# Dirty spatial econometrics

*Giuseppe Arbia*
*Catholic University of the Sacred Heart, Rome*

*Maria Michela Dickson, Giuseppe Espa and Diego Giuliani*
*University of Trento*

**Abstract:** Spatial data are often contaminated with a series of imperfections that reduce their quality and can dramatically distort the inferential conclusions based on spatial econometric modeling. A "clean" ideal situation considered in standard spatial econometrics textbooks is when we fit Cliff-Ord-type models to data where the spatial units constitute the full population, there are no missing data and there is no uncertainty on the spatial observations that are free from measurement and locational errors. Unfortunately in practical cases the reality is often very different and the datasets contain all sorts of imperfections: they are often based on a sample drawn from the whole population, some data are missing and they almost invariably contain both attribute and locational errors. This is a situation of "dirty" spatial econometric modelling. Through a series of Monte Carlo experiments, this paper considers the effects on spatial econometric model estimation and hypothesis testing of two specific sources of dirt, namely missing data and locational errors.

## 1. Introduction

Variables observed within territorial units are not randomly scattered and they are almost invariably characterized by spatial dependence. Statisticians have long been aware of this phenomenon and of its non-negligible impact on parameter estimation and hypothesis testing within a regression framework. Following this tradition in the last decades a wide class of spatial econometrics models have been introduced in the literature to properly accommodate for the various sources of bias and inefficiencies in the statistical inference based on spatial data. Such models almost invariably can be traced back to the autoregressive paradigm, sometimes referred to as Cliff-Ord-type models after the contribution of Cliff and Ord (1972).
A "clean" ideal situation considered in standard spatial econometrics textbooks (Anselin, 1988; Arbia, 2014) is when we fit Cliff-Ord-type models to data where: (i) the spatial units (whether they are points in space or regions) constitute a full population and not a sample, (ii) a complete cross section of territorial units is available with no missing data, (iii) variables are observed directly, (iv) there is no uncertainty on the spatial observations that are free from measurement error, and (v) the location of the observations is perfectly known. Unfortunately for spatial econometricians engaged in empirical analysis the reality is often very different and the datasets they have to fight with are made "dirty" with all sorts of imperfections: they are often based on a sample drawn from the whole population of spatial

locations, some data are missing, some variable only proxy the target variables and they almost invariably contain both attribute and locational errors.

Our aim is to show that such imperfections are not just incidental to the statistical analysis, but they can mask and hide the real phenomena up to the point of distorting dramatically the inferential conclusions.

Let us consider three paradigmatic cases.

A first example is the case of spatial health data. In a recent paper Deuchert and Wunsch (2014) analyze a set of Malawi data to assess the effectiveness of health policies to reduce infant mortality due to malaria. The observed database is constituted by a sample of a household systematically drawn from a list of enumeration areas defined in the population census. In this dataset, some of the sample units can be missing not at random for various reasons (non-response, death, cancellations, clusters of households refusing to answer). In the case of Malawi data (as in many other household surveys), field teams routinely use GPS receivers with a positional accuracy of 15 meters or less to geo-reference the location of the observational unit, but in order to preserve confidentiality the GPS coordinates are then displaced according to a "random direction, random distance" method as described, e. g., in Collins (2011).

A second example concerns plant locations in official statistics. For example, in Italy the National Statistical Institute (ISTAT) collects and disseminates data related to the active firms (ASIA archive. See e. g. Cozzi and Filipponi, 2012). At a firm level, the ASIA archive includes information concerning a set of economic variables (e. g. the firm code, the sector of activity, employees, legal status, firm's birth date and firm's termination date) together with the geographic location of plant in terms of latitude-longitude spatial coordinates. Spatial coordinates are identified automatically on the basis of the street address and so they contain a certain location error. Furthermore, for a non-negligible numbers of plants for which the address was missing, the geographic location, is approximated by the coordinates of the centroid of the municipality of the plant.

A third example is related to forestry. Forest inventories are important tools to monitor the state of the environment, to assess the quantity and quality of forestry resources, and to measure other important variables such as biomass, growth and production capacity. The inventory is often based on samples of trees distributed in the study area following a certain spatial design. For instance, the Italian National Forestry Inventory (see IFNC, 2015) collects data related to approximately 300,000 sample points, randomly located and covering the whole Italian territory. Some of these analysis involve collecting data about trees whose position is geo-masked. The reasons for masking the exact position of the observed trees are mainly connected with the need to preserve the information about the value of the trees and of their property and in order to avoid conflicts with the owners that may refuse to include their trees in the panel. For these reasons the marking of the trees included in the INFC sample are invisible (e. g. under the earth) and the information about the coordinates are disclosed only in terms of the south-west coordinate of a 1km-by-1km regular grid. As a consequence the information is geo-masked with a location error of up to 1.4 kms.

The previous examples are paradigmatic of "dirty" spatial econometric situations where both missing data and locational errors are present.

