11. Spatial econometrics on survey data

Raphaël Lardeux, Thomas Merly-Alpa
INSEE

11.1 First approach by simulation 280
11.1.1 Simulation of a SAR 280
11.1.2 Sampling procedures 282
11.1.3 Results and interpretation 284
11.1.4 A "size effect" 285
11.1.5 Robustness 287
11.2 Prospects for resolution 288
11.2.1 Moving to a higher scale by aggregation 288
11.2.2 Imputing missing data 289
11.3 Empirical application: the manufacturing sector in Bouches-du-Rhône 291
11.3.1 Data 292
11.3.2 Identification 292
11.3.3 Estimation 293
11.3.4 Spatial estimates on samples 294
11.3.5 Estimation on aggregated data 296
11.3.6 Imputation of missing data 297

Abstract

Spatial econometrics requires comprehensive data on a territory, which in principle prohibits the use of survey data. This chapter presents the pitfalls in estimating spatial autoregressive models (SAR) on sampled data and assesses potential corrections offered by the empirical literature. We identify two sources of bias: *(i)* a "size effect" resulting from the distortion of the spatial weighting matrix and *(ii)* a "sparsity effect" resulting from the omission of units spatially correlated with the units observed. Both effects tend to underestimate the magnitude of spatial correlations. However, the bias is lower in the case of a cluster survey and when the sample is large enough. Two types of methods are offered by the empirical literature to sidestep these pitfalls: imputing missing values (linear regression, hot deck) and aggregating data on a higher scale. These potential solutions are efficient under very restrictive conditions, part of the difficulty being the reconstruction of complex information based on few observations. The last part of the chapter illustrates this issue with the estimation of production externalities between plants of the manufacturing sector located in the French department of Bouches-du-Rhône.

Prior reading of Chapters 2: “Codifying the neighbourhood structure”, 3: “Spatial autocorrelation indices” and 6 “Spatial econometrics: common models” is recommended.

Introduction

Recent developments in spatial econometrics and geolocation have made it possible to analyse spatial phenomena at very local scales. Concepts derived from spatial econometrics analysis are used in an increasingly diverse range of fields such as geostatistics, economics and network analysis. However, the application of these spatial analysis methods requires comprehensive data, which is not always accessible (due in particular to non-response, excessively lengthy collection time, . . .) and cannot easily be processed within the limited time available. The extension of spatial econometrics to survey data would make it possible to take full advantage of a detailed piece of information to produce a fine-grained metric of the incidence of spatial correlations on econometric estimates[1].

In this chapter, we discuss recent developments in the application of spatial estimation methods when a part of the data set is missing, particularly in the case of survey data. We show that estimating spatial econometric models on survey data underestimates spatial correlations, in particular when data is collected according to simple random sampling and when the sample is small. Ignoring missing observations is never an appropriate solution, but other potential correction methods such as imputing missing data or switching to a coarser scale through data aggregation may be efficient under very restrictive assumptions. We will not be addressing the possibility of a spatial survey, which is particularly complex in the case of social data[2]. Neither will we develop the case of unknown location.

Why does spatial econometrics require comprehensive data? Traditional econometrics is based on the hypothesis of mutual independence between observations. Estimating a model on a data subset may affect the power of statistical testing but, in the absence of a selection problem, estimators remain unbiased and efficient. In contrast, in spatial econometrics models, observations are considered to be correlated with each other: each unit is influenced by its neighbours. Removing some observations amounts to omitting their links with nearby units, which biases the spatial correlation parameter as well as spatial effects. We show that this bias minimises the value of the spatial correlation parameter, since some neighbourhood links are no longer taken into account in the estimate.

Conceptually, spatial econometrics differs from traditional econometrics because of the assumptions about the data generating process. In traditional econometrics, observations are considered as a representative random sample of a population and are interchangeable. Spatial analysis conceives of them as the sole completion of a spatial process, each observation being necessary to estimate the underlying process[3]. Spatial econometrics has been developed within the framework of Cliff et al. 1972, characterised by exhaustive and perfect information on spatial units and by the absence of missing data (Arbia et al. 2016). In practice, these conditions are almost never met.Therefore, direct application of spatial estimation techniques to sampled data can significantly affect the results.

The estimation of spatial models on survey data induces several issues. First, estimates are disrupted by a "size effect". The existence of m missing data in a population of size n gives rise to a weighting matrix of size $(n-m) \times (n-m)$ instead of the effective weighting matrix of size $n \times n$. This dimension reduction skews in turn the estimated spatial correlation parameter (Arbia et al. 2016). Second, estimation on an incomplete data set implies that interconnections

[1] Pinkse et al. 2010 refer to this prospect as "the future of spatial econometrics".

[2] On these questions, readers are invited to refer to Chapter 10 "Spatial sampling".

[3] In this sense, spatial analysis is similar to time series, where the observed dataset is derived from a stochastic process.

between observed and missing units will be overlooked, which induces a measurement error on the neighbourhood (regressor $\mathbf{W}Y$) and biases estimated parameters. We compare the consequences of both effects estimating SAR models on different samples drawn from a population which has been simulated according to the exact same SAR specification. We show that beyond the "size effect", the "sparsity effect" has significant consequences.

While a number of corrections have been proposed, none of them are adequate to the current framework[4]. When the location of individuals is known, imputation is generally preferred (Rubin 1976; Little 1988; Little et al. 2002). However, naive imputation, for example through linear models, is not efficient to provide unbiased estimates (Belotti et al. 2017a). To get around this problem, Kelejian et al. 2010 develop estimators when only an incomplete subset of a population is available. Wang et al. 2013a suggest an imputation method through two-stage least squares, in a setting where the values of the dependent variable are randomly missing. In the same framework, LeSage et al. 2004 rely on the EM algorithm (Dempster et al. 1977). First, step "E" (expected) assigns a value to the missing data, conditionally on observables and parameters of the spatial model. Then step "M" (maximisation) determines the value of these parameters by likelihood maximisation. By iteration, this procedure makes it possible to draw from an estimated model all the information available to impute missing values. More recent work by Boehmke et al. 2015 extends this procedure to the case of missing observations (unknown dependent and independent variables).

Recent empirical papers illustrate the importance of these corrections. In a hedonic price model, LeSage et al. 2004 rely on the EM algorithm to predict the value of unsold housing. In a network model with space autocorrelation, Liu et al. 2017 show that the detection of a peer effect requires taking the sampling process into account. Yet, such complex imputation methods based on an estimated model (*model-based*) are still rarely applied. When some data is missing, the solution generally chosen is to remove the corresponding observations from the field of the estimation. The risk is that an attenuation bias in the spatial correlation may be generated. Some estimations are limited to a subset, in particular, a specific region or group, leading to a potential "size effect" as well as an underestimation of correlations at the edge of the considered area (Kelejian et al. 2010). Lastly, while most applications are performed on aggregated data to benefit from comprehensive data on a larger scale, this solution can induce positional errors[5] (Arbia et al. 2016) as well as an ecological bias (Anselin 2002b). We discuss the impact of these various methods on spatial estimates.

The issue of missing values in a framework where observations are correlated has been highlighted by other fields of statistics related to spatial econometrics: time series, geostatistics and network econometrics. Time series and geostatistics are similar to continuous spatial data processing. The issue of missing data was addressed very early in the field of time series (Chow et al. 1976, Ferreiro 1987). Jones 1980; Harvey et al. 1984 recommend using a Kalman filter to concurrently estimate a model and impute values. Geostatistic analysis corrects incomplete data sets either upstream using spatial sampling methods, or by predicting the value of a continuous spatial variable in an unknown position (spatial interpolation or kriging, see Chapter 5: "Geostatistics"). Longitudinal approaches combining kriging with the Kalman filter have also been developed (Mardia et al. 1998).

[4] In particular, these methods vary according to the underlying assumptions on missing data. Depending on the value and/or location of the observations, the dependent and/or independent variables are affected and depending on whether the likelihood of a piece of data going missing depends on the correlations with observable and/or unobservable data. The literature on the incidence of missing data thus establishes a distinction between *Missing at Random* (MAR), *Missing Completely at Random* (MCAR) and *Missing Not at random* (MNAR). cf Rubin 1976, Huisman 2014

[5] Arbia et al. 2016 develop this concept to refer to cases where the position of an observation (X,Y) is not known precisely. For example, lack of precision in measurement, metric blurred for confidentiality reasons, missing addresses.

However, continuous data methods cannot be transposed straightforwardly into economic and social analysis, where data is fundamentally discrete. Furthermore, the use of such spatial survey techniques would contradict fundamental principles of social data collection, such as equi-weighting and the use of deterministic sampling bases. Network econometrics focuses on the issue of missing observations (Burt 1987; Stork et al. 1992; Kossinets 2006). Estimation of spatial autocorrelation on a sample of a network is gaining momentum with the growing use of social networks (Zhou et al. 2017). However, practical solutions remain rare. As in spatial econometrics, the main difficulty is to reconstruct information on unobserved units based on observed data, without knowing the effect of the former on the latter (Koskinen et al. 2010). In particular, Huisman 2014 does not come out clearly in favour of any traditional imputation strategy and remains very cautious about extensions of imputation methods to network data. Solutions based on sampling methods have also been proposed in order to collect data on populations of interest (Gile et al. 2010).

This chapter focuses on two questions: which biases are generated by the estimation of spatial econometric models on survey data? What are the consequences of classic solutions aimed at correcting missing data (data deletion, imputation, aggregation)? These questions have been addressed by Arbia et al. 2016, who proceed by simulation and observe a stronger incidence of missing data when these are grouped in clusters, in which case all local phenomena may be lost. However, they consider cases where missing data accounts for at most 25% of the population, which is very low compared to survey data, where they generally reach more than 90% of the population.

