Università della Svizzera Italiana

Faculty of Economics

Doctoral Thesis

Statistical solutions for regressions-type models with big spatial data

Advisor: **Prof.ssa Antonietta Mira**

Co-advisor: **Prof. Giuseppe Arbia**

Candidate:

Chiara Ghiringhelli

Academic year : 2019 - 2020

Contents

List of Figures

List of Tables

Introduction

These last decades have witnessed an explosion of data collection and diffusion in all fields of human society. In many scientific fields researchers are becoming aware of a *big data* problem and of the need to manage it properly, combining the tools offered by statistics, computer, data science and informatics. This is particularly true for geo-located data, which can be collected in a comprehensive manner thanks to new technologies. Spatial econometric methods provide the right environment for spatial data modelling to identify causal mechanism and to assist empirically-supported decisions. However, such models become computationally prohibitive even with increasing power computing machines, when applied to very large datasets. The main features of big data are the so-called five V: volume, velocity, variety, veracity and value. All these characteristics can be easily detectable in the new geo-located data, instantly provided in a huge amount by smartphones, wearable, websites or other technologies.
From this perspective, in this thesis, we aim to identify the limits of the current routines and the issues which could occur dealing with big spatial dataset. Since, we are specifically interested in the study of causal relationships between phenomena through regression-type models, our entire research focuses on the most common spatial econometric models: the Spatial Lag Model and the Spatial Error Model.

We firstly investigate the main feature of big data: the volume. Are the current spatial econometric procedures able to handle such a massive quantity of data? We aim at evaluating the performances of the currently available estimation methods, based not only on an increasing number of samples, but also on the value of the parameters and the definition of the spatial topology.
In Chapter 1, *Estimation of Spatial Econometric linear models with large datasets: How big can spatial Big Data be?*, we report the results of our systematic Monte Carlo simulation study. We investigate the estimation of Spatial Lag Model with the Maximum Likelihood, which is the gold

standard in spatial econometric (Ord; 1975), the Two Stages Least Squares Kelejian and Prucha (1998), and Bayesian estimator. We test their performances for different sample sizes, different values of the parameters and different levels of sparsity of the weight matrices. We compare the methods referring to three performance indicators, namely: computing time, storage required and accuracy of the estimators. From our simulations exercise, we deduce that computational issues in estimation phase rise both from the sample size n both from high density of the weight matrix. We observe that computing time increases non-linearly with the sample size for all the procedures considered and it is also strongly related with the density of the weight matrix. Our conclusion is that the weight matrix density plays a key role in estimating the model both in terms of time required and accuracy.

As a second step, we consider the velocity at which data are acquired. The big data revolution is challenging the state-of-the-art statistical and econometric techniques in many different ways, not only in terms of volume, but also in terms of velocity. Handling a flow of observations in an efficient manner is a challenging thematic. From a spatial regression-type model perspective a focus in data streaming is lacking, nevertheless it is worth to explore this field since neglected spatial correlation could lead to inefficient estimators. Generally speaking, the presence of spatial correlation produces disturbance effects on statistical inference so that the classical estimation methods will lead to biased, inconsistent and inefficient parameter estimators (Cressie; 1993). In this case, the problem is more complex to solve than in the time series case since the spatial correlation between all observations must be taken into account. In Chapter 2 *Handling streaming of data in spatial econometric modelling* we propose an approximation to the available estimation method of the Spatial Error Model (Arbia; 2014) in order to make recursive writing of the estimation procedure possible. Since it would be time and space consuming to re-estimate the model all the time that a new piece of information becomes available, we try to find a recursive approach for the definition of the variables that combined with the Feasible Generalized Least Squares estimation method (Kelejian and Prucha; 1998) allows to reduce the number of algorithm steps and the storage required. In particular, we propose to approximate the spatial parameter ρ with an initial value and keep it fixed for the following steps of the streaming. Monte Carlo simulations under different values of the parameters allow us to conclude that, despite the important approximating assumption, the obtained results are comparable with those computed with the complete set of information and a recursive formulation of the model is essential to achieve a fast estimation procedure. Based on these results, we investigate the impact of the magnitude of the spatial correlation parameter ρ on the performance of the estimators of Spatial

Error Model. This investigation leads us to consider also the comparison with OLS, which can be considered a special case of Spatial Error Model with zero spatial correlation.

Chapter 3 *On the efficiency of estimators of the Spatial Error Model* contains our observations regarding the relative efficiency of the Feasible Generalized Least Squares and the OLS. As argued from Kramer and Donninger (1987), under some algebraic condition of the adjacency matrix W the Feasible Generalized Least Squares and the OLS have the same efficiency, in addition to being both consistent. In this framework, we further considerate algebraic structure of adjacency matrix and we investigate the effect on the size and the power of the models. We account for different structure of the W and values of the parameters involved.

Finally, we focus on one of the basic hypothesis of spatial models: the coefficients stationarity. In the specific case of large datasets, which are usually spread over large areas, this assumption is particularly unrealistic and unlikely. In most applications, the socio-economic variables considered follow specific patterns in the reference area, resulting in a non-homogeneous behaviour in the data. For this reason, in Chapter 4 *Modelling Non-Stationary Spatial Lag Model with Hidden Markov Random Fields* we propose a hierarchical model which, through the use of latent processes, is able to identify different groups within the observations Besag (1986). In this spirit, we introduce a latent process evolving according to a Potts model (Okabayashi et al.; 2011), which drives the coefficients values in a Spatial Lag Model. The Potts model structure allows us to consider similarities between neighbouring points, thus accounting for the intrinsic topology of the spatial model. We propose the estimation in a Bayesian framework with a Metropolis within Gibbs algorithm. Since this kind of models badly scales for large dataset, where needed, some parts of the code are programmed in C++ in order to obtain better computation performances. A simulation exercise was also performed to investigate the influence of individual model parameters on the accuracy of the results. Criteria to model comparison in order to establish the optimal number of clusters is also provided.

Bibliography

Arbia, G. (2014). *A primer for Spatial Econometrics*, Palgrave Macmillan.

Besag, J. (1986). On the statistical analysis of dirty pictures, *Journal of the Royal Statistical Society. Series B* **48**: 259–302.

Cressie, N. (1993). *Statistics for Spatial Data*, John Wiley & Sons, Inc.

Kelejian, H. and Prucha, I. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances, *The Journal of Real Estate Finance and Economics* **17**: 99–121.

Kramer, W. and Donninger, C. (1987). Spatial autocorrelation among errors and the relative efficiency of ols in the linear regression model, *Journal of the American Statistical Association* **82**: 577–579.

Okabayashi, S., Johnson, L. and Geyer, C. J. (2011). Extending pseudo-likelihood for potts models, *Statistica Sinica* **21**: 331–347.

Ord, K. (1975). Estimation methods for models of spatial interaction, *Journal of the American Statistical Association* **70**: 120–126.

Chapter 1

Estimation of spatial econometric linear models with large datasets: How big can spatial Big Data be?

1.1 Introduction

Spatial econometrics is currently experiencing the Big Data revolution both in terms of the volume of data and of the velocity with which they are accumulated. Regional data, employed traditionally in spatial econometric modeling, can be very large, the information being increasingly available at a very fine resolution at the level of census tracts, local markets, town blocks, regular grids or other small partitions of the territory. For example the US Census Bureau collects data at the level of about 75,000 census tracts. In the EU, Eurostat encourages the use of a regular grid to analyze and compare population distributions. The 1000-by-1000 square meters grid is the official one, thus involving some millions of observations, but many national institutes (like e.g. ISTAT in Italy) produce a regular grid population distribution at the level of 20-by-20 square meters involving many more data. When dealing with spatial micro econometric models referred to granular observations of the single economic agent, the number of observations can be a lot larger (Arbia, Espa and Giuliani; 2018). For example the US Census Bureau's Longitudinal Business Database (CBD) provides annually geo-coded observations for every private-sector firm, thus including some million establishments and employees. However, are the current spatial