When dealing with missing data a fundamental distinction is made between data missing at random (possibly completely at random) and data missing not at random (Roderick and Rubin, 2007). In the spatial case this distinction is particularly relevant. In fact, if data are missing not at random in space they can dramatically distort the picture by erasing entire spatial patterns.

Another fundamental distinction has to be made when dealing with locational errors of the kind discussed in the examples above. In fact in some instances the uncertainty about the position is due to imperfections in the process of data acquisition (as it is in the case of the ASIA archive of firms where the location is induced from the street address) while in some other case the uncertainty is induces by the data producer to preserve confidentiality (as it is the case with the health and the forestry data). In this second case the process of geo-masking is disclosed and can be used to improve the inferential procedure.

This paper aims at making researchers aware of problems of this kind when employing spatial econometrics standard techniques. In the present work we are not suggesting statistical solutions, a task which is left to some further future contributions. Here we limit ourselves to show how the inferential results can be affected by spatial data imperfections and to point out what are the limits within which we can expect our results to vary as a function of the dirt of the dataset. More specifically, in the present paper we focus on the analysis of the effects of missing data and locational error which are likely to occur in many practical circumstances.

## 2. Missing data and missing location

This paper discusses some issues related to spatial data quality emerging in many empirical situations undermining statistical analysis. In particular we will discuss issues related to missing data and positional uncertainty. It is important to remark right at the beginning that, when dealing with spatial data, there is still currently a certain degree of ambiguity in the literature, on the concept of uncertainty and missing data. In order to clarify this, let us distinguish preliminarily the case *of missing data* from the case of *missing location.* In fact, we can have cases when the individuals' location is uncertain or missing, cases where the observation is uncertain or missing and cases when both of them are uncertain or missing altogether. In practice we can encounter 4 different cases that is important to distinguish because the consequences (and the solutions) are intuitively different in the different situations. They are reported in the following table.

| | | Missing data | |
|---|---|---|---|
| | | Yes | No |
| Missing location | Yes | 1 | 2 |
| | No | 3 | 4 |

Case 1 relates to the situation of *Missing spatial data and spatial location* when both the location and some measurements are unknown. We know of the presence of some individuals in a certain area, but we ignore where exactly they are and, furthermore, we do not have information about some or all their characteristics. Some individuals are simply not observed on the study-area map. This situation is not uncommon in many surveys in developing countries.

Case 2 refers to *Missing spatial data* when the location of individuals is perfectly known without error, but we are unable to observe some or all individuals' characteristics. This happens, for instance, when we know of the presence of some individuals (e. g. a firm) and its exact GPS location, but some or all information are missing at a certain moment of time (e. g., the number of employees or the production realized by that firm in that location). It is important to remark that this case represents the traditional case of missing data as it has been treated at length in the statistical literature (Little, 1988; Little and Rubin, 2002; Rubin, 1976; Roderick and Rubin, 2007) where solutions have been suggested to replace the observations that are missing following different interpolating strategies (e. g. the EM algorithm (Dempster, Laird and Rubin, 1977) and multiple imputation methods (Rubin, 1987)). These approaches, however, (apart from few remarkable exceptions, like e. g. Bihrmann and Ersbøll, 2015), do not treat adequately the nature of spatial data and do not suggest solutions to the problem of allocating in the space the information that is artificially recovered.

Case 3 refers to what we will term *unintentional positional error* that is when observations on individuals are available, but their location is missing or not known with certainty. For instance, we have a list of firms in a small area (like e. g. a census tract) and we also have observations on some of their statistical characteristics, but we don't know their exact address within the area. In this case it is common to assign the individual to the centroid of each area, but this procedure generates a positional error. In this case, not only the traditional statistical procedures proposed in the literature to minimize the fallacies produced by missing data (such as multiple imputation etc.) are useless, but even their consequences on statistical modelling are still largely unknown (see Bennett, Haining and Griffith, 1984).

Finally Case 4 refers to what we can call *intentional positional error* where both location and measurement of the single individuals are known. This case is also interesting from the statistical point of view because, in some instances, the individuals positions might be geo-masked *a-posteriori* before letting them publicly available to the analysts, in order to preserve confidentiality.

In the present paper we will concentrate on Cases 1 and 4 where we believe there is the most urgent need to fill a huge gap in the literature. In particular we aim at (i) treating within the same methodological framework different data quality problems that were previously treated separately, (ii) introducing novel approaches to tackle them and (iii) examining their consequences on statistical modelling .

Neighbouring lists and weights matrices (*W*) are fundamental tools in spatial statistics (Cressie and Wilke, 2010) and, specifically, in spatial regression (Arbia, 2014). Very general definitions that can be used when treating individual information may involve some negative function of the inter-point distances (such as, e. g., $w_{ij} \in W$ ; $w_{ij} = d_{ij}^{-\alpha}$ ; $\alpha > 0$ with $d_{ij}$ the inter-point distance) or *k*-nearest neighbours distances (see, e. g. Arbia, 2014). All cases discussed above lead to uncertainty about the true *W* matrix which is, in turn, induced by uncertainty about the measurement of inter-individuals' distances due to either missing (or geo-masked) location or missing data or both.