Section 11.1 highlights the bias resulting from the application of spatial methods to a non-exhaustive sample and discusses its magnitude depending on the percentage of observations sampled and the sampling method. Section 11.2 shows the consequences of some usual solutions: shifting to a higher level by aggregation and the imputation of missing values. Section 11.3 illustrates these biases through the estimation of production externalities between industries of the French department of Bouches-du-Rhône.

11.1 First approach by simulation

In this section, we show that the estimation of a spatial autoregressive model (SAR) on sample data is biased. In order to do that, we proceed through Monte Carlo simulations. First, we simulate a spatial data set across a geographic area so that units are correlated according to a given value of the spatial correlation parameter. Second, we draw samples from this data set and estimate the value of the spatial correlation parameter for each one of them.

11.1.1 Simulation of a SAR

The geographical space chosen is a map of Europe[6], detailed at the administrative level NUTS3 (the lowest level in the NUTS hierarchy defined by Eurostat, which corresponds to small areas where specific studies can be carried out, such as the French departments) from which the furthest islands and Iceland have been removed in order to maintain a homogeneous and compact geographical space. From the *shapefile* of Europe, we build a neighbourhood matrix $\mathbf{W}$ based on distance, so that the weight associated with two neighbouring units decreases according to the square of the distance and is cancelled when this distance exceeds a limit threshold. Residuals and an explanatory variable are drawn from Gaussians: $\varepsilon \sim \mathcal{N}(0,1)$ and $X \sim \mathcal{N}(5,2)$, making it possible to ultimately simulate a variable Y following a SAR model (*Spatial Auto-Regressive*):

$$Y = (1-\rho\mathbf{W})^{-1}X\beta + (1-\rho\mathbf{W})^{-1}\varepsilon \tag{11.1}$$

[6] This card is shared on the site: `http://ec.europa.eu/eurostat/fr/web/gisco/geodata/reference-data/administrative-units-statistical-units`.

with $\beta = 1$ and $\rho = 0.5$, reference parameters which we try to find by estimating the exact same SAR model on samples. Data from simulated variables Y is shown in Figure 11.1. The presence of concentrated coloured areas is characteristic of the positive spatial autocorrelation resulting from the data-generating process.

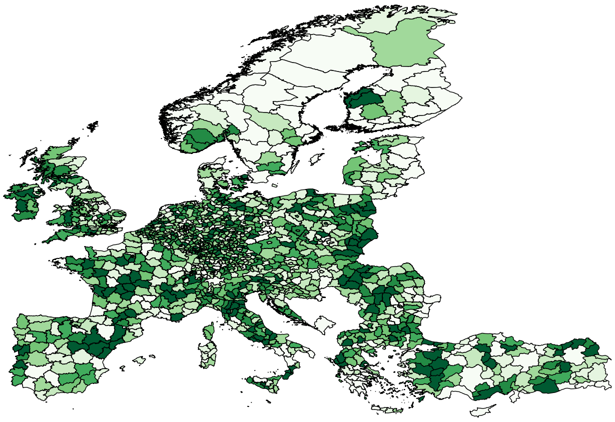

Figure 11.1 – Y simulated using a SAR model
Copyright: *EuroGeographics for administrative boundaries*

Table 11.1 shows the results of the estimation of a SAR model across all NUTS3 zones in Europe. They confirm the validity of this simulation, since the estimated parameters β and ρ are very close to the values initially calibrated.

β	ρ	Direct	Indirect	Total
0.989	0.494	1.043	0.860	1.902

Table 11.1 – Parameters estimated by SAR across all zones

Box 11.1.1 — Simulation of a SAR with with R. To simulate a SAR in R, the most important step is formatting its neighbourhood matrix $\mathbf{W}$:

```
D <- nb2listw(W, style="W", zero.policy=TRUE)
```

Once the neighbourhood matrix is in the format `listw`, $1-\rho\mathbf{W}$ must be inverted using the following function, of which ρ is one of the parameters:

```
InvD <- invIrW(D,rho)
```

Note: this step may be time-consuming. Then, all we need to do is simulate our variable Y:

```
Y <- (InvD %*% X) + (InvD %*% eps)
```

11.1.2 Sampling procedures

The challenge lies in examining the capacity of spatial models to correctly estimate ρ and β on samples drawn from this simulated data. In particular, we discuss the effect that sampling some of these areas may have on the estimation of the underlying model.

A survey consist in randomly selecting, using a procedure referred to as a sampling plan, a set of n units within a population of N, where n is often much smaller than N in order to limit the costs of collecting information. Survey theory states that estimates made using the sample extend without bias to the total population, but are more precise when the sample size increases and when the sampling plan is suited to the estimated variable. To explore questions around surveys, it is advised that readers refer to Ardilly 1994, Tillé 2001 or Cochran 2007.

In the rest of this section, we present a number of conventional sampling techniques and how they can be applied within the framework of the European NUTS3. However, we can already make a number of general hypotheses and comments, following the ideas developed in Goulard et al. 2013 regarding the new French population census. On the one hand, the effect should obviously not be the same according to size n of the selected sample. With just under a dozen zones, the initial spatial structure will not be able to be reconstructed, while sampling 95% or even 99% of the zones should make it easily recoverable. On the other hand, the question of the sampling method will also need to be addressed. Is the spatial dimension taken into account in the method? We can refer to Chapter 10 "Spatial sampling" to delve deeper into these issues.

Simple random sampling

Simple random sampling consists in drawing independently and without replacement a number n of marbles from a large bowl N. Under such conditions, all individuals have the same chance of being selected in the sample. Where one individual has been selected in a sample, this reduces the likelihood that others will also be included. In our case, n areas are selected in an entirely random manner. Figure 11.2 shows an example of a sample.

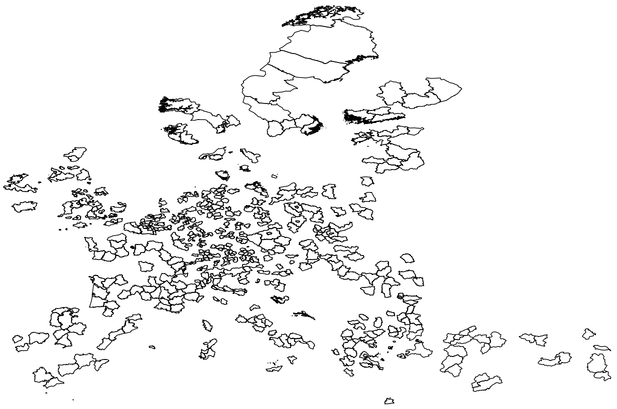

Figure 11.2 – A sample drawn according to simple random sampling ($n = 500$)
Copyright: *EuroGeographics for administrative boundaries*

Poisson sampling

Poisson (or Bernoullian) sampling consists in flipping a coin to determine, based on heads or tails, whether each individual should be included in the population. Under such conditions, all individuals have the same chance of being selected in the sample. While an individual's being selected for a sample does not affect the likelihood that the others will be included, the sample size is not set beforehand. In our case, each zone has a likelihood p of being retained in the sample. The resulting sample size is then pN in expected terms.

Cluster sampling

Cluster (or areolar) sampling consists in selecting groups of individuals together. Individuals always have the same chance of being selected in the sample. However, the selection of an individual within a sample has a strong impact on the likelihood that the others will also be included, as individuals of the same cluster are always selected together. Here, the process consists in combining NUTS3 zones into different clusters, then making random selection of some of these clusters. The main interest is to limit collection costs, at the expense of a loss in precision due to intra-cluster homogeneity.

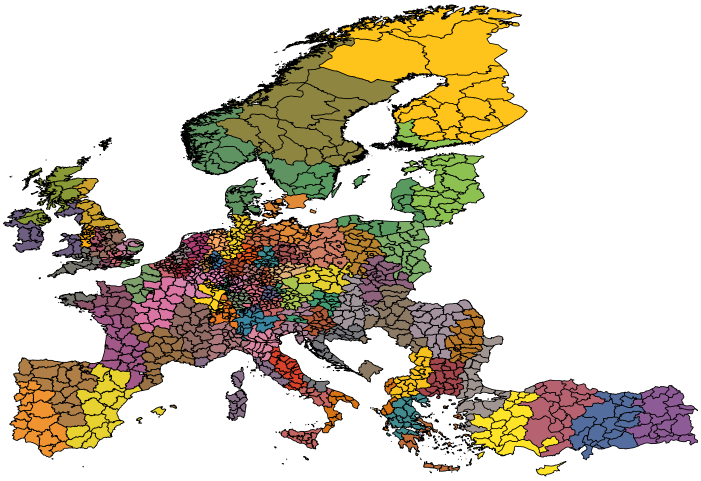

Figure 11.3 – Division of Europe into clusters
Copyright: *EuroGeographics for administrative boundaries*

It would be possible to use the different NUTS1 or NUTS2 levels as clusters. However, they are large in size and do not all include the same number of NUTS3. The problem is that large clusters limit the number of possible simulations. In contrast, clusters with different numbers of zones introduce either an issue of different sampling weights between individuals, which we do not wish to address here (see Davezies et al. 2009 for a discussion on the use of sampling weights in econometrics), or a problem of variable sample size, which may result in effects that are too complex to analyse. We therefore form clusters of the same size while maintaining a certain geographical consistency. As the weighting matrix is based on geographical distance, we give preference to the least-extensive clusters possible.

