3-2022

Locational Error in the Estimation of Regional Discrete Choice Models Using Distance as a Regressor

Giuseppe Arbia
Catholic University of Sacred Heart, giuseppearbia13@gmail.com

Paolo Berta

Carrie B. Dolan
William & Mary, cbdolan@wm.edu

Recommended Citation

Arbia, Giuseppe; Berta, Paolo; and Dolan, Carrie B., Locational Error in the Estimation of Regional Discrete Choice Models Using Distance as a Regressor (2022). *The Annals of Regional Science*. https://doi.org/10.1007/s00168-022-01116-y

The Annals of Regional Science
https://doi.org/10.1007/s00168-022-01116-y

ORIGINAL PAPER

Locational error in the estimation of regional discrete choice models using distance as a regressor

Giuseppe Arbia[1] · **Paolo Berta**[2] · **Carrie B. Dolan**[3]

Received: 11 May 2020 / Accepted: 19 January 2022

Abstract
In many microeconometric studies distance from a relevant point of interest (such as a hospital) is often used as a predictor in a regression framework. Confidentiality rules, often, require to geo-mask spatial micro-data, reducing the quality of such relevant information and distorting inference on models' parameters. This paper extends previous literature, extending the classical results on the measurement error in a linear regression model to the case of hospital choice, showing that in a discrete choice model the higher is the distortion produced by the geo-masking, the higher will be the downward bias in absolute value toward zero of the coefficient associated to the distance in the models. Monte Carlo simulations allow us to provide evidence of theoretical hypothesis. Results can be used by the data producers to choose the optimal value of the parameters of geo-masking preserving confidentiality, not destroying the statistical information.

JEL Classification C01 · C13 · C31

1 Introduction

In recent years we have observed an increasing interest in the use of individual data in regional economic studies. However, microeconometric studies typically suffer from several types of inaccuracies that are not present when dealing with the classical regional econometrics models which make use of aggregated data within regional partitions. Among the many forms of inaccuracy, missing data, locational

✉ Paolo Berta
paolo.berta@unimib.it

[1] Department of Statistics, Catholic University of Sacred Heart, Rome, Italy

[2] Department of Statistics and Quantitative Methods, University of Milan-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milan, Italy

[3] Department of Kinesiology and Health Sciences, William and Mary, Williamsburg, VA, USA

Published online: 09 March 2022

errors, sampling without a formal sample design, measurement errors and misalignment are the most common sources of errors that can affect the results and bias the conclusions in many different ways. A recent strand of the literature, in particular, concentrated on the distorting effects produced by locational errors on econometric modeling.

Arbia et al. (2015) analyzed intentional and unintentional measurement errors associated with the geo-masking of spatial micro–data. In some empirical circumstances, the exact geographical position of the individuals can be uncertain due to lack of information. This happens, for instance, when we have a list of firms in a small area (like e.g., a census tract), but we don't know their exact address within the area. A typical case is when we use GPS position of individuals derived from cell phone information where the position is known depending on the corresponding accuracy of the phone GPS. In this case, it is common to assign the individual to the centroid of each area, but this procedure obviously generates a locational error. Arbia et al. (2015) referred to this situation as to *unintentional locational error*. Deardon et al. (2012) found that when modeling the spread of disease at the individual level unintentional locational error can be accounted for through a random effects model. However, the biasing effect on the basic reproductive number is not easily alleviated. In other empirical cases, the individual's position is perfectly known, but they are geo-masked a priori before being made it publicly available to the analysts, in order to preserve confidentiality. In this second case we say that an *intentional positional error* was introduced. In their paper, Arbia et al. (2015) examined the instability of the results induced by locational errors in spatial regression models that include a distance as a predictor. Colenutt (1968) demonstrated how error can accumulate so that accurate prediction becomes challenging.

In regional economics it is a common practice to use spatial econometrics models which include distances as regressors. In healthcare the adoption of econometrics techniques in the study of the patients choices produced a large number of papers in which the distance between patients and hospitals is one of the most important predictors. Perucca et al. (2019) used spatial error models to examine spatial inequalities in access to care using distance as a key measure of patient mobility. Both unintentional locational errors (that is inaccuracy in the data collection due to approximate address or lack of sufficient information when individuals are located at the centroid of a small area) and intentional locational errors (a-posteriori geo-masking of the exact GPS coordinates to preserve confidentiality), are potential sources of errors that may undermine the results and mislead the substantive conclusions.

While in Arbia et al. (2015) the authors examined the effects of intentional locational errors induced by geo-masking in the case of continuous linear regression models when distance is used as a regressor, in this paper we aim to extend these results to the case of nonlinear models. We aim to shed light on the distortion effects due to locational error to make researchers aware of the possible limitations of their inferential conclusions. In contrast with the case of a continuous linear regression model, in the case of a nonlinear model, the ML estimators do not have closed form solutions and they are usually derived by numerical maximization. As a consequence, in this paper no formal result can be obtained and we have to resort to a Monte Carlo (MC) approach.

The rest of the paper is structured as follows. In Sect. 2 we summarize the main results of Arbia et al. (2015) and we present a motivating case study based on (Berta et al. 2016). Section 3 describes the effects of the Maximum Likelihood (ML) estimators on bias and efficiency, whereas Sect. 4 present the results of several MC experiments on spatial models affected by intentional locational errors. Finally, Sect. 5 concludes the paper.