Suppose that the population of interest is constituted by, say, *n* individuals distributed in the study area. In Case 1 the true *W* matrix is *n*-by-*n*, but the observed *W* matrix is, conversely, *(n-m)*-by-*(n-m)* if *m* data are missing (*m*<*n*) and this produces distortion in spatial inference. In addition, when examining Cases 3 and 4, the true *W* matrix is biased by the fact that the individuals' position is not observed correctly either due to lack of information or to a voluntary decision dictated by confidentiality. In this respect it is possible to treat all problems within the same methodological framework aiming at characterizing the sensitivity of missing points or point displacements to the definition of a spatial weights matrix and at developing methods to minimize their distorting effect on inference.

In the following we will introduce the methodological framework that we will use to illustrate the inferential consequences of using dirty datasets in spatial econometric modelling. Without loss of generality we will consider the case of a simple linear regression just for the sake of illustrating our point. Let us indicate with *y* the vector of observations of the dependent variable in *n* locations (*i* = 1, 2, …, *n*) and let us consider the simple linear regression model:

$$y = \alpha + x\beta + u \tag{1}$$

with *x* the vector of observations of the non-stochastic regressor and *u* a normal independent random error. The term "location" can refer either to point data, or to areal data summarized by their centroids. The impact of a variation in the independent variable observed at location *i* on the dependent variable in location *j* is given by $IMP_{ij} = \frac{\partial y_i}{\partial x_j}$ and, in this standard modeling framework, a variation in *x* at location *i,* has only an effect in that location ( $IMP_{ii} = \beta$ ) while no impact is observed in the neighboring. Consider now (more realistically) a spatial econometric model which allows spatial spillover effects. The model (known as spatial lag model) can be specified as follows (Arbia, 2006; 2014):

$$y = \lambda Wy + \alpha + x\beta + u \qquad |\lambda| < 1 \tag{2}$$

where in addition to the previous notation, *W* is the already mentioned exogenously given weights' matrix which specifies the topology of the spatial observations, and $\lambda$ is the spatial correlation parameter. Equation (2) describes the fact that a variable *x* produces a direct effect on location *i*, but also an indirect effect in some neighboring locations. In this new setting the calculation of the impact is different and it is now given by (LeSage and Pace, 2009):

$$S = (I - \lambda W)^{-1}\beta \qquad (3)$$

with $S$ the matrix of the cross-locations impacts such that $IMP_{ij} \in S$ . A common summary measure in this case is the Average Total Impact $ATI = n^{-1}i^{T}Si$ (LeSage and Pace, 2009) which highlights how in this new framework the impact depends on both $\lambda$ and $\beta$. In fact, if the spatial correlation parameter is positive (i. e. the positive effects spill over the neighborhood) it emphasizes the global impact; conversely, if $\lambda$ is negative (i.e the positive effects in one location produces a negative effect over the neighborhood), the global impact will be reduced.

The model presented here constitutes the framework within which we can monitor the distorting effects on regression estimation and hypothesis testing induced by missing data and locational errors as we will show in the next section.

## 3. The effects of "dirty" spatial data on econometric modelling

### 3.1 Simulation setup

In this section we will present the results of a series of Monte Carlo experiments run to illustrate the effects of missing data and locational error on spatial econometric modelling. All Monte Carlo experiments are built using the following procedure. Consider an initial set of, say *n*, spatial locations (points) randomly and independently generated on a unit square that will be regarded as the truth. As a data generating process for "clean data" we assumed the spatial lag specification reported in Equation (2). The single explanatory variable, *x*, is generated by a zero mean normal distribution with standard deviation equal to 1.5. The stochastic disturbances *u* are *i.i.d.* generated by a standard normal distribution. The intercept and slope are both uniformly set to 1 in all simulations. We simulated different sets of "clean" values of the independent variable *x* and of the dependent variable *y* according to different degrees of spatial autocorrelation as identified by different values of the spatial lag parameter. More specifically, for each given value of the parameter $\lambda$, we obtained a vector of $n$ = 100 "clean" values of the dependent variable, under the spatial lag model, as follows:

$$y = (I - \lambda W)^{-1}X\beta + (I - \lambda W)^{-1}u \qquad (4)$$

We generated increasing intensities of both negative and positive spatial autocorrelation by considering the following values for $\lambda$: -0.01, -0.25, -0.75, 0.01, 0.10, 0.25, 0.50, 0.75 and 0.90. In all cases *W* is a row-standardized binary weight matrix derived from the *k*-nearest criterion, setting $k$ = 2 (see Arbia, 2014). The sequence of *n* values of the variable *x* and *y* thus generated will be considered the "clean" true values to be contrasted with their dirty versions.