In order to obtain the same-size clusters, the number of clusters must be a divisor of the number of NUTS3 zones. In order to limit the size of the clusters, we bring together the 1,445 NUTS3 zones into 85 clusters of 17 zones each. For this purpose, we use an algorithm to build the clusters. From the area furthest away from the centre of the map, we aggregate the areas closest to it until we reach 17. As the clusters are built one by one, the most remote NUTS3 zones will already be assigned to the construction of the previous clusters, and the algorithm will continue with more central zones. The resulting clusters are shown in Figure 11.3.

Stratified sampling

In a stratified sampling, n units are also drawn. The difference is that n_1 units are drawn from a first stratum, n_2 from a second, etc. n_H in a Hth, where $n = n_1 + n_2 + \ldots + n_H$. To perform stratified sampling, it is important that the H strata be well defined, firstly, and that the allocation $(n_1, \ldots, n_H)$ be well-selected, secondly. A classical allocation is Neyman's allocation, the property of which is to minimise the variance of the estimator of the total of a variable of interest (see, for example Tillé 2001). The formula is as follows:

$$n_h = n \frac{N_h S_h}{\sum_{i=1}^{H} N_i S_i} \tag{11.2}$$

with n the size of the total sample, N_h the size of the stratum h and S_h the dispersion of the variable of interest within the stratum h. In some cases, when the behaviours with regard to the variable of interest are heterogeneous, this formula may give reason to run an exhaustive sampling of certain strata, *i.e.* to apply to them a 100% sampling rate.

11.1.3 Results and interpretation

In order to estimate the effect of sampling part of the European NUTS3, we use the Monte Carlo method. We thus carry out 100 simulations of Y based on a SAR model, then draw 100 samples for each of them. The exact same SAR model is estimated on each sample to retrieve the estimated values of the parameters. The parameters ultimately shown in the results are the averages of ρ and β out of the 10,000 samples, and their standard deviations are calculated on these 10,000 values.

For each of the 10,000 draws, we keep the values of X and Y and we reconstruct a spatial weighting matrix $\mathbf{W}_{\text{echantillon}}$ based on distance, as previously, but limited to the units included in the sample. Different sample sizes and sampling methods are considered.

Simple random sampling

Table 11.2 shows the results using simple random sampling for sample sizes n varying from 50 to 250 zones. Significant spatial autocorrelation may be detected for a sample size above $n = 150$, which corresponds in the present case to a sampling rate of $1/10$. Parameter β is estimated without bias, regardless of the size of the sample, but estimated parameter $\hat{\rho}$ is well below its true value $\rho = 0.5$ used for the simulation of the data set. Therefore, for small samples, the indirect effect does not significantly differ from zero and remains far lower than that observed across the entire population. The spatial autocorrelation is largely underestimated.

Cluster sampling

Cluster sampling makes it possible to maintain a strong geographical structure, which in our case appears beneficial for detecting spatial effects, particularly for small values of n. From the clusters shown in section 11.1.2, we carry out drawings of different numbers of clusters ranging from 3 to 15 clusters, *i.e.* 51 to 255 zones. Table 11.3 shows the results obtained for values of $n = 17p$, the sample size composed of p clusters.

With a cluster survey, the estimate $\hat{\rho}$ is closer to the true value of this parameter, which lies within its confidence interval. The accuracy of the estimate clearly improves when n increases, but the estimator remains biased. Thus, contrary to the case of simple random sampling, it is possible to capture spatial interactions even with a very low survey rate of around 3%. In fact, the units surveyed are highly concentrated in space and therefore highly representative of spatial correlations.

n	$\hat{\rho}$	$\hat{\beta}$	Direct	Indirect	Total
50	0.043	1.055^{***}	1.056^{***}	0.016	1.072^{***}
	(0.043)	(0.125)	(0.125)	(0.017)	(0.128)
100	0.058^{*}	1.050^{***}	1.052^{***}	0.032^{*}	1.083^{***}
	(0.031)	(0.087)	(0.087)	(0.019)	(0.091)
150	0.072^{**}	1.049^{***}	1.051^{***}	0.048^{**}	1.099^{***}
	(0.028)	(0.068)	(0.068)	(0.020)	(0.073)
250	0.101^{***}	1.051^{***}	1.054^{***}	0.080^{***}	1.135^{***}
	(0.026)	(0.051)	(0.052)	(0.023)	(0.060)

Table 11.2 – Estimation of a SAR model on samples drawn by simple random sampling
Note: *** denotes significance at 1%, ** significance at 5% and * significance at 10%. Standard deviations are shown between brackets. n: number of observations in the sample. These estimates come from 10,000 simulations.

n	p	$\hat{\rho}$	$\hat{\beta}$	Direct	Indirect	Total
51	3	0.309^{*}	1.015^{***}	1.051^{***}	0.441^{*}	1.492^{***}
		(0.237)	(0.091)	(0.097)	(0.262)	(0.310)
102	6	0.348^{***}	1.017^{***}	1.054^{***}	0.493^{***}	1.546^{***}
		(0.100)	(0.063)	(0.066)	(0.188)	(0.215)
153	9	0.363^{***}	1.017^{***}	1.054^{***}	0.516^{***}	1.571^{***}
		(0.078)	(0.052)	(0.055)	(0.152)	(0.176)
255	15	0.377^{***}	1.014^{***}	1.052^{***}	0.541^{***}	1.593^{***}
		(0.058)	(0.038)	(0.040)	(0.119)	(0.136)

Table 11.3 – Estimation of a SAR model on samples drawn by cluster
Note: *** denotes significance at 1%, ** significance at 5% and * significance at 10%. Standard deviations are shown between brackets. n: number of observations in the sample. p: number of clusters in the sample. These estimates come from 10,000 simulations.

However, if the number of units drawn is low, then the same is true for the accuracy of the spatial correlation estimate. Therefore, the indirect effect is effectively detected, even for small samples, and its value is closer to that obtained on the total population. The estimation of geographical effects therefore appears reasonable with surveys relying on cluster sampling.

Two questions remain. First, would this cluster sampling not lead to overestimating the detection of a spatially-controlled model, even if the effect is not major on the entire population? On the one hand, as there are few values X and Y, the term $\mathbf{W}Y$ is paradoxically quite well-known, which might encourage giving priority to this approach. Second, and this will be developed in Part 11.1.4, the gap observed between the estimated $\hat{\rho}$ and the real value used to generate the SAR can appear surprising, even though the spatial effects are clearly detected.

11.1.4 A "size effect"

The results derived from simulation may come as a surprise to econometricians. Simple random sampling can be related to the super-population model used in econometrics[7]. Therefore, the

[7] This term is connected with the difference between design-based and model-based approaches. Under a design-based approach, we assume that the population has deterministic Y values - the usual approach. Under a model-based

estimation of a population or model parameter is usually unbiased, as long as the sampling plan is correctly specified. However, spatial autocorrelation parameter ρ does not follow this "traditional law" of sampling theory[8].

Notwithstanding the question of the random selection method used for the zones on which information about Y is retrieved, we will restrict this analysis to a number of zones below that of the entire population, inducing a change in the underlying spatial structure. Intuitively, a spatial effect results from interactions between all units constituting a territory. When this effect is spatially homogeneous, omitting some units implies that we neglect their contribution to the total spatial effect, which is then underestimated. We call this first component a "size effect". Moreover, available data may be more or less scattered spatially. When observed units are too sparse, they hardly account for the structure of spatial correlations. This "sparsity effect" also leads to underestimating the spatial correlation parameter.

The question of ecological bias, *i.e.* estimation errors of spatial econometric models that come from poor spatial specification, whether in terms of data granularity (resolution) or boundary problems, is similar to this issue. Thus, it is entirely possible, when you are restricted to n zones, with $n < N$, to never achieve as strong a spatial effect as across the entire population.

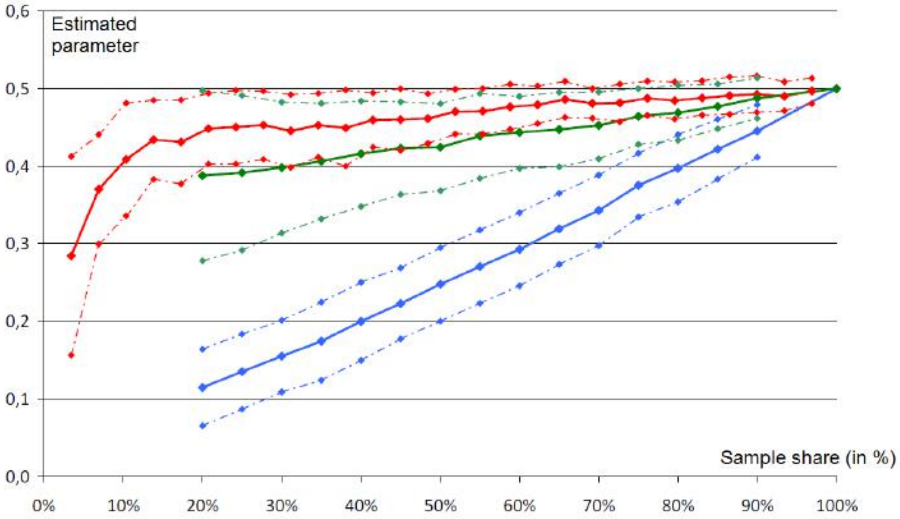

Figure 11.4 – "Size effect"
Note: Each point of a full-line curve represents an estimate of the parameter $\hat{\rho}$ for a sample size stated as a percentage of the comprehensive population. When the estimate is performed on exhaustive data, we find $\hat{\rho} = 0.5$. The blue curve is that of a simple random survey, the green curve a cluster sampling. The red curve reflects a deterministic selection of the regions, starting from an initial point, then moving away gradually. Dotted lines represent confidence intervals at 95%.