2 Effects of geo-masking in the regression analysis of healthcare competition

The theoretical motivation of this paper refers to Arbia et al. (2015), where the authors study the negative effects of the geo-masking, examining the measurement error introduced by geo-masking the individuals' true location, when distances are used as predictors in a linear regression. A very popular geo-masking mechanism is the uniform geo-masking (explained, e.g.,, in Burgert et al. 2013), in which the true coordinates are transformed by displacing the individuals' position along a random angle (say θ^*) and a random distance (say δ) both following a uniform probability law. The mechanism can be formally expressed through the following hypotheses:

$\mathbf{H_0}$ $\theta \sim U(0, \theta^*)$ and $\delta \sim U(0°, 360°)$, with θ^* the maximum distance error;

$\mathbf{H_1}$ θ and δ are independent.

Fig. 1 Random point displacement, Urban Malawi DHS Clusters

An example of uniformly geo-masked locations is illustrated in Fig. 1. Represented as points in Fig. 1 are groupings of urban households that participated in the 2014 Malawi Malaria Indicator Survey (NMCP and ICF International 2014). The buffer represents the maximum amount of random displacement introduced by the Demographic Health Survey (DHS) to ensure that respondent confidentiality is maintained. The true location of the respondent households are located within this 2km buffer.

An alternative geo-masking mechanism is the Gaussian geo-masking where points are randomly reallocated in the neighborhood of the true location, following a Gaussian bivariate density function with the mean vector coinciding with the true point and a given variance which can be expressed again as a function of the (practical, 99%) maximum displacement distance θ^*.

Let us now consider a simple linear regression model using a distance as a predictor:

$$y_{ih} = \alpha + \beta d_{ih}^2 + \varepsilon_{ih} \tag{1}$$

with d_{ih} the distance between point i and point h.

Considering a healthcare framework, patients living in a point are moved (geo-masking their true coordinates) within a circle with a maximum radius of, say, θ^* and randomly re-assigned in an erroneous position. In this way, since the coordinates associated to the geo-masked point are considered when calculating the distance of the patient from an hospital, we introduce a measurement error in the independent variable. Arbia et al. (2015) extended the classical error measurement theory (e.g., Verbeek 2008) to this specific case by showing that the greater the maximum displacement distance (θ^*), the larger will be both the loss in efficiency and the bias of OLS estimator of the β parameter, producing a reduction toward zero of its absolute value (known in the literature as the attenuation effect).

The loss in efficiency and the attenuation effect observed in the presence of geo-masking, are very important under a practical point of view. Figure 2 reports the theoretical behavior of the attenuation effect for Gaussian and uniform geo-masking of points as a function of the maximum displacement distance θ^* (Arbia et al. 2015). The inspection of the graph clearly shows that the attenuation increases dramatically already at small levels of θ^*, and the Gaussian geo-masking produces more severe consequences on the estimation of β than the uniform geo-masking.

A typical framework, where a measure of distance is used, is the study of competition in the healthcare sector. Hospital competition is one of the most widely studied topics in health economics and health econometrics. This research area follows the approach suggested by Kessler and McClellan (2000). In this seminal paper, the authors analyze the relationship between hospital quality and competition, modeling the patients' choices in order to build a competition index (Herfindahl-Hirschman index). This measure of hospital competition based on the predicted patients' choices is included as a predictor in a further model, where the hospital quality is the dependent variable. Following this approach, i.e.,, Berta et al. (2016) assumed that the discrete choice of the single patient i of choosing hospital h (say, y_{ih}) is related to the expected utility of the patient y_{ih}^*, with $y_{ih} = 1$ if $y_{ih}^* > 0$, and 0 otherwise.

The utility model was specified as follows:

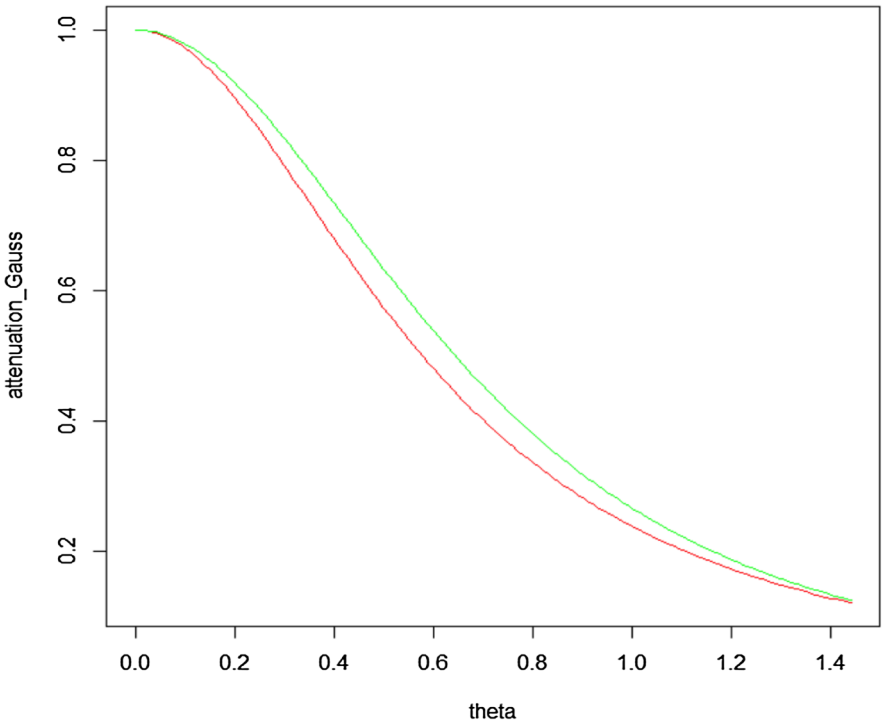

Fig. 2 Attenuation effect in the presence of geo-masking as a function of the maximum displacement distance θ^*. Gaussian geo-masking (red line). Uniform geo-masking (green line). *Source* Arbia et al. (2015)

$$y_{ih}^{*} = \rho d_{ih} + \delta_h Network_{ih} + \gamma_h GP_{ih} + \xi_h \boldsymbol{x}_{ih} \tag{2}$$

where d_{ih} is the distance between patient i and hospital h expressed in minutes of travelling, GP_{ih} is the percentage of patients living in the same zip code as patient i and sharing the GP with patient i and admitted to the hospital h, while $\boldsymbol{x}_{ih}$ is a set of patient-level characteristics. The variable $Network_{ih}$ is a continuous variable representing the share of people living in the same municipality as patient i and admitted in the same hospital h in the 12 months before the admission of patient i. Considering that the travel distance is strictly correlated with the hospital choice, the coefficients related to the distance was expected to be negative. The model was estimated using an administrative dataset related to 8627 patients admitted in the 20

cardiac surgery wards located in Lombardy (Italy) in 2014, obtaining a total number of 172,540 observations. In Berta et al. (2016), as in the majority of the papers adopting this empirical strategy, the patient location was not perfectly known, and it was approximated by the centroid of the municipality of the patient.

3 The theoretical efficiency loss of ML estimates of the parameters of a logit model

If we employ a ML strategy in the estimation process, the maximization has to be performed numerically due to the high degree of non linearity of the log-likelihood. As a consequence there is no closed form solution and so there is no possibility to derive a formal relationship of the attenuation effect as it is done in the classical error measurement theory for linear regression models. For this reason we will try to shed light on this aspect through a series of Monte Carlo experiments whose results will be reported in the next section. Before doing that, however, in this section we will examine some interesting formal results to quantify the loss efficiency associated to the geo-masking process

Let us consider the log-likelihood associated with the logit model in the case of perfect knowledge of the spatial coordinates. This can be expressed as:

$$\ell(\beta) = ln[L(\beta)] = \sum_{i=1}^{n} \left\{ y_i lnF(d_{ij}\beta) + (1 - y_i) ln[1 - F(d_{ij}\beta)] \right\} \tag{3}$$

where F is a cumulative probability distribution function to be specified (Greene 2016).

Now assume that the distance appearing in Eq. 3 is affected by a measurement error which in turn is due to a locational error introduced intentionally by geo-masking to preserve the respondent confidentiality. Let us further consider the associated, error-affected, likelihood that can be expressed as:

$$\bar{\ell}(\beta) = ln[\bar{L}(\beta)] = \sum_{i=1}^{n} \left\{ y_i lnF(\bar{d}_{ij}\beta) + (1 - y_i) ln[1 - F(\bar{d}_{ij}\beta)] \right\} \tag{4}$$

where, as before, $\bar{d}_{i,j}$ represents the error–affected distance.

Let us now assume, in particular, that the cumulative probability distribution is specified as a standardized *logistic* distribution, say Λ, characterized by 0 expected value and variance $\pi^2/3$. In this case, using Eq. 3, the associated score functions can be written as (Greene 2016):

$$\frac{\partial}{\partial \beta} \ell(\beta) = \sum_{i=1}^{n} (y_i - \Lambda_i) d_{ij} = 0 \tag{5}$$

with $\Lambda_i = \Lambda(d_{i,j}, \beta)$, and the second order derivative as:

$$\frac{\partial^2}{\partial^2 \beta^2} \ell(\beta) = \sum_{i=1}^{n} \Lambda_i (1 - \Lambda_i) d_{ij}^2 = 0 \tag{6}$$

with similar expressions for the first and the second likelihood derivatives, easy to obtain in the case of geo-masked coordinates. Using Eq. 6, we can obtain the elements of the Fisher Information Matrix related to the simulated true data as well as those related to the data after geo-masking. Indeed, from Eq. 6 we have:

$$\mathrm{Var}(\hat{\beta}) = -E\left[\frac{\partial^2}{\partial^2 \beta^2} \ell(\beta)\right] = -E\left[\sum_{i=1}^{n} \Lambda_i (1 - \Lambda_i) d_{ij}^2\right] \tag{7}$$

and similarly, for the likelihood under the geo-masked coordinates:

$$\mathrm{Var}(\hat{\bar{\beta}}) = -E\left[\frac{\partial^2}{\partial^2 \beta^2} \ell(\beta)\right] = -E\left[\sum_{i=1}^{n} \Lambda_i (1 - \Lambda_i) \bar{d}_{ij}^2\right] \tag{8}$$

Equations 7 and 8 show that the variance of the estimators depends essentially on the distance d. If this distance is inflated by geo-masking the precision of the estimators will be reduced.