### 3.2 Effects of missing data

We will start examining the results of a series of Monte Carlo experiments aiming at assessing the effects of missing data on regression estimation and hypothesis testing. In particular, through our Monte Carlo simulation, we will aim at assessing (i) the effects of missing data on the bias and the precision of the estimation of the various regression parameters, and (ii) the effects of missing data on hypothesis testing of regression parameters and in particular on the power of the significance tests.

Two major elements are relevant in this respect and need to be controlled for in the simulation, namely: (i) the *Proportion of Missing Points*, (the parameter $PMP \in (0,1)$ ), and (ii) the *spatial pattern* they display in the study area. This second aspect is systematically overlooked in the statistical literature on missing data, but it appears to be extremely relevant. In fact, if the missing data points are clustered in the study area, some of the geographical features (such as spill-over effects) could be hidden or cancelled out. In contrast, points missing randomly in the space are expected to produce milder effects on regression estimation and on hypothesis testing.

In our experiments we considered different situations of missing observations that can mimic those encountered in empirical cases. More specifically, we considered two spatial missing data mechanism. The first (which we will refer to as *Mechanism1*) reproduces the case where missing data tend to be concentrated in some specific zones of the study area. The second (termed *Mechanism2*), mimics the presence of pattern of spatial clusters of missing values around certain points.

For both mechanisms, we considered different degrees of spatial correlation of the dependent variable $y$ (parameter $\lambda$), different *Proportion of Missing Points (PMP)* and different intensities of clustering patterns. For each artificial "clean" dataset (generated as described in the previous Section 3.1) 10000 "dirty" versions were simulated according to different values of the *Proportion of Missing Points* (*PMP*) and the two missing mechanisms characterized by different spatial intensities.

In particular, with *Mechanism1* we cancel observations randomly with a probability that decreases with to the horizontal coordinate (see Figure 1a). The intensity of the data deletion is regulated by the parameter $\psi$ which regulates how steep the probability decreases. If $\psi = 0$ data are cancelled at random without any pattern, while higher values of $\psi$ imply a stronger spatial trend in the missing data structure and a concentration in the right side of the unitary square.

A step-by-step description of the simulation of *Mechanism1* is the following:

Step1: For each *i*-th observation, we derive the probability to be erased given by, $p_i = \left(PMPns_{1i}^{\psi}\right) \Big/ \sum_{i=1}^{n} s_{1i}^{\psi}$ , with $s_{1i}$ representing the horizontal coordinate of observation *i*.

Step2: We take a random sample of size *PMPn* of the observations to be erased from the elements of the clean data using without-replacement and the $p_i$'s as probability weights.

Step3: Using the “dirty” dataset obtained excluding the erased observations, we estimate a Spatial Lag Model using the ML estimator, we perform the 5% significance tests on parameters *α*, *β* and *λ* and we compute the impact measures.

After repeating steps 1 to 3 10000 times, we compute the expected value, bias and root mean square error of $\hat{\lambda}, \hat{\beta}$ and *ATI* and the power of the test for the null of $\lambda = 0$.