econometric procedures able to handle such a massive quantity of data? Two are the elements that produce computational problems in estimation phase: the sample size (n) and the level of density of the weight matrix W (δ). In spatial studies a linear increase in the number of observations corresponds to a quadratic increase in the dimension of the weight matrix thus exacerbating the problems. Furthermore, since all the estimation procedures involve the inversion of some function of the weight matrix, its level of sparsity also plays a crucial role. Under this respect when using the Maximum Likelihood (ML) framework, the eigenvalues decomposition suggested by Ord (1975) has been, for many years, the gold standard in spatial econometrics to solve computational problems occurring with large data when 'large' meant only few hundreds. Unfortunately the computation of eigenvalues may become highly inaccurate already for matrices with a dimension of 400×400 (Kelejian and Prucha; 1998). The accuracy improves if the weight matrix is symmetric, which, however, is not the case when dealing with row-standardized weights. Many approximations proposed in the literature are usually sufficiently accurate if the weight matrix is sparse, but only if the sample is not too large (for example see *Applying the Generalized-Moments Estimation Approach to Spatial Problems Involving Microlevel Data* (2000), where the case study has a dimension of 2,000 observations). If the weight matrices are, conversely, dense all approximations become quickly rather inaccurate. The literature reports numerous attempts to reduce the computational burden by suggesting new approximations (Smirnov and Anselin; 2001; Pace and LeSage; 2004; Griffith; 2000, 2004), new less time-consuming estimators with respect to ML (the Two Stage Least Squares - 2SLS - suggested by Kelejian and Prucha (1998) and the asymptotically optimal estimator proposed by Lee (2003)), or by redefining the models so as to tackle the problem at its very root (LeSage and R.; 2007; Arbia et al.; 2013; Arbia; 2014a; Arbia, Bee, Espa and Santi; 2018; Santi et al.; 2017). A different solution for the computation of the determinant was proposed by (Pace and Barry; 1997a) that illustrate how it is possible to accelerate dramatically the calculation in spatial econometric model estimation by exploiting the properties of sparse matrices. Faster solution for estimating Conditionally Autoregressive Models were also proposed in the presence of sparse W matrices by (Pace and Barry; 1997b). It is well known that dense W matrices should be avoided in spatial models in order to reduce computational problems and also due to the bias which they introduce in the estimator of the spatial parameters (Smith; 2009). Nonetheless fully dense matrices can emerge in practical circumstances when individual spatial units are strongly connected like in social and neural network studies (Novikov et al.; 2015) or when W matrices are specified as an inverse function of pair-wise distances (Qu and fei Lee; 2015). Spatial gravity models provide a further example of

very dense connectivity matrices (LeSage and Pace; 2008). In these cases what are the practical limitations imposed by the level of connectedness? Furthermore, several commercial software and Geographical Information Systems (GIS) create by default W matrices based on the inverse-distance specification (thus fully dense) without making the user aware of the dangers connected with this choice. Indeed there are only a few attempts in the literature to answer to the question that we pose in the title of our Chapter. Only (Walde et al.; 2008) evaluate the computing time associated to various matrix decomposition techniques with the Generalized Spatial Two Stages Least Squares (GSTSLS) estimators, while Bivand et al. (2013) focus on the accuracy of some approximate approaches when dealing with sparse matrices (i.e. when the computation of the determinant is performed on W matrices of dimensions up to 1,000,000×1,000,000, characterized by a rook contiguity criterion), but neglecting the case of very large and dense weight matrices. This Chapter reports the results of a systematic simulation study that aims at evaluating the limits within which the current methodologies are still feasible and reliable when estimating spatial econometric models with very large data sets and different levels of density of the W matrices. To allow a comparability of the results between different estimation strategies in our study we limited ourselves to the Spatial Lag Model which is estimated with 2SLS, ML and a Bayesian procedure. We test the performances of the various estimators at different sample sizes and for different levels of density by referring to three performance indicators: computing time, storage required and accuracy of the estimators. The rest of the Chapter is organized as follows. In Section 1.2 we present the simulation plan. In Section 1.3 we report the simulation results, while Section 1.4 is devoted to a summary of the main findings and to some concluding remarks.

1.2 The simulation setup

In our simulations we focus on Gaussian spatial models limiting ourselves, in particular, to the Spatial Lag specification which is frequently the choice of selection in regional and urban studies. Formally, the model can be written as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \rho W\boldsymbol{Y} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathcal{MVN}(0, \sigma^2 I).$$

A key role in the model is played by the spatial autoregressive component $\rho W\boldsymbol{Y}$, where W is the row-standardized weight matrix that describes the spatial structure of connectedness. The definition of W is crucial in the application of spatial models to real cases because not only it

determines the parameters' estimations, but it also dramatically influences the computational features. For more details about the interpretation of SLM see e.g. Arbia (2014b). Although in the literature a lot of different approaches were proposed in order to solve the estimation problems, the most popular methods are Maximum Likelihood and 2SLS, whose procedures are implemented in most econometric software such as R, Matlab or Python. Bayesian routines are also available.

Our simulation exercise aims at studying the behavior of three performance indicators, namely (i) computation time, (ii) storage required and (iii) accuracy of the estimations, with respect to three model characteristics, namely (i) the sample size, (ii) the density of the W matrix and (iii) the model's parameters. The simulation experiment is structured as follows. First of all, we consider a random variable $\boldsymbol{Y} = [Y_1, ..., Y_n]$ laid on a regular square lattices. In order to explore the effects of the different sample sizes, we consider regular lattice grids of increasing dimensions ranging from 10×10 up to 264×264, corresponding to a sample size that ranges between 100 and 69,696 observations[1]. The limit of allocation for the data depends on the computer features, but also on the method used to store them. In our case (for details of computational resources see Section 1.3) it was not possible to allocate more than $n = 100,000$ observations.

We also want to investigate how different density levels of the weight matrix W influence the results, where the density of W is defined as the proportion of non-zero elements in the matrix (say δ). As it has been proved (Smith; 2009), weight matrix with a huge number of links leads to bias in the estimation of the spatial correlation parameter. However, many empirical applications (especially those where the spatial setting depends on some function of the distance between individual units) require a highly-connected structure of the neighborhood. For this reason it is necessary to consider the effects of dense W matrices in the estimation procedures. In our simulations δ ranges over the values [0.001, 0.07, 0.49, 0.9]. While a density of 0.9 is unusual, and in any case not recommended given the distortions described by Smith (2009), we consider useful to report the results also for this limiting case. In order to control for the density of W, we defined different levels of neighborhood according to the queen's case contiguity criterion. Increasing levels of density are obtained considering higher orders of neighborhood. We compute the weight matrix, say $W(n, \delta)$, for every combination on n and δ using the above procedure with one remarkable exception. Indeed, when $n = 100$ and $\delta = 0.001$ the only neighbors are the cells to the left and to the right on the lattice, because in this case it is otherwise not possible to obtain the target density. Once $W(n, \delta)$ is computed, in order to obtain samples drawn from

[1] The sample sizes considered are precisely of 100, 1024, 5041, 10000, 30276, 50176 and finally 69696.

a random field characterized by that specific spatial structure, we proceed as follows. A vector of independent disturbances $\boldsymbol{\epsilon}$ is generated from a normal distribution with zero mean and a variance σ^2, where σ^2 can assume three different values, namely 0.1, 1 and 10. Secondly, a vector of observations from a covariate $\boldsymbol{X}$ is generated from a normal distribution with a mean arbitrarily fixed to 10 and a variance equal to 20. Since $\boldsymbol{X}$ is considered deterministic, its value is fixed in all simulations. We also fixed the regression slope coefficient β equal to 1, and we computed $\tilde{\boldsymbol{X}} = W(n, \delta)\boldsymbol{X}$, thus introducing spatial correlation into the covariate. Furthermore, we consider three levels for the autoregressive parameter ρ equal to -0.7, 0 and 0.7. Once all parameters are specified, we simulated the sample observations from the SLM as follows:

$$\boldsymbol{Y}(n, \delta, \rho, \sigma^2, \beta) = (1 - \rho W(n, \delta))^{-1}\tilde{\boldsymbol{X}}\beta + (1 - \rho W(n, \delta))^{-1}\boldsymbol{\epsilon}.$$

The practical details of the construction of W and the generation of $\boldsymbol{Y}$ are provided in Appendix A.