Previous estimations were the consequence of both the size and sparsity effects. To illustrate the former, we simulate data from a SAR model on the entire population and estimate spatial correlations on a sample of the most central NUTS3 n. The goal is to focus on a sub-section of

approach, we assume that there is a so-called *superpopulation* model, from which the Ys in the population are derived. Here, we are required to follow this approach in order to estimate our SAR models

[8] It should be noted that multiple parameters violate this law: for example, the maximum of variable Y on a population cannot be estimated without bias from a sample. Furthermore, in our case of simple random sampling or cluster sampling, there is no problem with under-coverage, *i.e.* units of the population that cannot belong to the sample for reasons often due to the quality of the registries. This angle cannot explain the bias on $\hat{\rho}$.

Europe, without letting it be randomly chosen or in a fragmented manner, as was the case with previous samples (Figure 11.2). Figure 11.4 compares the values of $\hat{\rho}$ resulting from three protocols for different percentages $P\%$ of the total population: the selection of the most central $P\%$ NUTS3; a cluster sampling in which each cluster of zones has $P\%$ chances of being selected; and a classic Poisson sampling, where each area independently has $P\%$ chances of being selected.

Just as in part 11.1.3, the Poisson sampling (similar to simple random sampling) yields estimated values $\hat{\rho}$ much lower than the cluster survey. The main contribution of this figure is in the red curve, which is based on a non-random selection of part of the zones. It converges quicker than the others toward 0.5, the true value of ρ. This appears to confirm the hypothesis of a bias linked to the distortion of the spatial structure or "size effect", resulting from a restriction to a subset of the total population.

11.1.5Robustness

To conclude this section, note that the choice of specification for the spatial model affects the results only marginally. The latter remain unchanged when the maximum distance threshold varies or when the concept of distance chosen is based on the closest neighbours (table 11.11 in Appendix 11.3.6). Lastly, the true value of parameter ρ does not affect the magnitude of the bias. Figure 11.5 shows that, at a given sampling rate, an estimate on a sample drawn by simple random sampling almost never makes it possible to find the true value of parameter ρ. In the case of a cluster survey, this value can be included in the confidence interval of the estimated parameter. However, the bias does not disappear when the magnitude or sign of this parameter vary. In any case, the bias mitigates the magnitude of the estimated spatial correlation.

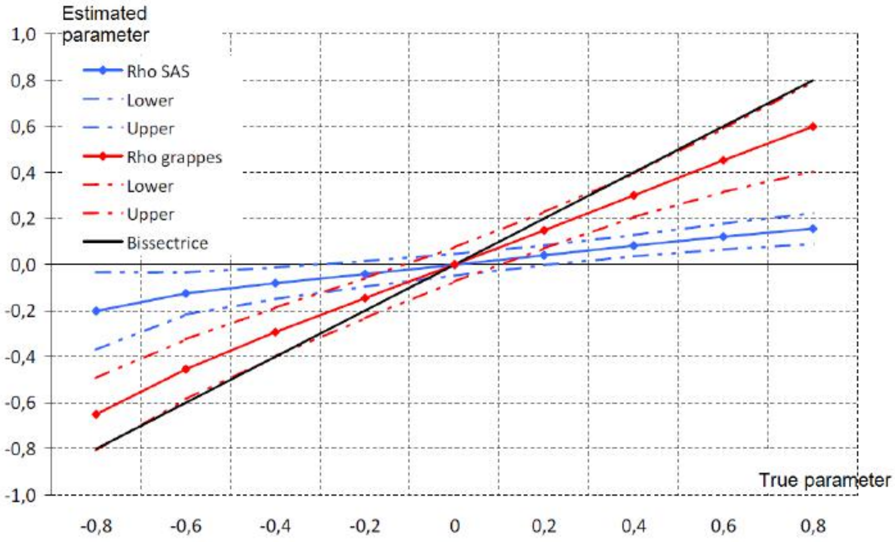

Figure 11.5 – Estimates of $\hat{\rho}$ for various values of ρ
Note: Full-line curves represent the estimated value $\hat{\rho}$ based on the effective value ρ set for data simulation. The blue curve shows the case of data driven by simple random sampling and the red curve the case of cluster sampling. Dotted curves represent 95% confidence intervals for estimator $\hat{\rho}$.

Lastly, considering a SEM model (*Spatial Error Model*): $Y_2 = X\beta + (1 - \lambda\mathbf{W})^{-1}\varepsilon$ does not radically affect the results (Table 11.12 in Appendix 11.3.6).

11.2 Prospects for resolution

One of the first positions one can adopt in the face of missing data is to ignore, consciously or not, this data and to directly apply the spatial model to the units observed. This mitigates the spatial correlation parameter relative to its true value, due to a "size effect" and a "sparsity effect".

The first effect comes from differences between the theoretical model and the estimated model regarding the dimension of the spatial weighting matrix. To remove it, the exhaustive data needs to be compared with the sample data according to a single geographical structure, and therefore on the same number of units. To compensate for the second, we need to be able to reconstruct the spatial correlations between observed and missing units. In this case, the location of units is always assumed to be known[9].

In this section, we discuss the impact of two solutions commonly applied to empirical work. First, switching to a higher scale by aggregating data and second, imputing missing data. Both methods maintain the geographical structure of the data, but are more or less effective in reconstructing spatial correlations.

11.2.1 Moving to a higher scale by aggregation

In the absence of comprehensive individual data, much research has been carried out on an aggregated scale of regions, departments or employment areas. This choice depends crucially on the relevant scale of the economic issue, assumes the availability of a good estimator of the local average and can lead to an ecological bias (see Anselin 2002b for further details). Intra-zone correlations are then omitted, to the benefit of correlations between zones.

To evaluate this solution, we simulate 6,000 points drawn from a uniform distribution on a square space and assign them, as in section 11.1, values of X and Yaccording to a SAR data generating process characterized by a spatial autocorrelation parameter $\rho = 0.5$. These points are depicted on the left-hand surface of Figure 11.6. Then, this square space is divided according to a grid of size $G \times G$ for different values of G, and to each centroid of each square is assigned the average of the points located inside this square. The centre and right panels of Figure 11.6 depict this configuration for $G = 50$ and $G = 20$ respectively.

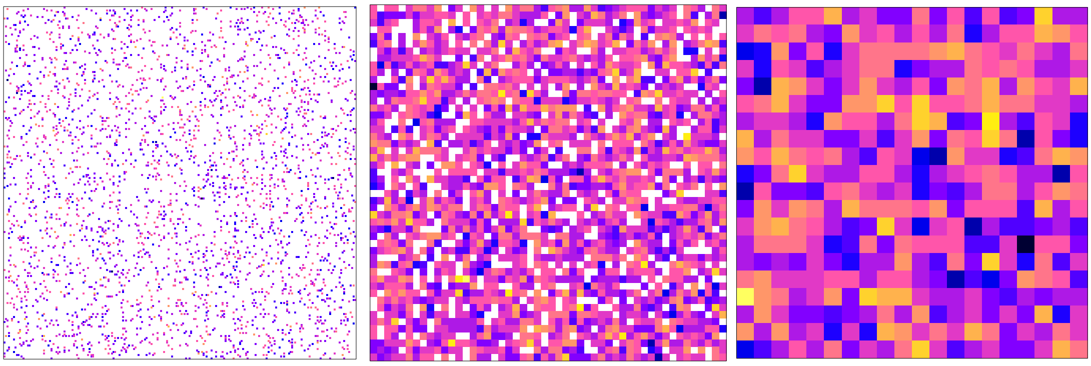

Figure 11.6 – Spatial data aggregation
Note: On the left panel, 6 000 items simulated from a uniform distribution, on the center panel, data aggregated in a grid of 50 × 50 squares and on the right panel, data aggregated in a grid of 20 × 20 squares

[9] The lack of information on the location of certain units is another challenge for current research on spatial econometrics (Arbia et al. 2016) which exceeds the scope of this chapter.

The estimation of a SAR model on comprehensive data aggregated with $G = 50$ provides a spatial correlation parameter $\hat{\rho} = 0.47$ with standard deviation of $\hat{\sigma}_\rho = 0.068$. This parameter is significantly positive and the estimate includes 0.5 in its confidence interval. Aggregating data on squares would limit the loss of spatial interactions and minimize the bias in estimating the spatial correlation parameter. Figure 11.7 shows that parameters ρ and β are estimated precisely and without bias when the grid on which the data is aggregated is relatively fine. The finer the grid, the closer we are to the spatial structure of exhaustive individual data and therefore, the closer the spatial correlation is to its true value.

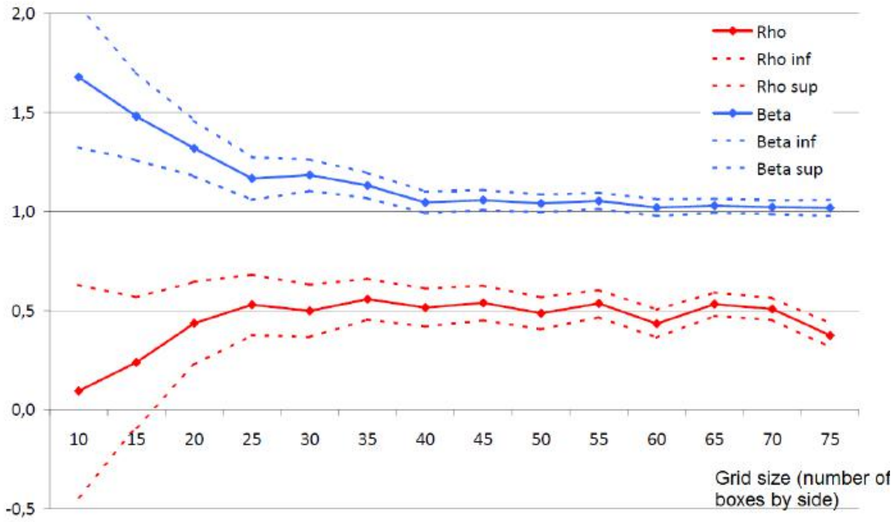

Figure 11.7 – Estimated parameters and fineness of the grid
Note: For each value of the size of the grid, the red curve displays the estimate $\hat{\rho}$ and the blue curve the estimate $\hat{\beta}$. Dotted curves stand for the 95% confidence intervals. Results from simulating 6,000 points from a uniform distribution.