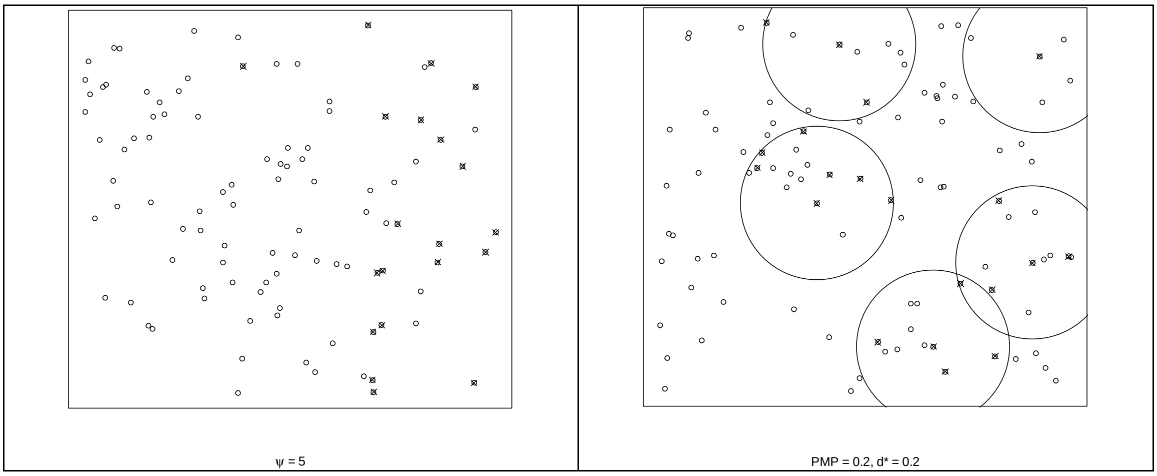


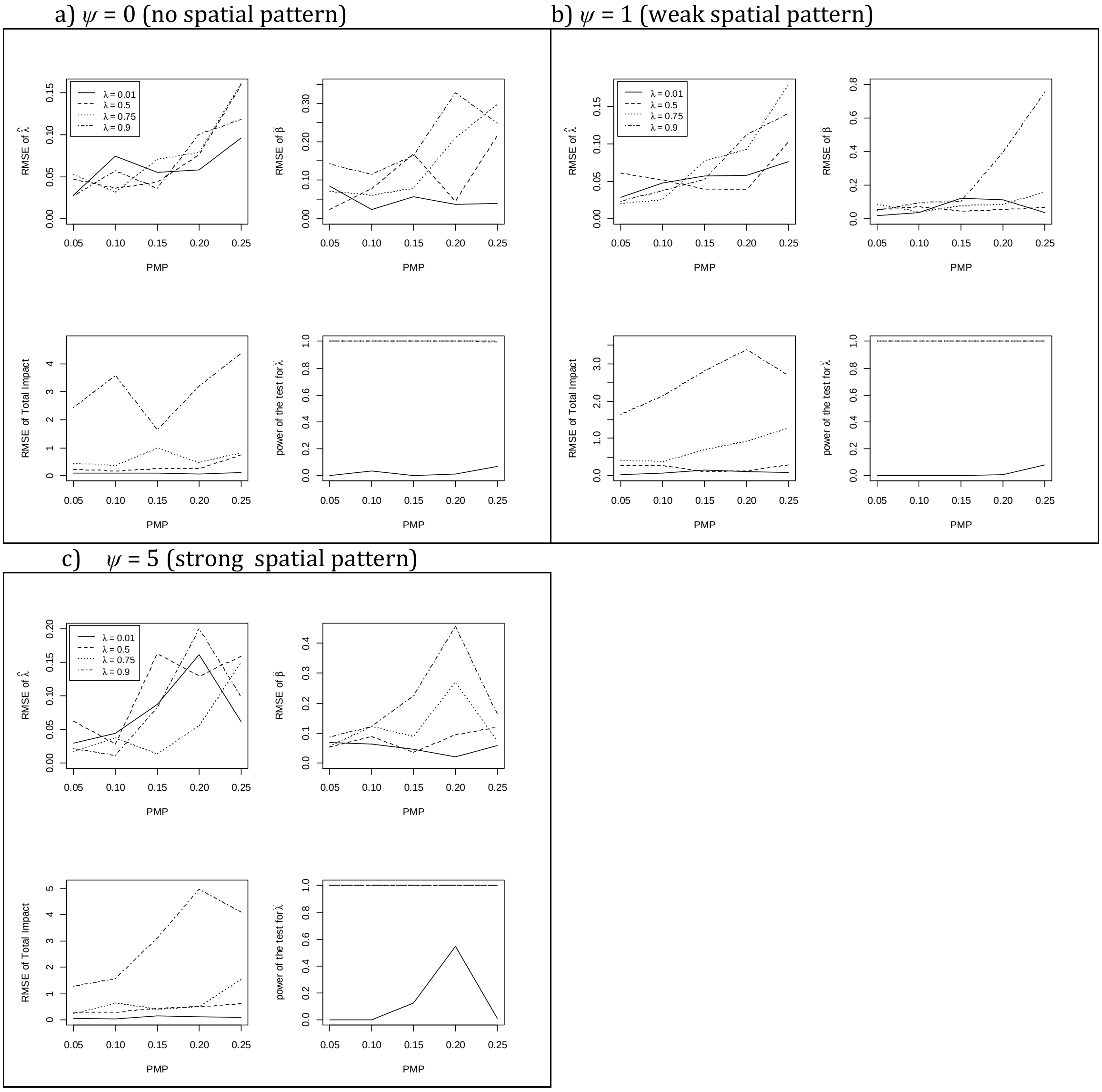
Figure 1: Mechanism 1 for the random deletion of points in the simulations (a) mechanism 1, (b) Mechanism 2

With *Mechanism2* we randomly identify 5 random points in the study area and we build up circular buffering zones centered on the selected points and with a given radius *d**. The values of the radius *d** were selected considering that the maximum distance in the unitary circle is $\sqrt{2}$. Then *PMPn* observations located within the circle are selected at random and eliminated (see Figure 1b).
Consequently, the steps used to simulate this second mechanism are the following:

Step1: Randomly select 5 observations (called *parent points*) among the clean data.
Step2: Select a set of points (called the *offsprings*) representing all observations located within a distance *d** from at least one of the five parents.
Step3: Randomly select *PMPn* observations to be erased amongst the parents and the offsprings.
Step4: Use the “dirty” dataset, obtained excluding the erased observations, to estimate a Spatial Lag Model using the ML estimator, perform the 5% significance tests on parameters *α*, *β* and *λ* and compute the impact measures.

The main results of or simulations are summarized in the following Figures 2 and 3. Only the results for $\lambda > 0$ are reported here for the sake of succinctness. Results for lambda $\lambda < 0$ are substantially symmetrical to those displayed here.