To show more explicitly this Efficiency Loss (EL), making use of Eqs. 7 and 8, we can calculate the loss in the efficiency of the ML estimator as the ratio as the variance of the MLE with the geo-masked coordinates, say $\hat{\bar{\beta}}$, and the variance of the MLE with the true coordinates, say $\hat{\beta}$:

$$\mathrm{EL} = \frac{\mathrm{Var}(\hat{\beta})}{\mathrm{Var}(\hat{\bar{\beta}})} = \frac{-E\left[\sum_{i=1}^{n} \Lambda_i (1 - \Lambda_i) d_{ij}^2\right]}{-E\left[\sum_{i=1}^{n} \Lambda_i (1 - \Lambda_i) \bar{d}_{ij}^2\right]} \tag{9}$$

Now let us concentrate our attention on the denominator of this expression, where we have:

$$\mathrm{Var}(\hat{\bar{\beta}}) = E\left[\sum_{i=1}^{n} \Lambda_i (1 - \Lambda_i) \bar{d}_{ij}^2\right] = \sum_{i=1}^{n} \Lambda_i (1 - \Lambda_i) E(\bar{d}_{ij}^2) \tag{10}$$

since the elements of $\Lambda_i = \Lambda(d_{i,j}, \beta)$ are by definition non-stochastic in our case. Furthermore, Arbia et al. (2015) show that, under the hypothesis of uniform geo-masking expressed in Sect. 2, we have:

$$E(\bar{d}_{ij}^2) = d_{ij}^2 + \frac{\theta^*}{3} \tag{11}$$

(see also Arbia (2016)). Expression 10, therefore, can be re-written as:

$$\mathrm{Var}(\hat{\bar{\beta}}) = E\left[\sum_{i=1}^{n} \Lambda_i (1 - \Lambda_i)\left(d_{ij}^2 + \frac{\theta^*}{3}\right)\right] \tag{12}$$

As a consequence, using Eq. 9, the efficiency loss due to geo-masking can be expressed as:

$$\mathrm{EL} = \frac{\mathrm{Var}(\hat{\beta})}{\mathrm{Var}(\hat{\hat{\beta}})} = \frac{E\left[\sum_{i=1}^{n} \Lambda_i(1-\Lambda_i)\left(d_{ij}^2\right)\right]}{E\left[\sum_{i=1}^{n} \Lambda_i(1-\Lambda_i)\left(d_{ij}^2 + \frac{\theta^*}{3}\right)\right]}. \tag{13}$$

Because the true distances d_{ij} are deterministic, Eq. 13 clearly shows that the efficiency loss of the ML estimates of the parameter β is an inverse function of the maximum displacement distance θ^*.

4 Monte Carlo evaluation of locational errors in discrete choice models

4.1 A first simulation experiment: the effect of "unintentional" locational error when allocating the individuals in the centroids of the areas

In a first MC experiment, we aimed to quantify the existence of distortion effects in discrete model estimation, specifically in the case of unintentional locational error induced by uncertainty on individual's location. In this MC study, we considered the data used by Berta et al. (2016) for their healthcare competition study and we assumed that the individuals' location observed by the authors was the true patient location known without error. We then estimate a logit model, randomly relocating 1000 times the 8627 patients of the original dataset using a uniform geo-masking (see Sect. 2) with a maximum distance θ^* which equals the radius of a circle with the equivalent surface area of each municipality. This radius represents the (approximate) maximum location error committed when an individual is allocated to the centroid. When a point is randomly relocated outside the study area the point it is randomly relocated a second time. The 1000 simulated relocations of the patients thus define 1000 new matrices of distances between the patients and the hospitals. Using these modified distances, we estimate 1000 discrete choice models, where the dependent variable is the patients' choice and the covariate is the distance. In this way, we obtained 1000 replications of the estimates of the parameter β concerning the effect of the distance on the patients' choice.

The results are reported in Figs. 3 and 4. Figure 3 reports the kernel density for the MC distribution of the estimates of the parameter β after geo-masking. The distribution assumes a symmetric shape. In Fig. 4 the kernel density of the parameters β is compared with the β parameter estimated with the not-distorted data (the straight red line). In addition, the two dashed red lines represent the confidence interval for the original β parameter. Comparing the value obtained by the MC experiment we observed that only in the 10% of the provided simulations do not statistically differ from the original β, whereas in the 90% the absolute value is lower and approaching 0.