1.3 Simulation results

In the current Section we present the main results of our simulation experiment. All the simulations were run using a cluster with 128 GB of RAM and a 3.50 GHz processor. The estimations are obtained with the `spdep` package, one of the most popular among researchers and practitioners. Bivand and Piras (2015) showed that the results of ML and 2SLS do not differ significantly when using different software routines like those contained in e.g. Matlab, Stata or R. Even if all our simulations were run using `spdep` , for comparison we have tested the procedures using also Matlab. The results are comparable to those reported here, although generally less time consuming in the case of Matlab. The package `spdep` contains user friendly routines for the ML procedure (the `lagsarlm` function), for the 2SLS procedure (the `stsls` function) and the corresponding Bayesian version (the `spBreg_lag` function). In the Bayesian framework, the prior and the likelihood are expressed as follows:

$$\begin{aligned}
&\boldsymbol{Y}|\boldsymbol{X}, \beta, \rho, \sigma^2 \sim \mathcal{N}((I - \rho W\boldsymbol{Y})^{-1}(\boldsymbol{X}\beta), \sigma^2(I - \rho W\boldsymbol{Y})^{-1}(I - \rho W\boldsymbol{Y})^{-T}) \\
&\beta \sim \mathcal{N}(0, 10^{12}) \\
&\sigma^2 \sim \Gamma(0.001, 0.001) \\
&\rho \sim Beta(1.01, 1.01)
\end{aligned}$$

In each simulation exercise we run the code for 15,000 iterations (1,500 of which were discarded as burnin) and with a thinning parameter of 10 we obtained a chain of length 5,000. It should be noticed that the execution time with the ML method is strongly influenced by the algorithm used for computing the determinant of the $(I - \rho W)$ matrix. While the differences between several methods proposed in the package (i.e. eigenvalues, LU decomposition, Cholesky decomposition and Monte Carlo approximation) are not significant with datasets of small dimensions, they become relevant when increasing the sample size. As suggested by Bivand et al. (2013), we choose the LU decomposition that has better performances with large datasets and can be applied to all kind of matrices.

Each simulation experiment was replicated 100 times. In our analysis we considered the average computing time, the average mean square error (MSE) of the parameters' estimation and the storage required.

1.3.1 Computing time considerations

In our simulation, first of all, we notice that the computing time is independent of the values of ρ and σ^2. For this reason Figure 1.1 reports only one specific combination of the parameters, namely $\rho = 0.7$ and $\sigma^2 = 1$. As it was expected, computing time essentially depends on the sample size. However this relationship is not linear: the time increases more sharply when n is greater than 5,000. Another remarkable feature is that computing time is strongly dependent on the density of the W matrix. We notice that timing problems appear also for relatively low values of δ, like 0.07. In this case a sample of 10,000 observations takes more the 30 minutes to be analyzed. The phenomenon is obviously exacerbated if we consider higher levels of density: the ML procedure takes 10 hours when dealing with $\delta = 0.9$ while with the Bayesian procedure the time goes up to 2 days. With our computing facilities, it was impossible to perform estimations with a sample size greater than 10,000 and with a high density of the W matrix due to the prohibitive amount of time required for the estimations. In the case of $\delta = 0.001$ it was possible to estimate models up to, approximately, $n = 70,000$ observations with the ML procedure and $n = 50,000$ with the Bayesian method. It is fair to notice that these problems do not affect the 2SLS method, which requires only few second in any simulation scenario. Figure 1.2 clearly displays the computing time required by the three procedures as a function of the sample size.

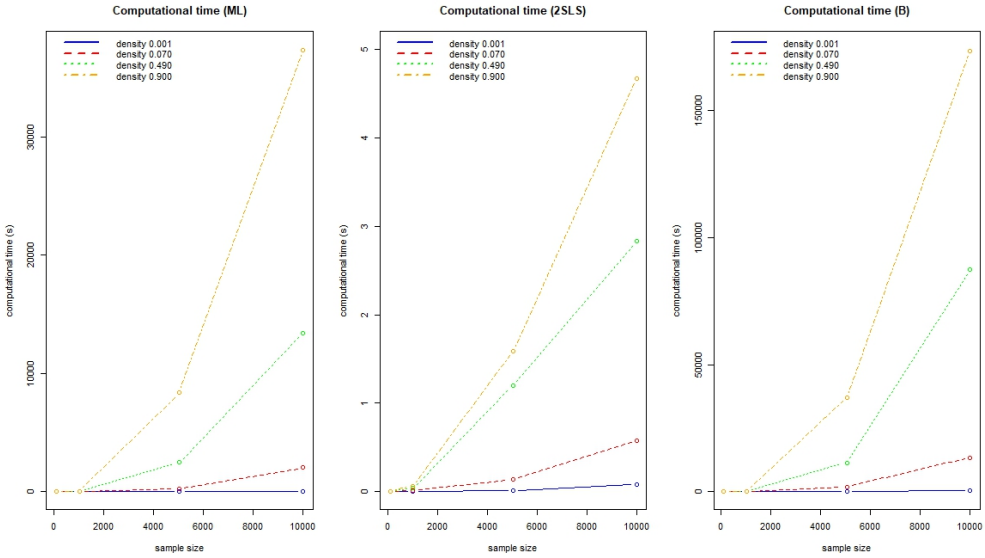

Figure 1.1: Computational time related to the estimation of the SLM model at different sample size and different density of W, where the three estimation methods are: ML= Maximum Likelihood, 2SLS= Two Stages Least Squares, B=Bayesian ($\rho = 0.7$ and $\sigma^2 = 1$).

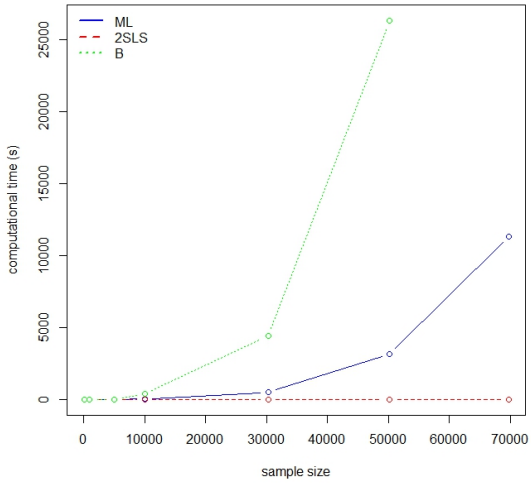

Figure 1.2: Computational time related to the estimation of the SLM at different sample sizes, where the three estimation methods are: ML= Maximum Likelihood, 2SLS= Two Stages Least Squares, B=Bayesian ($\rho = 0.7$, $\sigma^2 = 1$ and $\delta = 0.001$).

In our evaluation of computing time there are two important points to remark. First of all, the estimation procedures included in the R routines return different outputs. In the case of the ML procedure, statistics for significance and information about likelihood and optimization process are also provided, while they are not calculated in the 2SLS case. This results in an increase in computing time. Secondly, consistent results can be obtained with other software like Matlab, which however generally performs better.

1.3.2 Accuracy evaluation

We then proceed to evaluate the accuracy of the estimators using the three estimation procedures through the analysis of the mean square errors. For the sake of brevity, we only report here some examples in order to support our conclusions. Figure 1.3 and 1.4 report the error respectively for β and ρ. In all graphs $n = 1024$ while we consider two levels of variance: $\sigma^2 = 1$ and $\sigma^2 = 10$. More analytical results are reported in Appendix B. In general, all procedures are accurate with small σ^2, while, as expected, the error increases with the residuals error variance. This holds, in particular, for the 2SLS procedure where with "high densities" the error increases dramatically. Our simulation shows that, although with different patterns, accuracy depends crucially on the density of W for all estimation methods: estimators become less accurate for denser W matrices.

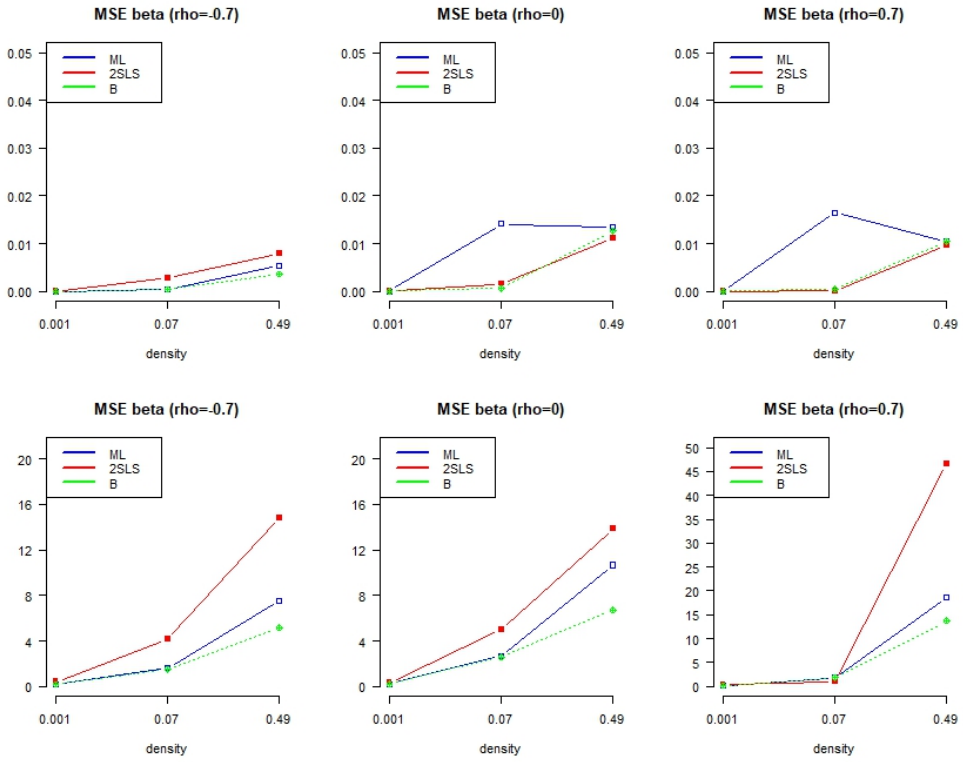

Figure 1.3: Comparison of the MSE of the estimator of the regression parameter β with $n = 1024$ and $\sigma^2 = 1$ (upper graphs), and $\sigma^2 = 10$ (lower graphs) as a function of the density of the W matrix. Vertical scales can be different in the various graphs.