Looking at Figure 2 it is evident how the RMSE of $\hat{\lambda}$ increases monotonically with *PMP* and is positively related to the true value of $\lambda$. The RMSE of $\hat{\beta}$ is less sensitive to the presence of missing data unless $\lambda$ or PMP are also very high. The effect of *PMP* and $\lambda$ on the precision of the estimates is more evident for *ATI* in all three situations of spatial pattern when the true $\lambda$ is very high. In all case examined, the increase of RMSE for all parameters becomes sharper when the proportion of missing points exceeds 15%. In contrast the power appears to be substantially unaffected by missing data at all levels of PMP and $\lambda$.

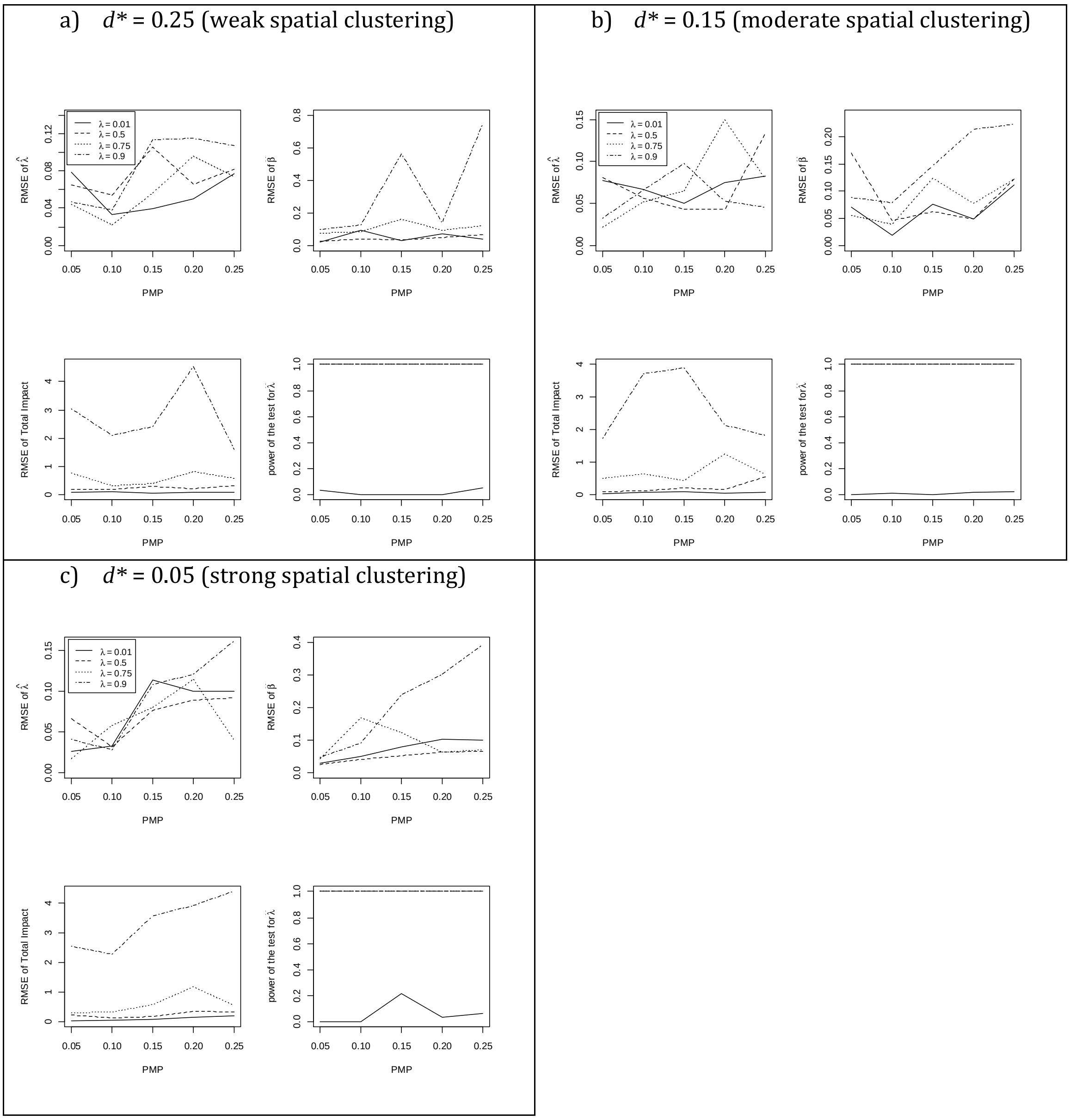

a) $\psi$ = 0 (no spatial pattern)

b) $\psi$ = 1 (weak spatial pattern)

c) $\psi$ = 5 (strong spatial pattern)

Figure 2: Root mean squared error of the estimation of λ, β and total impact (ATI*)* and power of the LR test of λ for various levels of the spatial correlation parameter λ (0.01; 0.5;0.75,0.9) and for various levels of *PMP* (0.05, 0.10, 0.15, 0.20, 0.25). Random cancellation is performed with Mechanism 1 with different intensities of clustering of missing data: a) $\psi$= 0, b) $\psi$ = 1, c) $\psi$ = 5.