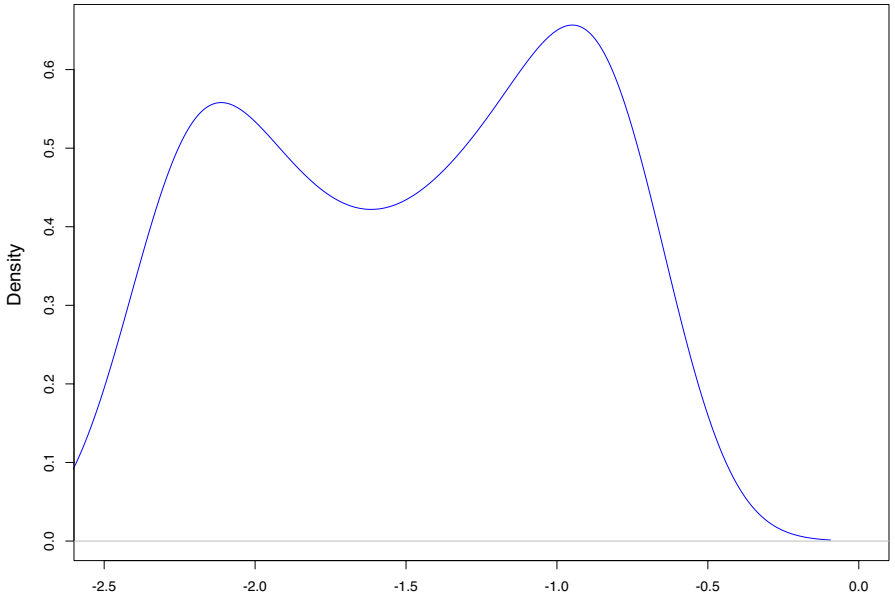

Fig. 3 MC distribution of the models' coefficients β after geo-masking

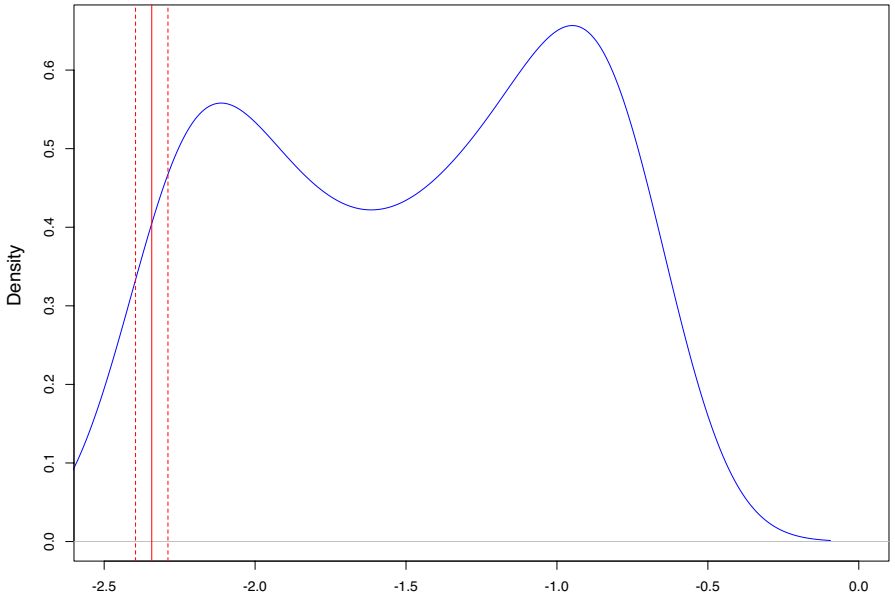

Fig. 4 MC distribution of the models' coefficients β after geo-masking compared with the β based on the "true" location

This first simulation experiment clearly shows that a locational error originated by randomly allocating the patients within the specific municipality of reference, affects significantly the estimates of the parameters of a logit model. Indeed, our results reinforce those included in Berta et al. (2016). In light of the effect reported here it is reasonable to believe that their results underestimate the spatial effect of distance in an hospital choice model.

4.2 A second simulation experiment: the effects of "intentional" locational error with different displacement distances

In the first simulation exercise the maximum displacement distance was considered fixed and (since we imposed the constraint that a patient should remain in the original municipality) it depends only on the surface area of the municipality. The larger the municipality the larger the maximum locational error that can be introduced with the geo-masking mechanism. In this second simulation experiment, we aim at a more general result by removing the constraint related to the dimension of the municipality so as to be able to explore (similarly to the study in Arbia et al. (2015) for linear models) how the distortion in the estimation is related to the maximum radius of geo-masking. This second MC exercise, therefore, tends to reproduce the case of intentional locational error. In this second instance examined, for simplicity but without loss of generality, we considered the presence of only one hospital in the study area. We thus consider the same set of individuals located in Lombardy and reported in the study by Berta et al. (2016), and the associated Euclidean distances (say d_{ih}) measured between each patient i and the hospital. In order to monitor the effects of geo-masking on the bias in the estimation of different displacement distances, we allowed the parameter θ^* to assume different values. In particular, we considered the following values: $\theta^* = 6.6(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1)$, the constant 6.6 representing the radius of the equivalent circle with a surface area equal to that of the entire Lombardy region (after coordinates standardization). For each value of θ^*, we repeat 1000 times a uniform geo-masking of the patients' coordinates. In this way, for each individual (and for each of the 10 maximum radius), we obtain 1000 new distances (say $\bar{d}_{ih}$) and, for each of these geo-masked distances, we estimate the patients' hospital choice. In particular, we assume that the observable choice of patient i of being admitted to the hospital h (y_{ih}) is related to the expected utility of i choosing h (y^*_{ih}), according to $y_{ih} = 1$ if $y^*_{ih} > 0$. As a consequence, our choice model is:

$$y_{ih} = \alpha + \beta \bar{d}_{ih} \tag{14}$$

We still expect the coefficient β to be negative, but we also expect again that the geo-masking forces the coefficients toward 0 when the radius of displacement increases. Furthermore, similarly to the continuous linear model case, this distortion depends on the maximum displacement distance θ^* adopted. The main results of this second simulation are reported in Fig. 5 which is the counter-part of Fig. 2 for the case considered in our study.