Generally speaking the estimations of the spatial correlation parameter ρ are more accurate than those of the regression parameter β. Furthermore another important evidence from the experiment is that the Bayesian estimators are less affected by the values of the various parameters and, even in the worst case, they are less sensitive to the density of the W matrix.

1.3.3 Storage issues

Finally, let us consider the storage required by the various procedures in the different simulation scenarios. The major results are summarized in Table 1.1. The storage required appears to be unaffected by the density of W and by the parameters' values and it seems to only depend on the sample size with the remarkable exception of the Bayesian estimator. Generally speaking ML is more space-consuming than the other methods, this behavior depending mostly on the

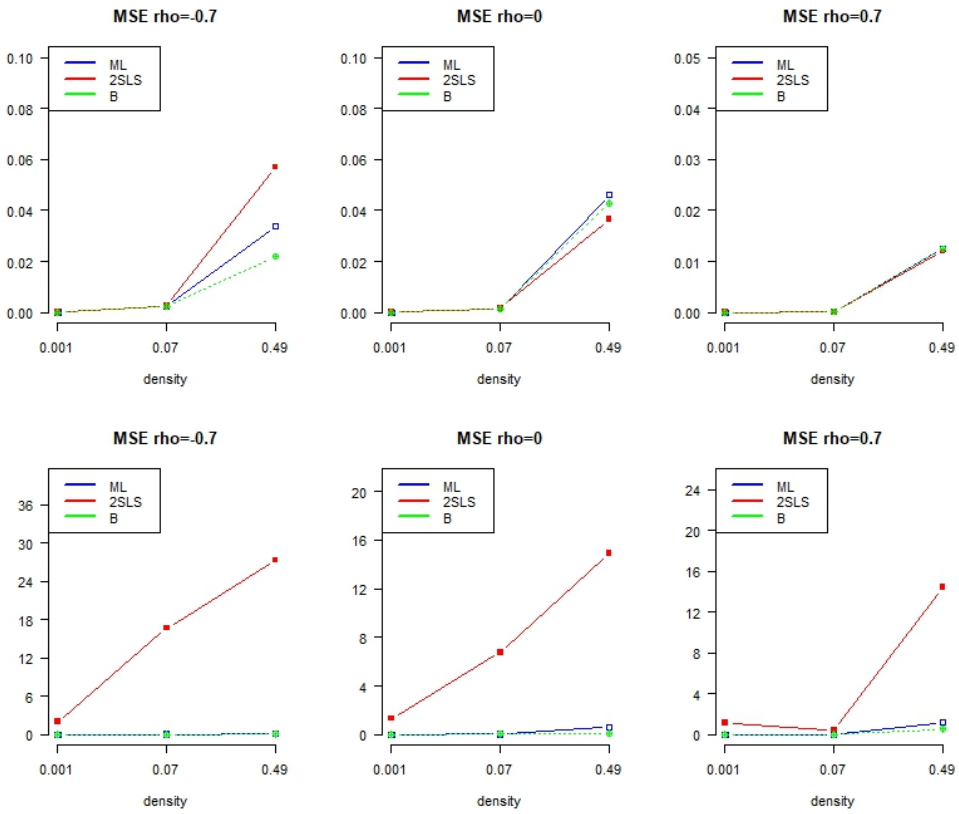

Figure 1.4: Comparison of the MSE of the estimator of the spatial correlation parameter ρ with $n = 1024$ and $\sigma^2 = 1$ (upper graphs), and $\sigma^2 = 10$ (lower graphs) as a function of the density of W matrix. Vertical scales can be different in the various graphs.

structure of the R function `lagsarlm` which saves in the output also many summary statistics and data matrices.

1.4 Summary and Conclusions

In this Chapter, we analyzed, through an intensive simulation study, what are the limits within which spatial econometric models can be accurately estimated when dealing with very large datasets. We observed that computing time increases non-linearly with the sample size for all the procedures considered and it is also related with the density of the weight matrix. As it is well known 2SLS largely outperforms ML and Bayesian procedure in terms of computing time. For the last two methods it was not possible to estimate models with more than $10,000$

	$n = 100$	$n = 1,024$	$n = 5,041$	$n = 10,000$
ML	0.053	0.423	2,209	4,013
2SLS	0.003	0.11	0.43	0.83
B	0.860	0.860	0.860	0.860

Table 1.1: Storage required by the different estimation procedures expressed in MegaBytes.

units when matrices are characterized by a high density, due to the prohibitive amount of time required, while with 2SLS a sample of dimension 70,000 could be easily accommodated. The density of the W matrix plays a key role in the estimation process since it negatively affects both the computing time and the accuracy of the results. For this reason in practical applications, it is suggested to avoid matrices with a very dense structure, as already remarked in papers like (Pace and Barry; 1997a,b) and Smith (2009). In cases where a dense structure is required by the problem, it is advisable to introduce a cut-off in order to reduce the computational problems. In terms of the accuracy of the estimators, by comparing the simulated MSE, we found that ML and Bayesian methods outperform 2SLS in the estimations of all parameters with all sample sizes and all densities of the weight matrices in most experimental situations. This results holds, in particular, when the variability of the disturbances is very high, in which case the MSE obtained with 2SLS increases dramatically. Furthermore, in most simulation circumstances, the estimation seems to be affected by the increase of the density of the W matrix. This is particularly true for the 2SLS procedure, while the Bayesian procedure provides more stable results. Finally, we observed that the computer storage increases linearly with the sample size for both ML and 2SLS estimators, while it is approximately constant in the case of the Bayesian estimator. ML is more space-consuming than 2SLS with an average storage of about 4 MB for a sample size of 10,000 units, against only 0,8 MB required by the 2SLS procedure. Since the increase is linear our results can help the users to identify what are the limits of the sample size that can be managed by the computational resources that they have access to. Our results can also help the researchers to choose between the various estimators in empirical applications. If time and computer storage is not a constraint it is always better to exploit an ML strategy. 2SLS is particularly not recommended with very large sample sizes when the weight matrix is very dense, as it happens in many social interaction applications. The Bayesian estimator seems to be a reasonable alternative especially when the W matrix is very dense.

Given our results, with the growing diffusion of very large databases including often millions (or

more) spatial interacting individuals (e.g. in social network applications) and with the state-of-the-art spatial econometric methodologies, it is easy to forecast that in the near future we will find ourselves in the practical impossibility to analyze data in a timely and accurate way. Hence the studies related to time-efficient and statistically-accurate procedures represent an area that requires urgent solutions and a discontinuity step with respect to the recent evolution of the methodology.

Appendix A

Below it is briefly described the algorithm to build a W matrix referred to the spatial structure of a grid of dimension $n \times m$ and k levels of neighborhood and the sampler scheme.

