *Collecting Spatial Data*. Springer.

Nardelli, V. 2019. *Spatial hedonic house price modelling using web data*. Master thesis. Catholic University of the Sacred Heart.

Pohjankukka, J., T. Pahikkala, P. Nevalainen, and J Heikkonen. 2017. “Estimating the Prediction Performance of Spatial Models via Spatial k-Fold Cross Validation.” *International Journal of Geographical Information Science* 31 (10): 2001-2019. https://doi.org/10.1080/13658816.2017.1346255

Poincare, J-H. 1908. *Science and method*. London: T. Nelson.

Wrede, M. 2022. “How Short-Term Rentals are Changing the Neighborhood.” *International Regional Science Review* 45 (4): 417-443.

Wei, Y., and Y. Cao. 2017. “Forecasting House Prices Using Dynamic Model Averaging Approach: Evidence from China.” *Economic Modelling* 61: 147-155.

Xiao, Y. 2017. *Urban Morphology and Housing Market*. Springer Geography.

Yang, S., J. K. Kim, and R. Song. 2020. “Doubly Robust Inference When Combining Probability and Non-Probability Samples With High Dimensional Data.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (2): 445-465.

Zoričák, M., P. Gnip, P. Drotár, and V. Gazda. 2020. “Bankruptcy Prediction for Small- and Medium-Sized Companies Using Severely Imbalanced Datasets.” *Economic Modelling* 84: 165-176.