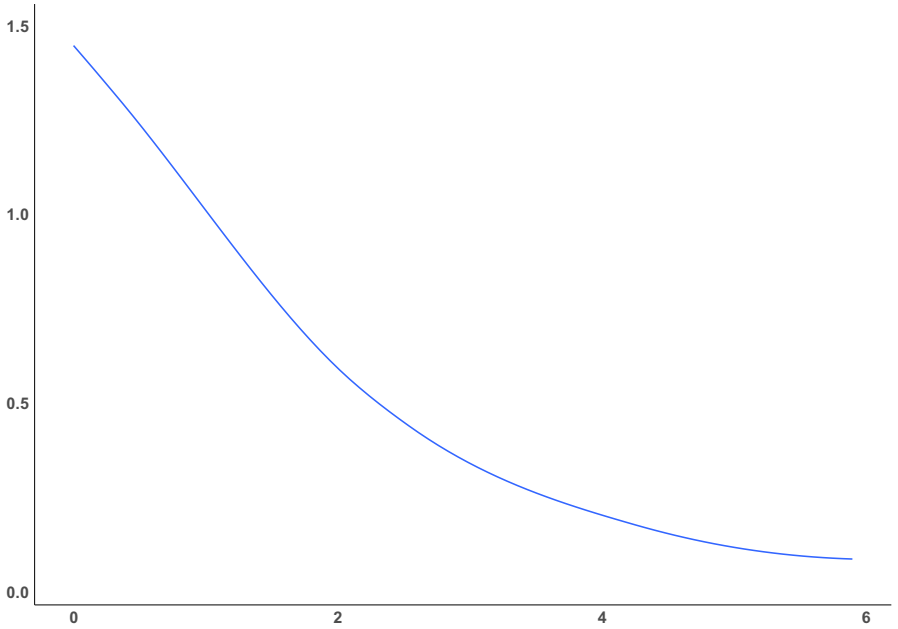

Fig. 5 Attenuation effect for 8627 patients in Lombardy affected by uniform geo-masking for increasing displacement θ^*

For increasing values of θ^* the coefficient related to the distance decreases monotonically towards 0, thus showing a similar effect to the attenuation effects observed in linear models (Arbia et al. 2015). A sharp decrease is observed already for low levels of θ^*. For instance, when $\theta^* = 2$ (corresponding to 30% of the maximum distortion distance) the estimated β is about 0.6 which is 40% of the original value. Furthermore, the 92% of the estimates provided in this simulation experiment differ from the original β. In terms of variability, the 42% of the coefficients are not significant.

4.3 Third simulation experiment: the effects of "intentional" locational error with 1000 simulated individuals

To obtain more general results then those described in Sect. 4.2, in the third MC experiment we still refer to the case of an intentionally induced locational error, but we now abandon the reference to real data and we simulate the behaviour of $n = 1000$ individuals randomly distributed in a unitary squared area centred on zero. Their distribution is reported in Fig. 6.

Using the Complete Spatial Randomness (CSR) model (Diggle 1983) the individuals' coordinates are generated as two independent uniform distributions $U(-0.5, 0.5)$. For each individual located in (i, j), we then calculate the Euclidean distance from the point (0, 0) assuming, without loss of generality, that the hospital

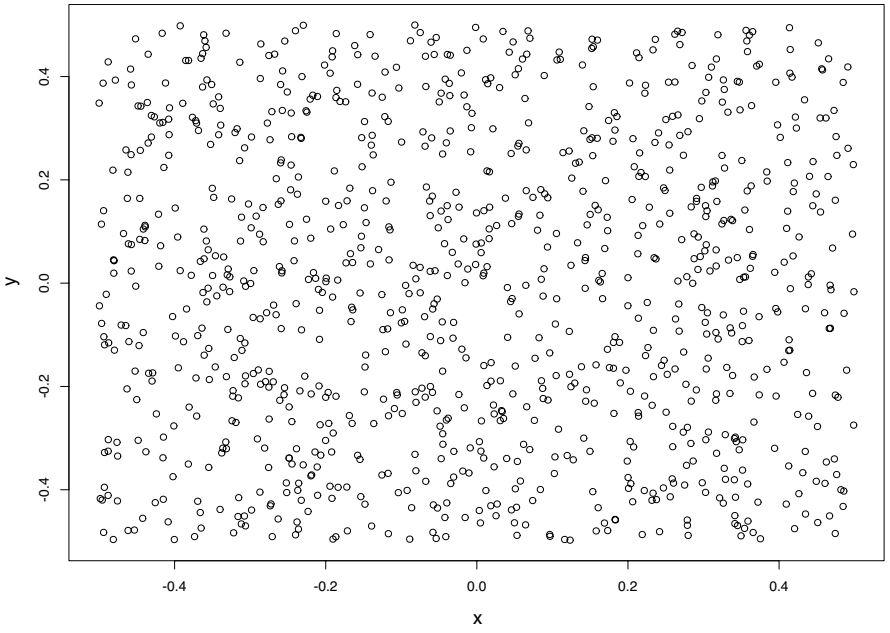

Fig. 6 1000 individuals randomly distributed (CSR) in a squared unitary study area centred in the origin