Algorithm 1 Data simulation and model estimation

Construction of W

1: fix n and δ

2: compute the level of neighborhood k

3: generate W following Algorithm 1

Generation of the dataset

4: fix the values of the parameters ρ, σ and β

5: sample $\epsilon \sim \mathcal{N}(0, \sigma^2)$

6: sample $X \sim \mathcal{N}(10, 20)$

7: compute $Y \leftarrow (I - \rho W)^{-1}(X\beta + \epsilon)$

Estimation of the model

8: $mod_{ML} \leftarrow lagsarlm(Y \sim X, listw = mat2listw(W, style =' W'), method =' LU')$

9: $mod_{2SLS} \leftarrow stsls(Y \sim X, listw = mat2listw(W, style =' W'))$

10: $mod_{B} \leftarrow spBreg\_lag(Y \sim X, listw = mat2listw(W, style =' W'), type =' lag')$

11: Extract the results and compute MSE, storage and computational time

Algorithm 2 Generation of W matrices

input: n, m, k

$d \leftarrow nm$

$W \leftarrow matrix(rep(0, d^2), nrow = d, ncol = d)$

for all in 1:d **do**

  $i \leftarrow t\%\%n$

  if i==0 **then** $i \leftarrow n$

  $j \leftarrow ceiling(t/n)$

  $k_1 \leftarrow min(i-1,k), k_2 \leftarrow min(j-1,k), k_3 \leftarrow min(m-j,k)$ and $k_4 \leftarrow min(n-i,k)$

neighbours on the right

  if j==m **then** {}

  else $W[t, seq(t+n, t+k_3 n, n)] \leftarrow 1$

neighbours on the left

  if j==1 **then** {}

  else $W[t, seq(t-k_2 n, t-n, n)] \leftarrow 1$

neighbours above

  if i==1 **then** {}

  else

    for l in $1 : k_1$ **do**

      $W[t, seq(t-l-k_2 n, t-l+k_3 n, n)] < -1$

neighbours below

  if i==n **then** {}

  else

    for l in $1 : k_4$ **do**

      $W[t, seq(t+l-k_2 n, t+l+k_3 n, n)] < -1$

$W \leftarrow W/rowSums(W)$

return : W

Appendix B

		$\boldsymbol{\rho}$			$\boldsymbol{\beta}$		
		$\rho = -0.7$	$\rho = 0$	$\rho = 0.7$	$\rho = -0.7$	$\rho = 0$	$\rho = 0.7$
$\delta = 0.001$	$\sigma^2 = 1$	0.00007	0.00006	0.00001	0.00002	0.00003	0.00003
	$\sigma^2 = 10$	0.00397	0.00369	0.00111	0.18289	0.19684	0.17015
$\delta = 0.07$	$\sigma^2 = 1$	0.00245	0.00167	0.00016	0.00040	0.00065	0.00048
	$\sigma^2 = 10$	0.03051	0.03456	0.00627	1.63291	2.68478	1.79517
$\delta = 0.49$	$\sigma^2 = 1$	0.03393	0.04605	0.01253	0.00531	0.01345	0.01041
	$\sigma^2 = 10$	0.10895	0.63154	1.20799	7.50964	10.6728	18.6957
$\delta = 0.90$	$\sigma^2 = 1$	0.09770	0.52813	1.58917	0.01805	0.01672	0.02190
	$\sigma^2 = 10$	0.09514	0.91970	2.51405	281.798	231.448	157.059

Table 1.2: MSE of the estimator of the spatial correlation parameter ρ and the regression parameter β with $n = 1024$, obtained with the Maximum Likelihood estimation method.

		$\boldsymbol{\rho}$			$\boldsymbol{\beta}$		
		$\rho = -0.7$	$\rho = 0$	$\rho = 0.7$	$\rho = -0.7$	$\rho = 0$	$\rho = 0.7$
$\delta = 0.001$	$\sigma^2 = 1$	0.00007	0.00006	0.00001	0.00002	0.00003	0.00004
	$\sigma^2 = 10$	0.00392	0.00359	0.00115	0.18321	0.19421	0.17100
$\delta = 0.07$	$\sigma^2 = 1$	0.00247	0.00153	0.00017	0.00040	0.00061	0.00052
	$\sigma^2 = 10$	0.01793	0.02934	0.00689	1.49957	2.60625	1.84144
$\delta = 0.49$	$\sigma^2 = 1$	0.02207	0.04290	0.01249	0.00360	0.01267	0.01041
	$\sigma^2 = 10$	0.11244	0.12268	0.53061	5.15733	6.73867	13.7950
$\delta = 0.90$	$\sigma^2 = 1$	0.24355	0.03645	0.53397	0.01596	0.01721	0.02095
	$\sigma^2 = 10$	0.35888	0.01892	0.64652	284.881	240.097	142.201

Table 1.3: MSE of the estimator of the spatial correlation parameter ρ and the regression parameter β with $n = 1024$, obtained with the Two Stages Least Squares estimation method.

		$\boldsymbol{\rho}$			$\boldsymbol{\beta}$		
		$\rho = -0.7$	$\rho = 0$	$\rho = 0.7$	$\rho = -0.7$	$\rho = 0$	$\rho = 0.7$
$\delta = 0.001$	$\sigma^2 = 1$	0.00007	0.00006	0.00001	0.00002	0.00003	0.00003
	$\sigma^2 = 10$	1.95619	1.30728	1.13793	0.40572	0.30270	0.41406
$\delta = 0.07$	$\sigma^2 = 1$	0.00271	0.00164	0.00015	0.00040	0.00065	0.00046
	$\sigma^2 = 10$	16.6949	6.78245	0.45504	4.18173	4.98586	0.90919
$\delta = 0.49$	$\sigma^2 = 1$	0.05699	0.03654	0.01199	0.00801	0.01113	0.00971
	$\sigma^2 = 10$	27.3203	14.8862	14.3977	14.7748	13.9130	46.4945
$\delta = 0.90$	$\sigma^2 = 1$	2.21343	1.08794	0.77866	0.02536	0.01768	0.02202
	$\sigma^2 = 10$	114.767	268.4957	3686.95	1168.96	1033.68	2015.23

Table 1.4: MSE of the estimator of the spatial correlation parameter ρ and the regression parameter β with $n = 1024$, obtained with Bayesian estimation method.

Bibliography

Applying the Generalized-Moments Estimation Approach to Spatial Problems Involving Microlevel Data (2000). *The Review of Economics and Statistics* **82**: 72–82.

Arbia, G. (2014a). Pairwise likelihood inference for spatial regressions estimated on very large datasets, *Spatial Statistics* **7**: 21 – 39.

Arbia, G. (2014b). *A primer for Spatial Econometrics*, Palgrave Macmillan.

Arbia, G., Bee, M. and Espa, G. (2013). Testing isotropy in spatial econometric models, *Spatial Economic Analysis* **8**: 228–240.

Arbia, G., Bee, M., Espa, G. and Santi, F. (2018). Fitting spatial regressions to large datasets using unilateral approximations, *Communications in Statistics - Theory and Methods* **47**: 222–238.

Arbia, G., Espa, G. and Giuliani, D. (2018). *Spatial microeconometrics*, Routledge.

Bivand, R., Hauke, J. and Kossowski, T. (2013). Computing the jacobian in gaussian spatial autoregressive models: An illustrated comparison of available methods, *Geographical Analysis* **45**: 150–179.

Bivand, R. and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics, *Journal of Statistical Software* **63**: 1 – 36.

Griffith, D. A. (2000). Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses, *Linear Algebra and its Applications* **321**: 95 – 112.

Griffith, D. A. (2004). Extreme eigenfunctions of adjacency matrices for planar graphs employed in spatial analyses, *Linear Algebra and its Applications* **388**: 201 – 219.

Kelejian, H. and Prucha, I. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances, *The Journal of Real Estate Finance and Economics* **17**: 99–121.

Lee, L. (2003). Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances, *Econometric Reviews* **22**: 307–335.

LeSage, J. P. and Pace, R. K. (2008). Spatial econometric modeling of origin-destination flows, *Journal of Regional Science* **48**: 941–967.

LeSage, J. P. and R. (2007). A matrix exponential spatial specification, *Journal of Econometrics* **140**: 190 – 214. Analysis of spatially dependent data.

Novikov, A., Podoprikhin, A. and Vetrov, D. (2015). Fast cars, *Tensorizing Neural Networks* pp. 442 – 450.

Ord, K. (1975). Estimation methods for models of spatial interaction, *Journal of the American Statistical Association* **70**: 120–126.

Pace, K. and Barry, R. (1997a). Sparse spatial autoregressions, *Statistics & Probability Letters* **33**: 291 – 297.

Pace, R. and Barry, R. (1997b). Fast cars, *Journal of Statistical Computation and Simulation* **59**: 123 – 147.

Pace, R. and LeSage, J. P. (2004). Chebyshev approximation of log-determinants of spatial weight matrices, *Computational Statistics & Data Analysis* **45**: 179 – 196.

Qu, X. and fei Lee, L. (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix, *Journal of Econometrics* **184**: 209 – 232.

Santi, F., Arbia, G., Bee, M. and Espa, G. (2017). A frequency domain test for isotropy in spatial data models, *Spatial Statistics* **21**: 262 – 278.

Smirnov, O. and Anselin, L. (2001). Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach, *Computational Statistics & Data Analysis* **35**: 301 – 319.

Smith, T. E. (2009). Estimation bias in spatial models with strongly connected weight matrices, *Geographical Analysis* **41**: 307–332.

Walde, J., Larch, M. and Tappeiner, G. (2008). Performance contest between mle and gmm for huge spatial autoregressive models, *Journal of Statistical Computation and Simulation* **78**: 151–166.