Application to a sample

This procedure is replicated on data sampled by simple random sampling. The fineness of the grid fulfils the need for bias-variance arbitrage: fine squares reflect the distance between observations more accurately but lead to estimates of local averages that are less precise for each variable. Subject to assigning null weight and null values of the dependent and explanatory variables to squares without observation, we are able to identify the simulated spatial effect.

Table 11.4 shows the results of this procedure for different sample sizes and various spatial grids. In most cases, the true value of ρ is well within the confidence interval of the estimated parameter. For a small sample, an overly-coarse grid flattens out the spatial effects while an overly fine grid provides a poor estimate of individual variables. As before, the larger the sample, the more accurate the estimate.

These simulations tend to statistically validate the aggregation approach, provided that interpretation is not made directly at the individual level, but rests on strong hypotheses (coordinates of units drawn from a uniform distribution, homogeneous SAR process), rarely met in practice.

11.2.2 Imputing missing data

To stay on the scale of the available data, the solution is to impute values to missing observations. This is another way of ignoring the "size effect": ensuring consistency between the spatial structure

	$\hat{\rho}$				$\hat{\beta}$			
G n	10	30	50	60	10	30	50	60
100	0.487^{***}	0.494^{***}	0.483^{***}	0.478^{***}	1.016^{***}	1.007^{***}	1.027^{***}	1.030^{***}
	(0.070)	(0.060)	(0.068)	(0.072)	(0.134)	(0.115)	(0.129)	(0.134)
200	0.482^{***}	0.499^{***}	0.495^{***}	0.489^{***}	1.020^{***}	0.998^{***}	1.006^{***}	1.011^{***}
	(0.064)	(0.045)	(0.046)	(0.050)	(0.126)	(0.084)	(0.088)	(0.095)
500	-0.093	0.488^{***}	0.483^{***}	0.489^{***}	1.035^{***}	1.022^{***}	1.031^{***}	1.021^{***}
	(0.701)	(0.032)	(0.030)	(0.032)	(0.121)	(0.060)	(0.055)	(0.059)
1000	-0.982	0.487^{***}	0.485^{***}	0.491^{***}	1.048^{***}	1.024^{***}	1.028^{***}	1.019^{***}
	(0.159)	(0.024)	(0.020)	(0.021)	(0.119)	(0.045)	(0.038)	(0.040)

Table 11.4 – SAR model estimated on data aggregated by squares of the grid
Note: Each line reflects the size n of the sample drawn from the 6,000 simulated points and each column reflects the fineness of the grid in terms of number of squares (a size 30 grid breaks the initial square in 900 boxes). *** denotes significance at 1%.

of survey data and administrative data. In the face of missing values in a survey or census, assigning "plausible" values to these units makes it possible to have a sample or even a complete population.

Imputation methods

This section lists a number of traditional imputation methods. Interested readers may refer to a comprehensive handbook on survey theory, for example Ardilly 1994 or Tillé 2001, which provides more information, theoretical context, as well as other more advanced methods. In the case of imputation by ratio or by hot deck, explanatory variables X are assumed to be known exhaustively.

Imputation by the mean. The method of imputation by the mean (or by the median, or by the dominant class in the case of qualitative variables) is a common method consisting of replacing all missing values by the mean of the observed values. This method does not respect a possible econometric structure between different variables of the survey and may lead to false results in the estimation of such models.

Imputation by ratio. The attribution by ratio method involves mobilising the auxiliary information X available on the entire population, including the units for which the information of interest Y is missing, in order to impute plausible Y values. To do this, we assume the existence of a linear model $Y = \beta X + \varepsilon$. $\hat{\beta}$ is estimated by ordinary least squares, after which the value $Y_{\text{ratio}} = \hat{\beta}X$ is imputed for the missing Y. The ratio of the Y over the X, in the case of quantitative data, is the same between the units observed and the units for which no information is available. This method can be refined by adding constraints to the units for which the estimate of β is calculated, for example on a specific domain or stratum.

Hot deck imputation. The hot deck method randomly connects a donor to a missing value, in contrast to the cold deck, which establishes this link deterministically. A donor here is an individual statistically "close" to the missing individual (they share similar values of auxiliary variable X, belong to the same stratum, to the same domain, or possibly are located in the same spatial position). The application of hot deck is based on the definition of a distance criterion, from which k neighbours of the valueless individual Y are determined. One individual is randomly selected from the k neighbours, uniformly or otherwise, to give its value to the new Y_{hotdeck}. Variants can be introduced, for example by limiting the number of times a single individual can be a donor, or by performing the hot deck sequentially.

We illustrate the proposed methods with a simple example. We simulate the geographic position of $N = 1,000$ points to which we assign variables X and Y following a SAR structure with $\beta = 1$ and $\rho = 0.5$. We then draw samples by simple random sampling for different sizes n. For each sample, the $N - n$ units not drawn are imputed by one of the methods mentioned above: imputation by the X ratio, imputation by statistical hot deck (neighbours have values close to X) and imputation by geographic hot deck (neighbours are spatially close). Table 11.5 compares the results of these different methods to direct exploitation of sample.

	Direct		**Ratio**		**Statistical**		**Geographic**	
					hot deck		**hot deck**	
n	ρ	β	ρ	β	ρ	β	ρ	β
100	0.06	1.14***	0.05***	1.13***	0.03	1.06***	0.19***	0.12**
	(0.06)	(0.14)	(0.01)	(0.12)	(0.03)	(0.13)	(0.05)	(0.05)
200	0.09*	1.10***	0.11***	1.11***	0.06**	1.08***	0.22***	0.22***
	(0.04)	(0.08)	(0.02)	(0.08)	(0.02)	(0.09)	(0.04)	(0.05)
500	0.19***	1.08***	0.25***	1.09***	0.19***	1.09***	0.31***	0.55***
	(0.02)	(0.04)	(0.02)	(0.05)	(0.02)	(0.05)	(0.03)	(0.04)

Table 11.5 – Imputation methods
Note: Parameter ρ of SAR model, estimated by Monte Carlo, for a 100-size sample, after imputation by ratio, is 0.05, with empirical standard deviation of 0.01. *** indicates a 1% significance, ** 5% and * 10%.

The choice of method has a significant impact on the results. Imputation by ratio seems to work well for both parameters even though it underestimates parameter ρ for small samples. On the other hand, the geographic hot deck method gives good results on the auto-correlation parameter but implies a very strong bias on parameter β. Finally, the statistical hot deck method seems to give similar results to the direct estimation on the sample. These results illustrate these methods on a very simple example and show their inability to find the initial parameters of the model.

Further reading

Imputation methods may bias estimates. In particular, the link between Y and X on which the imputation is based can be passed on to parameters estimated from a regression of Y on X (see Charreaux et al. 2016 for a discussion on this point). Similarly, imputation methods can create a spatial structure *ex-nihilo* or, to the contrary, break the spatial correlations which are not taken into account.

Lastly, as mentioned in introduction, more refined methods of imputation using EM algorithms have been developed (LeSage et al. 2004; Wang et al. 2013a). However, they are complex, very specific to the type of information missing and still remain rarely applied.

11.3 Empirical application: the manufacturing sector in Bouches-du-Rhône

In order to illustrate the issues of estimating spatial models on survey data, we estimate a production function on plants from the manufacturing sector. The spatial approach makes it possible to measure the impact of interactions between the production processes of selected firms. Such *spillovers* between firms have already been highlighted by a significant literature on conurbation economies (see in particular LeSage et al. 2007,Ertur et al. 2007, López-Bazo et al. 2004, Egger et al. 2006).

11.3.1 Data

The SIRUS register (identification system in the register of statistics units [10]) lists all French firms, groups and their entities, contained in SIRENE (computerised system for the national register of enterprises and institutions [11]), the administrative register used for registering legal units. For each firm, information on turnover, main activity (accessible *via* the APE code (principal activity code), following the French nomenclature), total balance sheet, exports (administrative and full-time equivalent), physical address and list of plants is available.

Using the geographical information available (cadastral reference, road or city centre), the (x,y) coordinates of each entity have been successfully geolocated by the INSEE Methodology and Geographical Guidelines Division. This geographical data, combined with the economic data available in the SIRUS register, makes it possible to model econometric relations by taking into account the spatial interactions.