The results of the simulation generated using Mechanism 2 are reported in Figure 3 and substantially confirm the findings, but with the effects on the precision of the estimates that become more evident for all parameters. In particular, we notice a sharper increase of the RMSE of $\beta$ when *PMP* increases in the estimation especially in the case of high clustering of missing data (Figure 3c).



a) $d*$ = 0.25 (weak spatial clustering)

b) $d*$ = 0.15 (moderate spatial clustering)

c) $d*$ = 0.05 (strong spatial clustering)

Figure 3: Root mean squared error of the estimation of λ, β and total impact (ATI*)* and power of the LR test of $\lambda$ for various levels of the spatial correlation parameter λ (0.01; 0.5;0.75,0.9) and for various levels of *PMP* (0.05, 0.10, 0.15, 0.20, 0.25). Random cancellation is performed with Mechanism 2 using different cut-off radius for the clustering circles: a) *d** =0.05, b) *d**=0.15, c) *d**=0.25. d* represents the distance measured in the unitary square where the maximum distance is $\sqrt{2}$ .

We also notice that in this case the effects on the efficiency of the estimates of ATI are very strong as soon as a certain degree of *PMP* is introduced when the true $\lambda$ is high. Values of the RMSE(ATI) are 0.0389 in the case of low spatial correlation ($\lambda$ =0.01) and small PMP (=0.05) but they become about hundred times bigger (RMSE(ATI))=4.4064) in the opposite case of high spatial correlation ($\lambda$=0.9) and high PMP (=0.25).

### 3.3 Effects of positional error

To identify the effects of a positional error in spatial regression, we use again the same dataset artificially generated as described in Section 3.1. So the observations on the variables $x$ and $y$ are assumed to be fixed in all experiments but, at each simulation run, we displace the observations using a random mechanism.

In particular we consider the case of intentional positional error where the mechanism of random displacement is often known. We consider the following procedure: at each simulation run, for each location we select a random angle and a random distance. The random angle is generated from an uniform distribution $U(0,360)$, whereas the random distance is generated from a uniform $U(0,\vartheta)$, where $\vartheta$ is a further simulation parameter that will be left free to vary in a given range. In particular, we considered $\theta$ to vary between 0.05 and 0.25 in a study area defined in the unitray square where the maximum distance between points is represented by the diagonal and thus equal to $\sqrt{2}$. This mechanism is often used in empirical case to preserve confidentiality (See, e. g. USAID, 2013). We then displace each point with the random angle along the random distance thus generated. Similarly to the missing data case, at each step of the replication we estimate the parameters of model (2) and we calculate the RMSE of these estimates. We then average the RMSE obtained over the simulation runs. We expect the RMSE to increase with the distance $\theta$ and we study how this behavior is related to the other simulation parameters. Similarly, in order to analyze the effects of displacement on hypothesis testing in each simulation run we calculate the significance of the test and again we reject the null if $\alpha \leq 0.05$. We can then compute how many times we wrongly accepted the null and monitor how the empirical power thus evaluated changes by increasing the parameter $\theta$. The main results are summarized in Figure 4.

The effects of location error on the precision of the estimates of all parameteres ($\lambda, \beta$ and $ATI$) are dramatic. First of all the RMSE of all parameters increase sharply with the radius of the displacement effect ($\theta$). Secondly, while this effect is almost absent when $\lambda$ is close to zero, it increases dramatically when $\lambda$ gets bigger. The effects of locational errors are relevant on the efficiency of the estimates of the Average Total Impact which become extremely unreliable in the case of a strong spatial pattern of data and of a high intentional locational error. For instance, in the case of very weak spatial correlation ($\lambda$ = 0.01) and small displacement ($\theta$ = 0.05) the effect is negligible (in this case RMSE(ATI) = 0.1733), in contrast, in the case of very high spatial correlation ($\lambda$ = 0.9) it explodes to values that are very high even in the presence of very small locational errors (e. g. RMSE(ATI)= 4.4580 when $\theta$ = 0.05) becoming 43 times higher (RMSE(ATI)= 7.5908) in the case of a radius of displacement of $\theta$ = 0.25.