Chapter 2

A recursive spatial regression model to handle streaming of geo-located data

2.1 Introduction

Starting a few years ago, the big data revolution has been challenging the state-of-the-art statistical modeling techniques in many different fields of application ranging from computer science to medicine, to the social sciences to name only a few. As discussed in Chapter 1, these challenges are produced not only by the huge volume of data which are generated and the consequent obvious increased computational effort, but also by many other factors that include, for example, the speed with which data become available.

Of particular interest is the case in which data are distributed over space. With the support of the new information technologies, a wide variety of geo-located data is becoming increasingly available to researchers. Some common (but certainly not exhaustive!) examples include data generated through sensors, meters or telecommunication networks (like those related to cell phone conversations or industrial electricity consumption); data generated by credit/debit cards or fidelity cards transactions; data automatically scraped from the web; and, last but not least, crowdsourced data voluntarily transmitted regularly through smart phones, in order to measure phenomena that are otherwise difficult (or unreliable) to quantify with traditional sur-

veys. Crowdsourced collection of food prices in developing countries (Arbia et al.; 2018) and epidemiological data (Créquit et al.; 2018) are among the most common example of this last category.

In this chapter we focus on the particular case related to the streaming of data, i.e. a flow of data that are continuously received by some automatic process (Bar-Yossef et al.; 2004). For their very nature, data streams are often too large and the increase in volume is too rapid to be properly dealt with using traditional tools. Simple statistics (such as variance or correlation) can be easily calculated in a recursive way (Felber; 2015), but more complex models are far more difficult to analyze. In a nutshell, the general idea of a recursive approach is to summarize part of the information and use it when new pieces of information become available. The state-space models (Ord; 1975), the Kalman-Filter (Welch and Bishop; 2006) and the recurrent neural network (Hochreiter and Schmidhuber; 1997) are all examples of this approach. Recursive forms equations were derived in a time series context (Hiriotappa et al.; 2017; Escobar and Moser; 1993) as well as in a linear regression framework (Wermuth; 1992; Seidou and Ouarda; 2007). Most of the research in spatial field is based on the trajectories study (Potamias et al. (2006); Tang et al. (2014) and Muckell et al. (2010) just to mention a few). An increasing focus can be notice on spatio-temporal applications (Oud et al.; 2012; Morales and Klette; 2013). Most of the research in the spatial field is based on the trajectories study (Potamias et al. (2006); Tang et al. (2014) and Muckell et al. (2010) just to mention a few),even if an increasing interest can be notice on spatio-temporal applications (Oud et al.; 2012; Morales and Klette; 2013). Among the various fields of application of the use of spatial data streaming, in addition to those already mentioned, we can remember the transportation industry (Putatunda; 2017; Patroumpas and Sellis; 2012), raster image data continuously streaming down to Earth from satellites (Gertz et al.; 2006). Speaking of a spatial data stream, the major focus is on the management of a huge flow of geospatial information from an informatic point of view (Fox et al.; 2006; Galić et al.; 2014) in order to improve the performance in storing and querying spatial database (Thakur et al.; 2015; Zhang; 2006). It is commonly recognized the importance of developing new tools for the management and analysis of spatial data, especially in view of the increasing speed at which data is collected (Armstrong et al.; 2019; Raman and Ali; 2010).

From a spatial regression model perspective, techniques to handle data streaming have not yet been developed. Indeed, the spatial case presents some peculiar difficulties which are related to the fact that, when a new piece of information becomes available, one needs to update not

only the dataset, but also the entire topology of the system. In this chapter we consider the problem of estimating a linear spatial regression model using continuously generated streams of spatial data. In particular, we restrict our attention to the popular Spatial Error Model (henceforth SEM - Anselin (1988)). We propose a recursive approach which is based on two basic simplifying assumptions: First of all, we assume that the spatial correlation parameter does not change significantly during the process of data acquisition and hence we set it to a value which is based on the initial regression estimated when the dataset is first made available. Secondly, we assume a particular specification for the spatial weighting matrix which simplifies substantially the upgrading process and we derive recursive formulas to estimate the parameters in the model. Our Monte Carlo results show that, despite the limitation imposed by our two basic assumptions, both the estimation results and the statistical inference on the model are comparable with those obtained using the complete set of information in each moment of the streaming.

The rest of the chapter is organized as follows: Section 2.2 explains in details the procedure adopted. In Section 2.3 we present the result of the Monte Carlo simulation. Concluding remarks follow in Section 2.4.

2.2 Modeling approach

In the field of geo-located streaming, we can find several typologies of information. The first possible situation is the case where data are available only for a range of time. Examples can be data collected from the smartphone of the peoples location, the position of the available shared vehicles (cars, bikes or scooters). As expected, the positions of data change quite quickly and each time step points appear an disappear on the map. In any case, the number of available information hardly drastically increases. For this reason, a huge quantity of data is not reached. A different problem is an information flow where the number of observations is continuously increasing. Real cases are for example the posts on social networks with a spatial tag or the positions of ill people during and emergency among the others. Especially if the stream of information is fast, the dimension of the dataset becomes significant. Therefore, it is this second case that we consider in this paper. In addition, we do not impose restrictions of any kind on the possible localizations of information, without reducing us to the case of spatial panels where we have the same statistical units for each time step. We now briefly review the SEM specification and discuss alternative estimation procedures in order to propose a recursive formulation.

The SEM specification and estimation

The SEM model is defined as

$$\begin{aligned} y &= X\beta + u \\ u &= \rho W u + \varepsilon \\ \varepsilon &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \end{aligned} \tag{2.1}$$

where X is an $n \times k$ matrix of regressors, β is a $k \times 1$ vector of parameters, W is an $n \times n$ exogenous spatial weighting matrix that describes the topology of the space, and ρ is the corresponding spatial parameter. As it is widely known, the SEM cannot be efficiently estimated by OLS and, therefore, a GLS approach has to be employed. Generally, either Maximum Likelihood (ML) or generalized method of moments (GMM) are implemented in order to make the GLS feasible (FGLS). One of the main advantages of the GMM with respect to ML has to do with computational time. As the sample size increases the computation of the so-called Jacobian term becomes an obstacle for ML. In truth, it should be noticed that many exact as well as approximated solution have been developed for the computation of the Jacobian term (Ord; 1975; Bivand et al.; 2013; LeSage and Pace; 2009; Barry and Pace; 1999; Pace and LeSage; 2004; Griffith; 2004; Smirnov and Anselin; 2009, 2001; Pace and Barry; 1997). For this reason, GMM is a natural candidate when the number of spatial units is particularly large, as it is assumed in this chapter. This is why we briefly review the estimation procedure described in Kelejian and Prucha (1998, 2010) and Drukker et al. (2013). The procedure develops through the following four different computational steps:

1. Estimate the first equation in (2.1) by OLS to get an initial estimate of β (say $\tilde{\beta}$)
2. Calculate the estimated residuals as $\tilde{u} = y - X\tilde{\beta}$
3. Use the estimated residuals to solve the moment conditions (Drukker et al.; 2013) in order to obtain an estimate of ρ, (say $\hat{\rho}$)
4. Estimate $\hat{\beta}$ and $\hat{\sigma}^2$ from a linear regression on the spatially filtered variables

$$\tilde{y} = \tilde{X}\beta + \varepsilon \quad \varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

where $\tilde{y} = (I - \hat{\rho}W)y$ and $\tilde{X} = (I - \hat{\rho}W)X$.

The recursive estimation of the SEM

In data streaming, the number of observations can increase quite rapidly. Assuming that the new information is geo-located, the spatial weighting matrix should be re-defined and the model re-estimated whenever a new piece of information becomes available. If the amount of incoming data is substantially large (or if data arrive at a particularly fast rate), it might be wise to introduce some assumptions to speed up and simplify the estimation procedure.
Our suggestion is twofold: on the one hand, under the assumption that spatial correlation does not change dramatically when new data are available, we propose to estimate the spatial parameter on the initial set of data and keep it fixed throughout the estimation process. Of course, since the real value of ρ may vary, at some point the initial estimate should be re-calibrated when changes become significant. There could be several ways to assess if, and when, the value of ρ should be re-estimated. One way could be monitoring the normality of the residuals (Sargan and Bhargava; 1983; Shapiro and Wilk; 1965). Another way could be defining a-priori a maximum number of new observations after which the entire model should be used to re-estimate the spatial parameter. Although interesting, we leave this point for future researches. On the other hand, we suggest a recursive procedure to update the spatial weighting matrix when a set of new geo-located data become available. In order to implement such a recursive procedure, W should be a row-standardized binary matrix defined in terms of some cut-off distance (Arbia; 2014). It is intuitive that the distance between units remain unchanged when new observations arrive. Hence, each set of neighbors can only increase (or, at most, remain unchanged), when new observations are available. As we will discuss in the Monte Carlo section, this iterative procedure presents some relevant advantages in terms of computational speed.