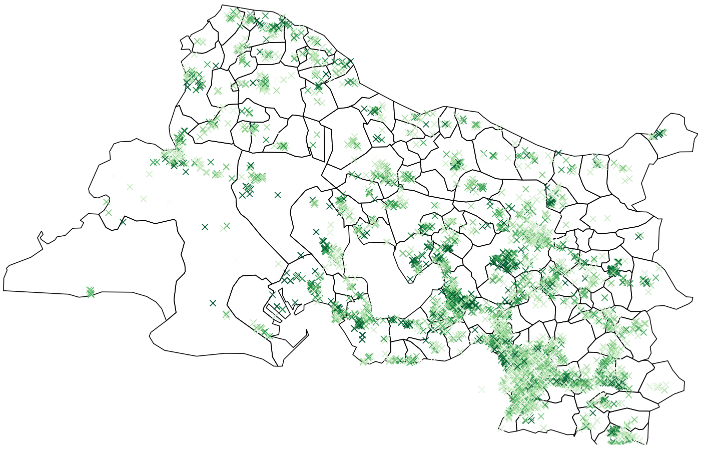

Figure 11.8 – Industrial companies in Bouches-du-Rhône
Source: *SIRUS register, 2015*
Scope: Companies of the manufacturing sector in Bouches-du-Rhône
Note: Darker green matches higher turnover

11.3.2 Identification

The production of a firm can be influenced by the geographical proximity of neighbouring companies. These interactions are designated by the concept of "externalities" which can be positive when the neighbourhood has a favourable impact on production (complementarities between sectors, integration of production chains, relationship with suppliers, transport, sharing of knowledge, etc.) or negative when they damage production (competition, pollution, traffic jams, etc.).

First of all, in order to carry out this analysis, it is necessary to choose coherent field, type of spatial links, and territory, so that the units considered maintain relations between them but not (or little) with the outside. In this case, we assume to have exhaustive data as regards the spatial effects on this territory. The choice to study the manufacturing sector in the Bouches-du-Rhône is didactic, but not meaningless. The presence of the port of Fos-sur-mer, the axis of the Rhône valley and the road nodes towards Toulouse and Italy make it a territory of interest (Figure 11.8 clearly illustrates the establishment of plants according to the transport networks). Of course, this estimate will not take national or even international trade relations into account, but will be exhaustive concerning local relations.

[10] In French, *Système d'Identification au Répertoire des Unités Statistiques*.

[11] In French, *Système Informatisé du REpertoire National des Entreprises et des établissements*.

The production level Y_i of a plant i can be stated according to a Cobb-Douglas production function: $Y_i = AL_i^{\beta_L} K_i^{\beta_K}$, where workforce L_i and capital K_i are input factors while A refers to general productivity of factors. Parameters β_L and β_K represent, respectively, the share of earned income and capital in production[12]. Traditionally, the term A refers to all mechanisms that influence production (human capital, technological progress, complementarities...) without being directly measurable. It can also be conceived of as representing the positive production externalities and be written: $A = \exp(\beta_0) \prod_{j \in v_i} Y_j^{\rho \omega_{ij}}$, where v_i refers to the neighbours of plant i, and Y_j the production level of a neighbouring plant j. The production function may be expressed in logarithm as:

$$\log(Y_i) = \beta_0 + \rho \sum_{j \in v_i} \omega_{ij} \log(Y_j) + \beta_L \log(L_i) + \beta_K \log(K_i) + \varepsilon_i \tag{11.3}$$

or:

$$\widetilde{Y} = \beta_0 + \rho \mathscr{W} \widetilde{Y} + \beta_L \widetilde{L} + \beta_K \widetilde{K} + \varepsilon$$

Terms with a tilda refers to the logarithm of these variables. $\mathscr{W}$ is the spatial weighting matrix, such that $\mathscr{W}_{i,j} = \omega_{ij}$. This equation may be estimated using a SAR model. The parameter of spatial autocorrelation ρ captures the complementarities common to all units while ω_{ij} captures specific complementarities resulting from the impact of the production of plant j on the production of plant i. The term $\rho \omega_{ij}$ refers to the elasticity of the plant i's production with respect to the production of plant j: when a plant j neighbouring i increases production by 1%, the production level of plant i increases by $\rho \omega_{ij}$%through direct effects. Deriving equation 11.3 with respect to $\log(Y_k)$, we have:

$$\frac{d\log(Y_i)}{d\log(Y_j)} = \underbrace{\rho \omega_{ij}}_{\text{Direct effect from j on i}} \underbrace{+\rho \sum_{k \neq j} \omega_{ik} \frac{d\log(Y_k)}{d\log(Y_j)}}_{\text{Indirect effect } via \text{ k}}$$

Likewise, summing this expression over j, ρ appears as the direct elasticity of plant i's production with respect to the production of neighbouring plants:

$$\sum_j \frac{d\log(Y_i)}{d\log(Y_j)} = \rho + \rho \sum_j \sum_{k \neq j} \omega_{ik} \frac{d\log(Y_k)}{d\log(Y_j)}$$

11.3.3 Estimation

The equation 11.3 is estimated over 6,306 plants geolocated in Bouches-du-Rhône, belonging to the manufacturing sector[13]. This sector is particularly appropriate to a spatial estimate, since the geographical location is not directly part of its production activities (unlike trade, transport or agriculture), it is not too concentrated (as is the case in high technology) and does not make particular use of network logics other than spatial (contrary to finance or communications for instance).

Production Y_i from plant i is given by its turnover. The total balance sheet of the plant, which is a measure of its assets, is used as a proxy for capital K_i. These two variables, which are only available at firm level, are divided by the number of plants within the company. Lastly, workforce L_i is available at plant level in SIRUS.

Figure 11.8 shows the location of these plants. The intensity of the green in these cross-signs denotes their turnover: the darker the colour, the higher the turnover. Groups of plants with strong

[12] These parameters can also be interpreted, respectively, as the elasticities of production with respect to work and capital.

[13] The manufacturing sector encompasses companies whose main business belongs to Divisions 10 to 33 of NAF Rev 2. 2008.

turnover appear clearly, for example near Aix-en-Provence or around Fos-sur-Mer. As in the simulations in section 11.1, neighbourhood proximity is depicted by a weighting matrix based on distance. According to our definition, each plant has on average 109 neighbours, and 76 plants do not have neighbours [14].

β_0	β_L	β_K	ρ
0.422	0.535	0.769	0.051
(0.050)	(0.015)	(0.009)	(0.009)

Table 11.6 – SAR model estimation: all plants
Source: *SIRUS register, 2015*
Scope: All plants from the manufacturing sector located in the Bouches-du-Rhône department, whose turnover and total balance sheet are strictly positive
Note: Estimated parameters are significant at 1%.

Table 11.6 shows the results of the SAR model estimated on all of plants within the manufacturing sector in Bouches-du-Rhône. Labour income and capital income shares are close to values generally estimated (roughly one-half to two-thirds for the latter, and one-third to two-thirds for the second, the high marginal return on capital being attributable to the choice of the industrial sector). There is a positive and significant spatial correlation: when the average turnover of the neighbouring units of plant *i* increases by 1%, the turnover of plant *i* increases by 0.05%.

11.3.4 Spatial estimates on samples

Sampling plans

As in Section 11.1, we replicate the model's estimate 11.3 on a sample of plants. Simple random sampling is used as a reference point, but is not common in the context of corporate surveys. Stratified sampling methods are more frequently employed in studies identifying the effect of the labour force and capital on turnover. These sampling methods have been presented in Section 11.1.2.

Stratification is carried out according to this workforce variable, assuming a correlation between headcount and turnover. Table 11.7 shows the strata created using a Neyman Allocation, based on dispersion of turnover within each of the strata. The dispersion within Stratum 4 is much higher than that of the other strata. As a consequence, we consider Stratum 4 to be exhaustive, *i.e.* we will always sample the 67 companies of Stratum 4 in order to limit the variance of estimation.

Results

In this section, we compare the results secured using a simple and stratified random sampling plan, varying the sample size: $n \in \{250, 500, 1000, 2000\}$.

Table 11.8 shows the parameters of the SAR model estimated from 1,000 draws by simple random sampling (*on the left*) and stratified sampling (*on the right*). In the case of simple random sampling, as well as in section 11.1, the traditional regression parameters β_L and β_K are correctly estimated. In contrast, spatial correlation parameter ρ is significant only for a sample size greater than 1,000 and always remains lower than the value it would take on full data.

The stratified survey plan applied to non-reweighted data skews traditional estimators β_L and β_K when regression is unweighted (Davezies et al. 2009). On the other hand, the bias on spatial

[14] The units without neighbours, also referred to as "islands", do not participate in estimating the spatial correlation parameter ρ. The threshold is determined by a trade-off between minimising the number of neighbours and the number of islands.

Number of strata	Number of employees	Number of companies
1	0	3 628
2	1 to 9	2 742
3	10 to 99	770
4	100 and more	67

Table 11.7 – Constitution of strata
Source: *SIRUS register, 2013*
Scope: All plants from the manufacturing sector located in Bouches-du-Rhône, whose turnover and total balance sheet are strictly positive

	Simple random sampling			Stratified sample		
n	ρ	β_L	β_K	ρ	β_L	β_K
250	0.011	0.554***	0.768***	0.015*	0.311***	0.813***
	(0.021)	(0.104)	(0.078)	(0.009)	(0.072)	(0.056)
500	0.017	0.545***	0.773***	0.020**	0.371***	0.796***
	(0.016)	(0.073)	(0.052)	(0.008)	(0.053)	(0.041)
1000	0.024**	0.542***	0.774***	0.024***	0.410***	0.793***
	(0.012)	(0.051)	(0.039)	(0.007)	(0.039)	(0.029)
2000	0.034***	0.541***	0.770***	0.036***	0.457***	0.790***
	(0.010)	(0.031)	(0.023)	(0.007)	(0.028)	(0.022)

Table 11.8 – Model 11.3 estimated on a random sample and on a stratified sample
Source: *SIRUS register, 2015*
Scope: All plants from the manufacturing sector located in Bouches-du-Rhône, whose turnover and total balance sheet are strictly positive

correlation parameter ρ appears less. This is because large companies likely to have significant spatial influence are all taken into account in the sample, due to this stratified survey plan.