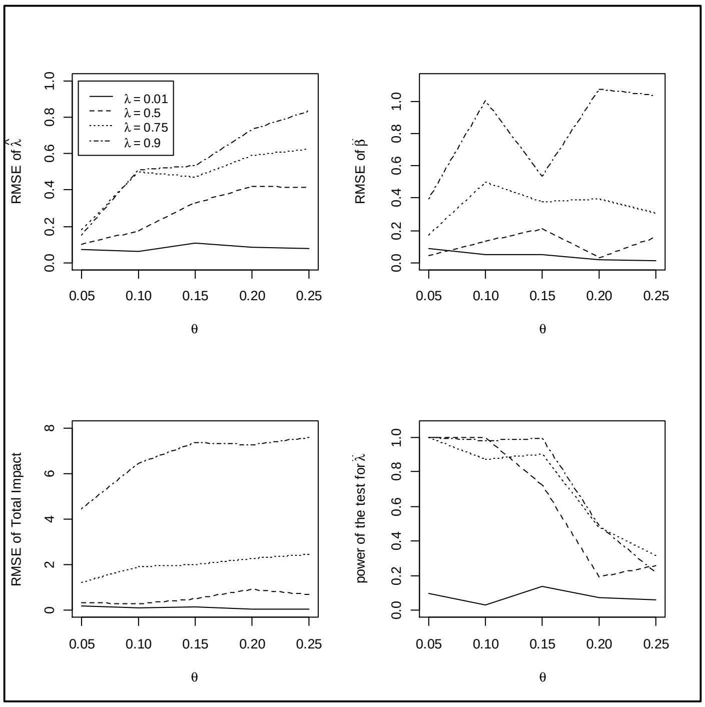


Figure 4: Root mean squared error of the estimation of $\lambda$, β and total impact (ATI*)* and power of the LR test of λ for various levels of the spatial correlation parameter $\lambda$ (0.01; 0.5;0.75,0.9) and for various levels of the dislocation radius $\theta$ (0.05, 0.10, 0.15, 0.20, 0.25).

The effects on the LM test of $\lambda$ are equally dramatic with a clear trend to decrease its discriminating power when $\theta$ increases. From the graphs it is evident how the test can tolerate small values of dislocation errors only if $\lambda$ is very close to zero, but, as soon as the dislocation radius $\theta$ is greater than 0.15 and $\lambda$ is greater than 0.01 the power drops dramatically towards zero.

## 4. Conclusions

This paper discussed some of the consequences on estimation and hypothesis testing procedures in spatial econometric modelling of having "dirty" spatial datasets contaminated by missing data and locational errors.

The presence of missing data reduces the precision of the estimates and this reduction in efficiency is emphasized by the presence of strong spatial correlation. Furthermore the effects are more relevant when data are missing in clusters in which case entire geographical features, like e. g. spatial spillovers, tend to disappear. The practice of intentionally geo-masking individual data for protecting confidentiality has also strong effects on the estimation and hypothesis testing. These effects are directly related with the entity of the induced displacement and to the degree of spatial correlation in the data. We observed a sharp reduction of the efficiency of the estimators of all models' parameters after a displacement distance that is 15% of the side of the study area. Similarly after this distance the LR test of significance on the spatial parameter becomes highly unreliable.

The aim of suggesting procedures to reduce the consequences of dirty data on spatial econometric modelling is left to a future study. The results presented in this paper aim at making researchers aware of the possible consequences of the presence of missing data and locational error while running empirical analyses. Ideally any empirical analysis should contain a discussion of the experimental situation that led to the data collection, e.g. in terms of the proportion of missing points and the amount of positional error present in the dataset, in order to be able to stress whether the results obtained can be considered robust to the observed data imperfections of the data or if, conversely, their credibility is dramatically undermined by them.

## References

Anselin L. (1988) *Spatial econometrics*, Kluwer Academic Press, Dordrecht.

Arbia G. (2014) *A primer for spatial econometrics*, Palgrave MacMillan, London.

Cliff A.D., Ord J.K. (1972) *Spatial autocorrelation*, Pion, London.

Collins B. (2011) Boundary Respecting Point Displacement, Python Script, Blue Raster, LLC.

Cozzi S., Filipponi D. (2012) The new geospatial Business Register of Local Units: potentiality and application areas, *3rd Meeting of the Wiesbaden Group on BusinessRegisters-International Roundtable on Business Survey Frames*, Washington, D.C. 17 – 20 September 2012.

Deuchert E., Wunsch C. (2014) Evaluating nationwide health interventions: Malawi's insecticide-treated-net distribution programme, *Journal of the Royal Statistical Society*, A, 177, 523–552.

Dubin R.A. (1992) Spatial autocorrelation and neighborhood quality, *Regional Science and Urban Economics*, 22, 3, 433–452.

IFNC (2015) http://www.sian.it/inventarioforestale/jsp/home_en.jsp.

Little R.J.A. (1988) Missing-data adjustments in large surveys, *Journal of Business and Economic Statistics*, 6, 3, 287–296.

Little R.J.A., Rubin D.B. (2002) *Statistical analysis with missing data, 2nd edition*, Wiley & Sons, New York.

Roderick J.L., Rubin D.B. (2007) *Statistical analysis with missing data*, Wiley, New York.

Rubin D.B. (1976) Inference and missing data, *Biometrika*, 63, 581–592.

Rubin D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, Wiley & Sons, New York.

USAID (2013) Geographical displacement procedure and georeferences data release policy for the demographic and health surveys, *DHS Spatial Analysis Report*, 7, September 2013.