Let $s = 1, ..., S$ be the index denoting the availability of new observations at subsequent streaming steps. Similarly, let $n = n_1, ..., n_S$ represent the dimension of the sample at the steaming step s. In this context, $y^{(1)}, X^{(1)}$ and $W^{(1)}$ relate to the initial sample size n_1; while $y^{(s)}, X^{(s)}$ and $W^{(s)}$ are the elements of the new information available at the s-th step of the steaming. Clearly, X and y can be updated easily, but the recalculation of the weighting matrix can be challenging. For this reason, a definition of W which allows a recursive formulation would improve enormously the computational performance. Once the cut-off distance has been fixed at a certain value (say d^*), the elements of $W^{(1)}$ can be defined as

$$W_{i,j}^{(1)} = \begin{cases} 1 & \text{if } d_{i,j} < d^* \\ 0 & \text{otherwise} \end{cases}$$

where $d_{i,j}$ is the distance between units i and j, with $i = 1, ..., n_1$ and $j = 1, ..., n_1$. Let $nn_i^{(1)}$ be the $i - th$ element of the $n_1 \times 1$ vector defined as

$$nn_i^{(1)} = \sum_{j=1}^{n_1} W_{i,j}^{(1)}.$$

A row-standardized version of $W^{(1)}$ can be defined by dividing each of its rows by the corresponding element of $nn^{(1)}$. For the sake of simplicity, suppose that when $s = 2$, only one observation is added (say o_2) and the sample size increases from n_1 to $n_2 = n_1 + 1$. Then, $W^{(2)}$ is obtained from $W^{(1)}$ as

$$W^{(2)} = \begin{bmatrix} W^{(1)} & w \\ w^T & 0 \end{bmatrix}$$

where w is a $n_1 \times 1$ binary vector containing the neighbors of the new observation. At this point, the vector $nn^{(2)}$ can be easily obtained by updating $nn^{(1)}$ in the following way:

$$nn_i^{(2)} = \begin{cases} nn_i^{(1)} & i = 1, ..., n_1 \quad \text{if } i \text{ and } o_2 \text{ are not neighbors} \\ nn_i^{(1)} + 1 & i = 1, ..., n_1 \quad \text{if } i \text{ and } o_2 \text{ are neighbors} \\ \sum_{k=1}^{n_1} w_k & i = n_1 + 1. \end{cases}$$

Once $W^{(2)}$ and $nn^{(2)}$ are obtained, and with $\hat{\rho}$ based on the initial sample of size n_1, the filtered dependent variable at $s = 2$ can be calculated as

$$\begin{aligned} \tilde{y}_i^{(2)} &= y_i^{(2)} - \hat{\rho}^{(1)} \frac{1}{nn_i^{(2)}} \sum_{j=1}^{n_1+1} W_{i,j}^{(2)} y_j^{(2)} \qquad i = 1, ..., n_1 + 1 \qquad (2.2) \\ &= y_i^{(2)} - \hat{\rho}^{(1)} \frac{1}{nn_i^{(2)}} \left(\sum_{j=i}^{n_1} W_{i,j}^{(1)} y_j^{(1)} + w_i y_{n_1+1}^{(2)} \right) \\ &= y_i^{(2)} - M_i^{(2)}. \end{aligned}$$

Since all the operations are linear, the last line of (2) can be written in a recursive formulation:

$$M_i^{(2)} = \begin{cases} M_i^{(1)} & i = 1, ..., n_1 \quad \text{if } i \text{ and } o_2 \text{ are not neighbors} \\ \frac{nn_i^{(1)}}{nn_i^{(2)}} M_i^{(1)} + \hat{\rho}^{(1)} \frac{Y_{n_1+1}^{(2)}}{nn_i^{(2)}} & i = 1, ..., n_1 \quad \text{if } i \text{ and } o_2 \text{ are neighbors} \\ \hat{\rho}^{(1)} \frac{w_i y^{(1)}}{nn_{n_1+1}^{(2)}} & i = n_1 + 1 \end{cases}$$

with $M_i^{(1)} = \hat{\rho}^{(1)} \sum_{j=1}^{n_1} W_{i,j}^{(1)} y_j^{(1)}$.

A similar procedure is needed to update $\tilde{X}^{(2)}$. In the next section we will test through Monte Carlo experiment the reliability of the proposed methodology when we want to approximate the analysis of the whole dataset saving computing time and memory.

2.3 Simulation setting and results

In this section we report the results of a series of Monte Carlo experiments designed to evaluate the effect of the approximation on ρ and to investigate the time gain achieved by employing our recursive approach.

Monte Carlo design

Using again a spatial error framework, our data generating process is the following:

$$y = \alpha + \beta x + (I - \rho W^{(S)})^{-1}\varepsilon. \tag{2.3}$$

where the values of x are generated from a Uniform distribution over the interval $(0, 10)$ and they are kept fixed over the different Monte Carlo samples. The value of the intercept α is one, the value of the slope parameter β is two; and ρ increases from -0.8 to 0.8 by 0.2. We consider two different scenarios for the error term. In the first scenario the errors are homoskedastic $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ while the second scenario allows for heteroskedasticity $\varepsilon_i \sim \mathcal{N}(0, (\sigma + 0.6x_i)^2)$. In both cases the value of σ is set to 7. The total number of spatial units $n_S = 7,000$ are generated from a set of coordinates that are independent draws from a Uniform distribution over the interval $(0, 10)$. In order to mimic the streaming, we recreate a flow of data in six intervals $s = 1, ..., 7$, with an initial sample size of $n_1 = 1,000$. Observations are added in groups of 1,000 until we reach a total number $n_7 = 7,000$. The observations for the initial sample size as well as those that are added at each interval are chosen randomly.

In the following, we compare the results obtained with the recursive approach described in the previous section with those obtained by using, at each step of the streaming, the FGLS on the entire dataset available at the considered step. Figures 2.1-2.4 report the results related to 1,000 Monte Carlo samples for the homoskedastic model. Since they lead to similar conclusions, the results when the error is heteroskedastic can be found in appendix A. The first two figures illustrate the bias of the slope parameter β (Figure 2.1) and the spatial parameter ρ (Figure 2.2) over the streaming. In both Figure 2.1) and 2.2), the blue line corresponds to the recursive method, the red line to the FGLS and the orange dashed line is the true value of the parameters.

Consistently with the theory, Figure 2.1 shows that there is no bias in estimating β, both using the FGLS and the recursive approach. It is important to note that the averages of the estimated values of β are the same only for the initial sample $n_1 = 1000$ where the filtered variables are based on the same estimated value of ρ. For the other intervals, the parameter ρ is

estimated in every step and then used in the FGLS. In contrast, the recursive approach calculates the filtered variables by keeping the spatial parameter constant at the initially estimated value.

The pattern described by Figure 2.2 is also somewhat predictable. The blue line, indicating the recursive approach, remains constant at the value of ρ which was estimated on the initial set of data (i.e., $\hat{\rho}^{(1)}$). It is worth noticing that given our setup, we do not know a priori what is the true parameter value and not even if it will be close or far away from the "final" value estimated on the entire set of data (i.e., $\hat{\rho}^{(7)}$). Figure 2.2 also shows that the red line starts from the same value as the blue one but consistently tends to move towards the true value of the spatial parameter.