The decision not to weight the regression is made by default. In traditional econometrics, it is of use to weigh observations before estimating an econometric model when the structure of the sampling plan is linked to the estimated variables. However, the question of using sampling weight as part of a SAR model has not been definitively addressed by the current literature [15]. In the current state of affairs, unweighted regression appears to be the safest and easiest choice. We do not explore this question further in this chapter.

11.3.5 Estimation on aggregated data

As discussed in section 11.2, an approach commonly used to circumvent the problem of missing data consists in moving to a wider scale by aggregating the sampled data. In order to move away from the administrative zonings, we divided the Bouches-du-Rhône department according to a $G \times G$ grid (Figure 11.9).

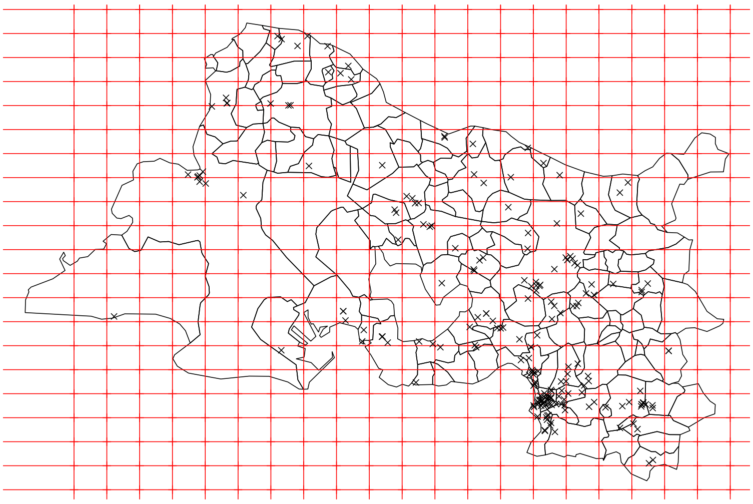

Figure 11.9 – Bouches-du-Rhône breakdown in a grid of 20 × 20 squares
Source: *SIRUS register, 2015*
Scope: Plants from the manufacturing sector in Bouches-du-Rhône

From this grid, observations of a sample are averaged on each cell and the spatial analysis is carried out at the scale of the grid, with the distance being defined between the centroids of the cells. Null values are assigned to the variables and spatial weights of cells without observation, thus excluding them from the estimate without distorting the size of the spatial weighting matrix. Table 11.9 shows parameter ρ estimated for different sample sizes and various grid sizes.

Estimating spatial models based on aggregated data appears to circumvent the problem of missing data within a very simple context of data simulated uniformly across a territory. However, the application of this method to actual data is not straightforward. In particular, in this specific case, the parameter of spatial autocorrelation is still underestimated and is never significant. This could be due to the high concentration of plants in Bouches-du-Rhône, as the intra-cell distances are not, by definition, taken into account in this method. Spatial estimates on aggregated data thus require that the estimated phenomenon not be specific to a finer geographical scale.

[15] For example, it is not clear whether it is necessary to involve the sampling weights in the spatial weighting matrix calculation **W**; this could also induce additional endogeneity, linked to the sample structure.

n \ G	20	30	50	60
100	0.007	0.009	0.014	0.015
	(0.018)	(0.022)	(0.022)	(0.024)
200	0.013	0.007	0.015	0.018
	(0.021)	(0.019)	(0.018)	(0.018)
500	0.024	0.023	0.012	0.013
	(0.031)	(0.023)	(0.014)	(0.013)
1000	0.031	0.057*	0.021*	0.014
	(0.026)	(0.040)	(0.015)	(0.012)

Table 11.9 – Parameter ρ estimated on aggregated data
Source: *SIRUS register, 2015*
Scope: Plants from the manufacturing sector in Bouches-du-Rhône
Note: Parameters are estimated on a size-n sample aggregated on a $G \times G$ grid. For the sake of clarity, we show here only parameter values ρ.

11.3.6 Imputation of missing data

Implementation

The second approach, referred to in section 11.2.2, consists in imputing the missing data, *i.e.* to attribute estimated Y_i values to plants for which none are available. We consider three types of imputations for Bouches-du-Rhône plants: *(i) imputation by ratio*, which uses variables L and K representing workforce and capital as explanatory variables of the model, *(ii)* imputation by *statistical hot deck*, in the sense that the distance is calculated based on the values of L and K, that is to say that the neighbours of a plant are the plants that share similar staff and capital and *(iii)* imputation by *geographic hot deck*, where one is associated with an individual in the geographic sense.

The implementation of these techniques requires, in the first case, to estimate a linear model (function `lm` in R), and in the next two cases, to define the neighbours (function `knn` of package *class*), then randomly make a draw amongst them (function `sample` in R). These three approaches are tested on the manufacturing sector in Bouches-du-Rhône. 1,000 size-n samples are drawn according to the principles of simple random sampling, after which the imputation process assigns values Y to the $N-n$ companies not sampled. The results obtained are presented in table 11.10. As a reminder, the results on the entire population can be found in table 11.8.

Results

The results are highly variable, depending on the method used. Imputation by ratio makes it possible to maintain the linear structure between turnover, workforce and capital, resulting in unbiased and accurate estimates for coefficients β_L and β_K. On the other hand, estimation of ρ is very low, even more so than in the case of a direct estimation on a random sample (see table 11.8). Imputation does not take into account the spatial structure, which is deleted when the model is estimated on the data completed. Therefore, it is not relevant to apply spatial econometric models to data imputed using this method.

Imputation by statistical hot deck appears more promising. The estimators are on the right order of magnitude with respect to the resulting values for the population and are estimated accurately. A comparison with table 11.6 reveals a bias when $\hat{\rho}$, $\hat{\beta}_L$ and $\hat{\beta}_K$ are estimated on small samples. Thus, the imputation by hot deck skews the estimators (Charreaux et al. 2016) but enables the structure of the spatial correlations to be brought out. This is because the connection between donor

	Ratio			**Statistical hot deck**			**Geographic hot deck**		
n	ρ	β_L	β_K	ρ	β_L	β_K	ρ	β_L	β_K
250	0.002	0.560^{***}	0.768^{***}	0.043^{***}	0.664^{***}	0.646^{***}	0.419^{***}	0.028	0.104^{***}
	(0.002)	(0.112)	(0.080)	(0.009)	(0.083)	(0.059)	(0.046)	(0.034)	(0.023)
500	0.004	0.548^{***}	0.774^{***}	0.042^{***}	0.613^{***}	0.698^{***}	0.412^{***}	0.061^{*}	0.149^{***}
	(0.003)	(0.077)	(0.058)	(0.008)	(0.061)	(0.044)	(0.035)	(0.034)	(0.022)
1000	0.008^{**}	0.546^{***}	0.774^{***}	0.040^{***}	0.577^{***}	0.734^{***}	0.389^{***}	0.116^{***}	0.217^{***}
	(0.003)	(0.051)	(0.037)	(0.007)	(0.040)	(0.028)	(0.035)	(0.035)	(0.023)
2000	0.017^{***}	0.542^{***}	0.773^{***}	0.040^{***}	0.562^{***}	0.751^{***}	0.333^{***}	0.203^{***}	0.338^{***}
	(0.004)	(0.032)	(0.024)	(0.007)	(0.031)	(0.023)	(0.022)	(0.034)	(0.022)

Table 11.10 – Imputation methods
Source: *SIRUS register, 2015*
Scope: Plants from the manufacturing sector in Bouches-du-Rhône, whose turnover and total balance sheet are strictly positive

and recipient implicitly maintains the structure of spatial interactions. It is also possible that the underlying spatial structure Y also exists for L and K and is recovered by imputation. Thus, the use of this imputation method for econometric analysis involves a trade-off between bias and variance on parameters β_L and β_K, traditional in sampling theory. However, in this instance, the method also offers the advantage of substantially reducing the pre-existing bias on ρ. These results, which are tested only on this data set and with a simple sampling plan, are to be used with caution. In any case, the effectiveness of this method is not based on the spatial proximity between the donor and the recipient, as the last example shows.

The imputation method using geographic hot deck results in aberrant estimations. Based directly on the spatial proximity between donor and receiver, it gives rise to very strong overestimation of the spatial effect (the estimated $\hat{\rho}$ is greater than its value estimated on the full population), at the expense of the effect of other variables in the model (estimated $\hat{\beta}$'s are well below the values of these parameters estimated on the full population). In fact, according to this method, nearby spatial structures will have the same turnover Y, which creates *ex-nihilo* a very high positive spatial correlation. The use of the spatial dimension to remedy the problem of missing data is not for the immediate future. Table 11.13 in Appendix 11.3.6 shows the results achieved for a geographical hot deck imputation by limiting itself to plants with similar workforce. Parameter ρ is less overestimated but the results remain far removed from the estimate based on full data. It may be possible to use geographic information parsimoniously for imputation, but this would require more extensive analysis of the data set and good knowledge of its spatial structure.

Conclusion

This chapter highlights the difficulties associated with the application of spatial econometrics to sampled data. There are two pitfalls in particular: *(i)* a "size effect" by which the estimate on a remote sample distorts the spatial weighting matrix, and *(ii)* a "sparsity effect" resulting from the omission of units spatially correlated with the units observed. Both effects tend to underestimate the magnitude of spatial correlations. However, the bias is lower in the case of a cluster survey and when the sample is larger.

Empirical studies typically resolve this problem by ignoring missing observations, aggregating data on a larger scale or imputing the missing values. The first solution is never desirable. The other two are far from perfect, because it is hard to rebuild a complex set of information from few observations. The imputation by statistical hot deck is promising, but we do not demonstrate its validity in a general case.