is located in the centred of the study area. We then simulate a conditional binomial choice of the hypothetical hospital located at the origin (0, 0), obtaining a simulated *logit* model, as follows:

$$logit(\pi_{ij}|d_{ij}, \boldsymbol{\varphi}) = 1 - 2 * d_{ij} \tag{15}$$

where $logit(\pi_{ij}|d_{ij}, \boldsymbol{\varphi})$ is the logit of the probability associated to a random conditional binomial distribution given the regression parameter vector $\boldsymbol{\varphi} = (1, 2)$.

Starting from this set of data we fix again 10 maximum radius of distortions such that $\theta^* = 0.707(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1)$, the constant $\sqrt{2}/2 = 0.707$ now representing half of the diagonal of the unitary square on which the data are laid.

For each of the 10 maximum θ^* we again replicate 1000 times for each individual a uniform geo-masking mechanism. In this way, for each individual and for each of the 10 levels of the maximum radius, we obtain 1000 new distances. Using the polar coordinates, the distorted distances from the origin ($\bar{d}_{ij}$) are calculated as follow:

$$\bar{d}_{ij} = \sqrt{(i - \theta\cos(\delta))^2 + (j - \theta\sin(\delta))^2} \tag{16}$$

where, as discussed in Sect. 2, θ is the distance from the true point and δ the angle of the segment joining the true point with the displaced one. For each replication, we then re-estimate the logit model in Eq. 15 using the distorted distances instead of the true ones:

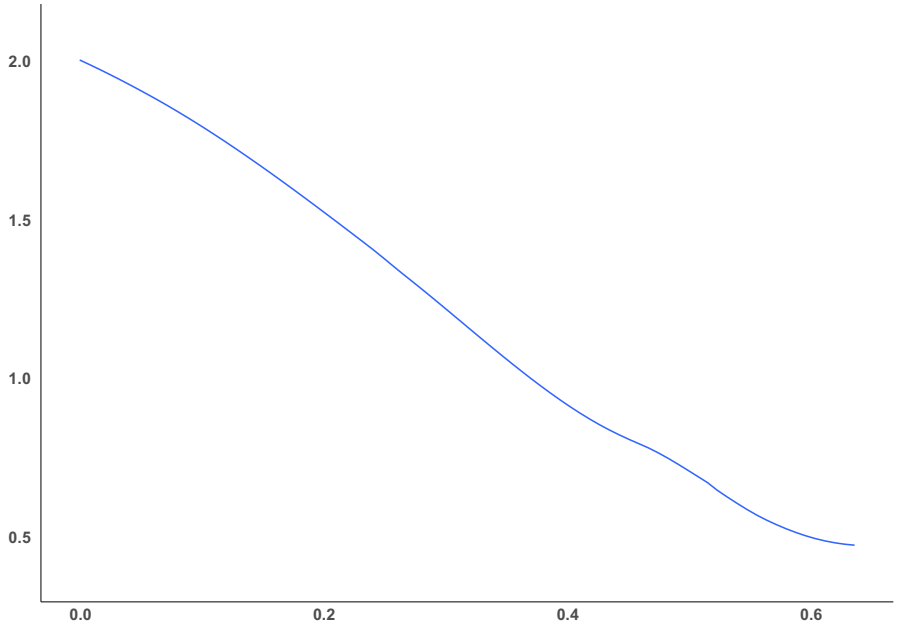

Fig. 7 Attenuation effect for 1000 simulated individuals in the presence of uniform geo-masking as a function of the maximum displacement distance θ^*

$$logit(\pi_{ij}|\bar{d}_{ij}, \boldsymbol{\varphi}) = 1 + 2 * \bar{d}_{ij} \tag{17}$$

We expect again to observe an attenuation effect reducing $\hat{\beta}$ in absolute value towards 0 as the maximum radius of distortion increases. The mean of the estimates of $\hat{\beta}$ for each θ^* can be plotted over the increasing values of θ^*, representing, once again, the attenuation effect on $\hat{\beta}$. Results are reported in Fig. 7.

As expected, for increasing values of θ^* the coefficient β decreases towards 0 thus erroneously suggesting a not-significant relationship of the hospital choice based on the distance between the hospital and the place where the patient lives. Again, the decrease is very sharp. When only a small amount of distortion is imposed (e.g. $\theta^* = 0.07 = 10\%$ of the maximum distortion) the value of β is almost unaffected, but as soon as it increases further, we observe a sharp decline. In this case, only the 22% of the estimates are different from the original β, and almost the 70% is statistically equal to 0.

5 Discussion and conclusion

In this paper, we have examined the effects of measurement error introduced in a logistic model by random geo-masking, when distances are used as predictors. We have exploited the data used in Berta et al. (2016) as case-study, where the authors studied the determinants of the patients' choice for hospital admissions.