While this issue is bound to become more important as social media and geolocated data become more prominent, estimating spatial models based on sampled data remains rare. For the time being, it is preferable to consider comprehensive data. This chapter warns against overly-expeditious solutions, such as aggregating data on a higher scale, calling upon simplistic imputation methods or removing any mention of missing data. When a relatively large sample is available, or derived from a cluster survey, a spatial estimate could then be considered, bearing in mind that the resulting spatial correlation parameter will probably be underestimated.

Appendix

Choice of model and neighbourhood matrix

Here, we consider the same procedure as in section 11.1 for different values of the parameters. Table 11.11 shows estimates comparables to those from Table 11.2 for different neighbourhood matrices. Table 11.12 displays results in the case of a Spatial Error Model (SEM) instead of a SAR model.

	ρ			β		
$\mathscr{M}$ n	2 neighbours	5 neighbours	Distance	2 neighbours	5 neighbours	Distance
50	0.020	-0.003	0.042	1.107^{***}	1.050^{***}	1.054^{***}
	(0.110)	(0.172)	(0.043)	(0.115)	(0.095)	(0.125)
100	0.063	0.069	0.058^{*}	1.112^{***}	1.056^{***}	1.054^{***}
	(0.076)	(0.111)	(0.031)	(0.079)	(0.065)	(0.086)
150	0.097^{*}	0.115	0.073^{**}	1.107^{***}	1.052^{***}	1.049^{***}
	(0.060)	(0.088)	(0.028)	(0.062)	(0.051)	(0.068)
250	0.150^{***}	0.189^{**}	0.101^{**}	1.105^{***}	1.050^{***}	1.053^{***}
	(0.047)	(0.065)	(0.026)	(0.049)	(0.040)	(0.052)

Table 11.11 – SAR Model - Monte Carlo Estimation
Source: *SIRUS register, 2015*
Scope: Plants from the manufacturing sector in Bouches-du-Rhône
Note: Standard deviations are in brackets.

	λ			β		
$\mathscr{M}$ n	2 neighbours	5 neighbours	Distance	2 neighbours	5 neighbours	Distance
50	-0.025	-0.110	0.008	1.003^{***}	1.003^{***}	1.002^{***}
	(0.167)	(0.287)	(0.193)	(0.115)	(0.113)	(0.112)
100	0.008	-0.027	0.024	1.003^{***}	1.004^{***}	1.003^{***}
	(0.113)	(0.182)	(0.124)	(0.080)	(0.078)	(0.078)
150	0.023	0.002	0.034	0.998^{***}	0.998^{***}	0.998^{***}
	(0.090)	(0.144)	(0.099)	(0.065)	(0.063)	(0.063)
250	0.047	0.042	0.052	1.000^{***}	1.000^{***}	1.000^{***}
	(0.069)	(0.108)	(0.079)	(0.051)	(0.050)	(0.050)

Table 11.12 – SEM Model - Monte Carlo Estimation
Source: *SIRUS register, 2015*
Scope: Plants from the manufacturing sector in Bouches-du-Rhône
Note: Standard deviations are in brackets.

Imputation by stratified geographical hot deck

Table 11.13 provides the results using geographical hot deck imputation, restricted to plants with similar workforce, *i.e.* those of the same stratum (defined in table 11.7) as the plant with a missing value.

	Geographic stratified hot deck		
n	ρ	β_L	β_K
250	0.137	1.216	0.029
	(0.037)	(0.100)	(0.026)
500	0.148	1.192	0.071
	(0.031)	(0.077)	(0.025)
1000	0.156	1.121	0.149
	(0.026)	(0.061)	(0.025)
2000	0.148	0.542	0.279
	(0.019)	(0.048)	(0.025)

Table 11.13 – Imputation by geographic stratified hot deck
Source: *SIRUS register, 2015*
Scope: Plants from the manufacturing sector in Bouches-du-Rhône

References - Chapter 11

Anselin, Luc (2002b). « Under the hood : Issues in the specification and interpretation of spatial regression models ». *Agricultural Economics* 27.3, pp. 247–267.

Arbia, Giuseppe, Giuseppe Espa, and Diego Giuliani (2016). « Dirty spatial econometrics ». *The Annals of Regional Science* 56.1, pp. 177–189.

Ardilly, Pascal (1994). *Les techniques de sondage*.

Belotti, F., G. Hughes, and A. Piano Mortari (2017a). « Spatial panel-data models using Stata ». *Stata Journal* 17.1, 139–180(42).

Boehmke, Frederick J., Emily U. Schilling, and Jude C. Hays (2015). *Missing data in spatial regression*. Tech. rep. Society for Political Methodology Summer Conference.

Burt, Ronald S. (1987). « A Note on Missing Network Data in the General Social Survey ». *Social Networks* 9, pp. 63–73.

Charreaux, C et al. (2016). « Econométrie et Données d'Enquête: les effets de l'imputation de la non-réponse partielle sur l'estimation des paramètres d'un modèle économétrique ».

Chow, Gregory C. and An-Loh Lin (1976). « Best Linear Unbiased Estimation of Missing Observations in an Economic Time Series ». *Journal of the American Statistical Association* 71.355, pp. 719–721.

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

Cochran, William G (2007). *Sampling techniques*. John Wiley & Sons.

Davezies, L. and X. D'Haultfoeuille (2009). *To Weight or not to Weight? The Eternal Question of Econometricians facing Survey Data*. Documents de Travail de la DESE - Working Papers of the DESE g2009-06. INSEE, DESE.

Dempster, A.P., N.M. Laird, and D.B. Rubin (1977). « Maximum likelihood from incomplete data via the EM algorithm ». *Journal of the royal statistical society* 39.1, pp. 1–38.

Egger, Peter and Michael Pfaffermayr (2006). « Spatial convergence ». *Papers in Regional Science* 85.2, pp. 199–215.

Ertur, Cem and Wilfried Koch (2007). « Growth, technological interdependence and spatial externalities: theory and evidence ». *Journal of Applied Econometrics* 22.6, pp. 1033–1062.

Ferreiro, Osvaldo (1987). « Methodologies for the estimation of missing observations in time series ». *Statistics and Probability Letters* 5.1, pp. 65–69.

Gile, Krista J. and Mark S. Handcock (2010). « Respondent-driven sampling: an assessment of current methodology ». *Sociological Methodology* 40.1, pp. 285–327.

Goulard, M., T. Laurent, and C. Thomas Agnan (2013). « About predictions in spatial autoregressive models: Optimal and almost optimal strategies ». *Toulouse School of Economics Working Paper* 13, p. 452.

Harvey, A. C. and R. G. Pierse (1984). « Estimating Missing Observations in Economic Time Series ». *Journal of the American Statistical Association* 79.385, pp. 125–131.

Huisman, Mark (2014). *Imputation of missing network data*. Ed. by Reda Alhajj and Jon Rokne. Vol. 2. Springer, pp. 707–715. ISBN: 978-1-4614-6169-2.

Jones, Richard H. (1980). « Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations ». *Technometrics* 22.3, pp. 389–395.

Kelejian, H.H. and I.R. Prusha (2010). « Spatial models with spatially lagged dependent variables and incomplete data ». *Journal of geographical systems*.

Koskinen, Johan H., Garry L. Robins, and Philippa E. Pattison (2010). « Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation ». *Statistical Methodology* 7.3", pp. 366–384.

Kossinets, Gueorgi (2006). « Effects of missing data in social networks ». *Social Networks* 28.3, pp. 247–268.

LeSage, James P., Manfred M. Fischer, and Thomas Scherngell (2007). « Knowledge spillovers across Europe: Evidence from a Poisson spatial interaction model with spatial effects ». *Papers in Regional Science* 86.3, pp. 393–421. ISSN: 1435-5957.

LeSage, J.P. and R.K. Pace (2004). « Models for spatially dependent missing data ». *The journal of real estate finance and economics* 29.2, pp. 233–254.

Little, Roderick J. A. (1988). « Missing-Data Adjustments in Large Surveys ». *Journal of Business and Economic Statistics* 6.3, pp. 287–296.

Little, Roderick J. A. and Donald B. Rubin (2002). *Statistical analysis with missing data*. 2nd. Wiley, Hoboken.

Liu, Xiaodong, Eleonora Patacchini, and Edoardo Rainone (2017). « Peer effects in bedtime decisions among adolescents: a social network model with sampled data ». *The Econometrics Journal*.

López-Bazo, Enrique, Esther Vayá, and Manuel Artís (2004). « Regional Externalities And Growth: Evidence From European Regions ». *Journal of Regional Science* 44.1, pp. 43–73.

Mardia, Kanti V. et al. (1998). « The Kriged Kalman filter ». *Test* 7.2, pp. 217–282.

Pinkse, Joris and Margaret E. Slade (2010). « The Future of Spatial Econometrics ». *Journal of Regional Science* 50.1, pp. 103–117.

Rubin, Donald B. (1976). « Inference and missing data ». *Biometrika* 63, pp. 581–592.

Stork, Diana and William D. Richards (1992). « Nonrespondents in Communication Network Studies ». *Group & Organization Management* 17.2, pp. 193–209.

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies: cours et exercices avec solutions:[2e cycle, écoles d'ingénieurs]*. Dunod.

Wang, W. and L.-F. Lee (2013a). « Estimation of spatial autoregressive models with randomly missing data in the dependent variable ». *The Econometrics Journal* 16.1, pp. 73–102.

Zhou, Jing et al. (2017). « Estimating Spatial Autocorrelation With Sampled Network Data ». *Journal of Business and Economic Statistics* 35.1, pp. 130–138.