Extending the classical results on the measurement error in a linear regression model, our MC experiment on hospital choices showed that the higher the distortion produced by the geo-masking, the higher is the downward bias in absolute value towards zero of the coefficient associated to the distance in a regression model. This effect is the discrete choice counter-part of the familiar attenuation effect well known in the literature on linear regression models (Greene 2016).

In particular, when a certain degree of locational error is produced by arbitrarily allocating individuals in the centroid of a geographical partition (as it is customary to do in many empirical studies), according to intuition, the larger is the surface area of the geographical partition the larger will be the downward bias. When data are intentionally displaced according to a random mechanism in order to protect confidentiality, we also observe a similar form of the attenuation effect.

There is a growing literature aimed to integrate information on the precision of geographic data to improve the accuracy of modeling efforts (Aerts et al. 2003). Several solutions to this problem from the efficient estimation of probability density (Lilburne and Tarantola 2009) to sensitivity models have been proposed (Lilburne and Tarantola 2009).

Two different approaches have recently been explored by the literature which seek to mitigate potential errors associated with the use of spatial data. First, a simulation and extrapolation method (Marty et al. 2019) has sought to overcome imprecision in spatial measurement - i.e.,, when the exact latitude and longitude coordinates of an intervention may not be known. It operates by intentionally introducing imprecision into known data, and then backward-extrapolating to estimate what true coefficients and standard errors would be under cases of no imprecision. While it represents an improvement relative to model averaging, relatively little research has explored the utility of SIMEX approaches using functional model forms other than ordinary regression. Building on this work, a second avenue of literature has begun to explore how to mitigate bias that may be caused due to spatial spillover between control and treatment units, as well as the arbitrary selection of a distance away from treatments that are considered “treated” (Runfola et al. 2020). This class of approach provides estimates of the distance-decay function of treatment effects, allowing policymakers to examine the geographic distance away from interventions that a treatment effect might be expected. While both of these approaches represent novel contributions, in both cases the authors note the nascent nature of the field and highlight the need for further research

In summary, the results obtained in this paper could be used by the data producers to choose an optimal value of θ^* that preserves confidentiality while not destroying the statistical information.

The MC results show that this attenuation effect increases dramatically already at small levels of distortion θ^*. Furthermore we also proved formally that the MLE of the logistic regression parameters looses efficiency when estimates are based on individuals whose position is geo-masked.

The results could be used to communicate to the practitioners the resulting level of attenuation and of efficiency loss they should expect from a logistic regression analysis.

Authors' contributions All authors contributed equally to the statistical analysis and the writing of the manuscript.

Funding None.

Availability of data and materials Simulated data available by the R-code.

Code availability R-code available upon request.

Declarations

Conflict of interest We declare no competing interests.

References

Aerts JC, Goodchild MF, Heuvelink G (2003) Accounting for spatial uncertainty in optimization with spatial decision support systems. Trans GIS 7(2):211–230

Arbia G et al (2016) Spatial econometrics: a broad view. Found Trends Econom 8(3–4):145–265

Arbia G, Espa G, Giuliani D (2015) Measurement errors arising when using distances in microeconometric modelling and the individuals' position is geo-masked for confidentiality. Econometrics 3(4):709–718

Berta P, Martini G, Moscone F, Vittadini G (2016) The association between asymmetric information, hospital competition and quality of healthcare: evidence from Italy. J R Stat Soc A Stat Soc 179(4):907–926

Burgert CR, Colston J, Roy T, Zachary B (2013) Geographic displacement procedure and georeferenced data release policy for the demographic and health surveys

Colenutt R (1968) Building linear predictive models for urban planning. Reg Stud 2(1):139–143

Deardon R, Habibzadeh B, Chung HY (2012) Spatial measurement error in infectious disease models. J Appl Stat 39(5):1139–1150

Diggle PJ (1983) Statistical analysis of spatial point processes. Academic, London

Greene WH (2016) Econometric analysis, 8th edn. New York University, Stern School of Business

Kessler DP, McClellan MB (2000) Is hospital competition socially wasteful? Q J Econ 115(2):577–615

Lilburne L, Tarantola S (2009) Sensitivity analysis of spatial models. Int J Geogr Inf Sci 23(2):151–168

Marty R, Goodman S, LeFew M, Dolan C, BenYishay A, Runfola D (2019) Assessing the causal impact of chinese aid on vegetative land cover in burundi and rwanda under conditions of spatial imprecision. Dev Eng 4:100038

NMCP and ICF International (2014) Malawi malaria indicator survey (MIS). Technical report, National Malaria Control Programme (NMCP) [Malawi] and ICF International, Lilongwe, Malawi, and Rockville, Maryland, USA

Perucca G, Piacenza M, Turati G (2019) Spatial inequality in access to healthcare: evidence from an italian alpine region. Reg Stud 53(4):478–489

Runfola D, Batra G, Anand A, Way A, Goodman S (2020) Exploring the socioeconomic co-benefits of global environment facility projects in uganda using a quasi-experimental geospatial interpolation (qgi) approach. Sustainability 12(8):3225

Verbeek M (2008) A guide to modern econometrics. John Wiley & Sons