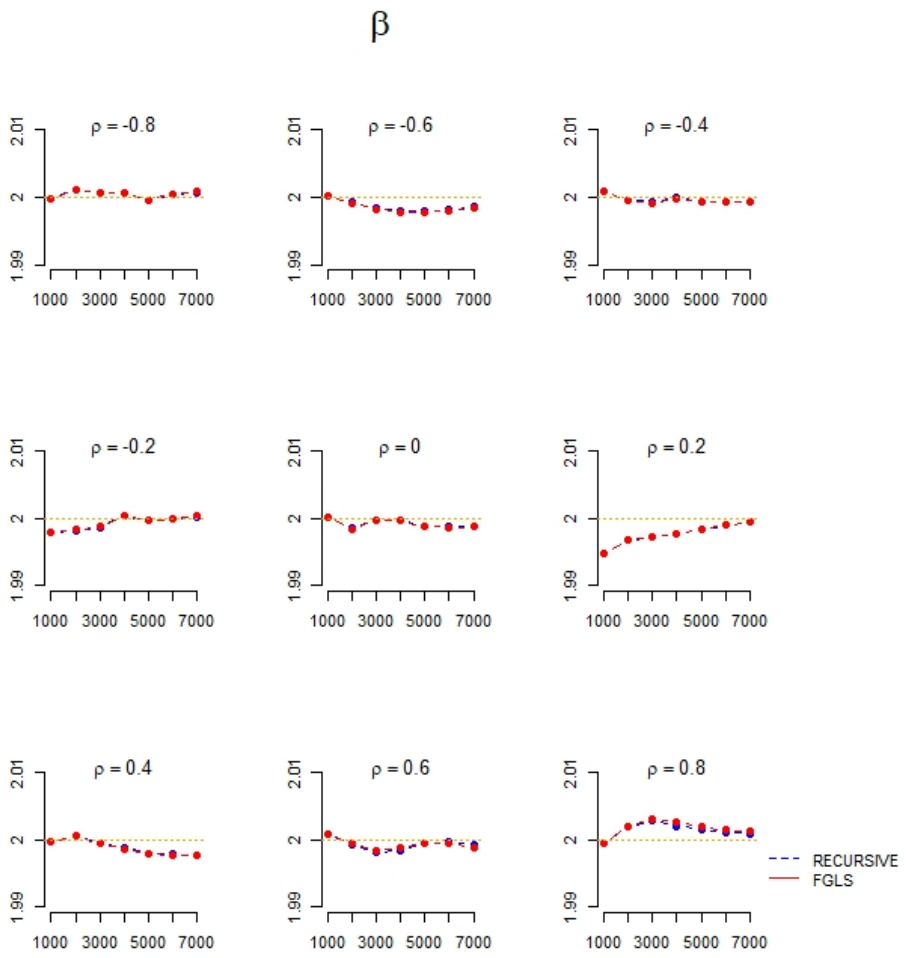

Figure 2.1: Bias of the slope parameter β obtained with the recursive formulation and the Feasible GLS (homoskedastic case).

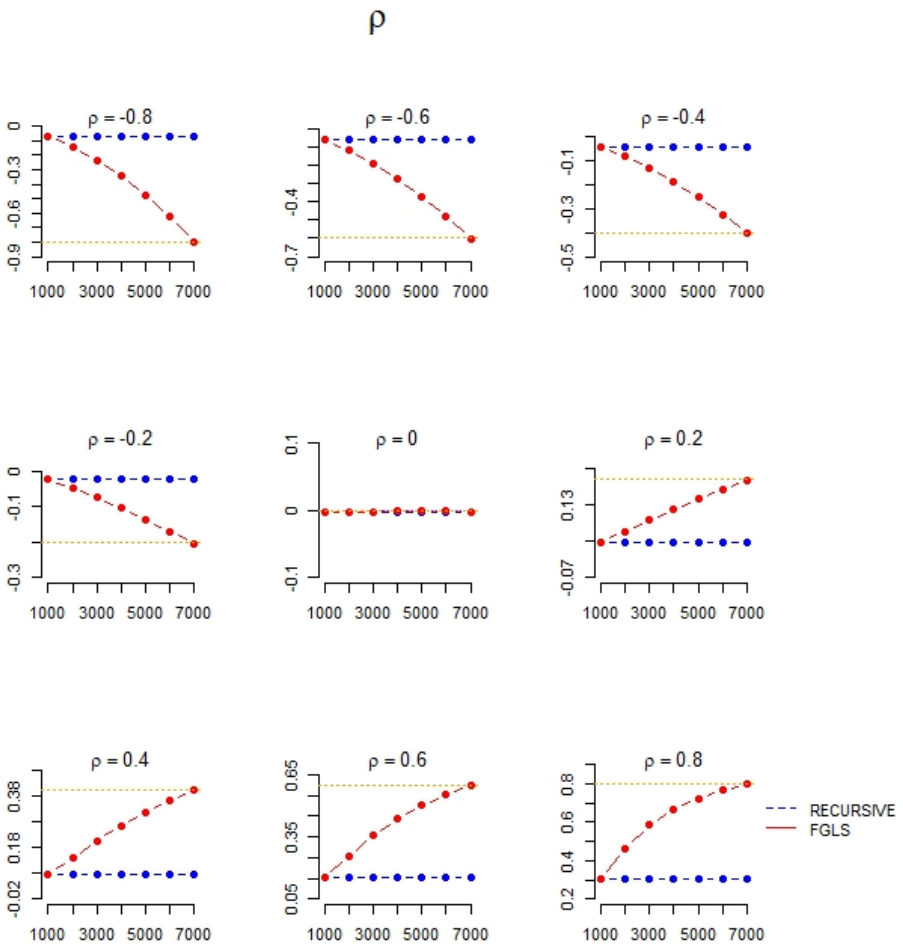

Figure 2.2: Bias of the spatial correlation parameter ρ obtained with the recursive formulation and the Feasible GLS (homoskedastic case).

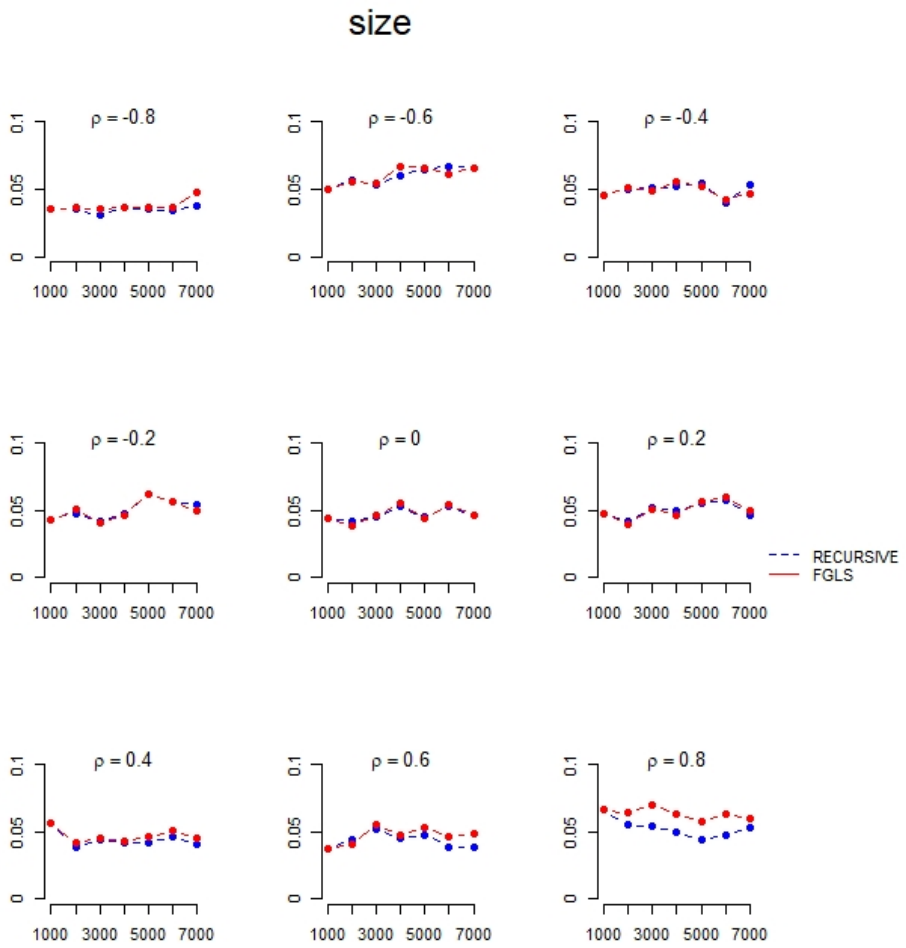

Figure 2.3: Comparison of the size of the test on β obtained with the recursive formulation and the Feasible GLS (homoskedastic case).

The plots reported in Figure 2.3 represent the size of a test on the slope coefficient β. Each plot corresponds to a different value of ρ. It is important to keep in mind that the size of the test depends upon the variance of β, that, in turns, depends on the error variance. The estimated variance of the error is based on the spatial weighting matrix as well as the estimated value of the spatial parameter. In light of this, the evidence reported in Figure 2.3 is quite surprising in that the size of the test is very close to the nominal 5% level for both approaches! While this should be expected for the FGLS, the excellent performance is not obvious for the recursive approach that keeps ρ fixed over the different intervals. However, this conclusion seems to be consistent

with the findings Kramer and Donninger (1987), who demonstrated that the relative efficiency of SEM is not affected by the spatial correlation parameter under certain regularity conditions.

Finally, let us turn our attention to the power of the test. The power of the test on β (Figure 2.4) is only reported at the end of the stream (i.e., when $s = 7$). Similar conclusions can be inferred since also the power of the test does not seem to be affected by the approximation of ρ.

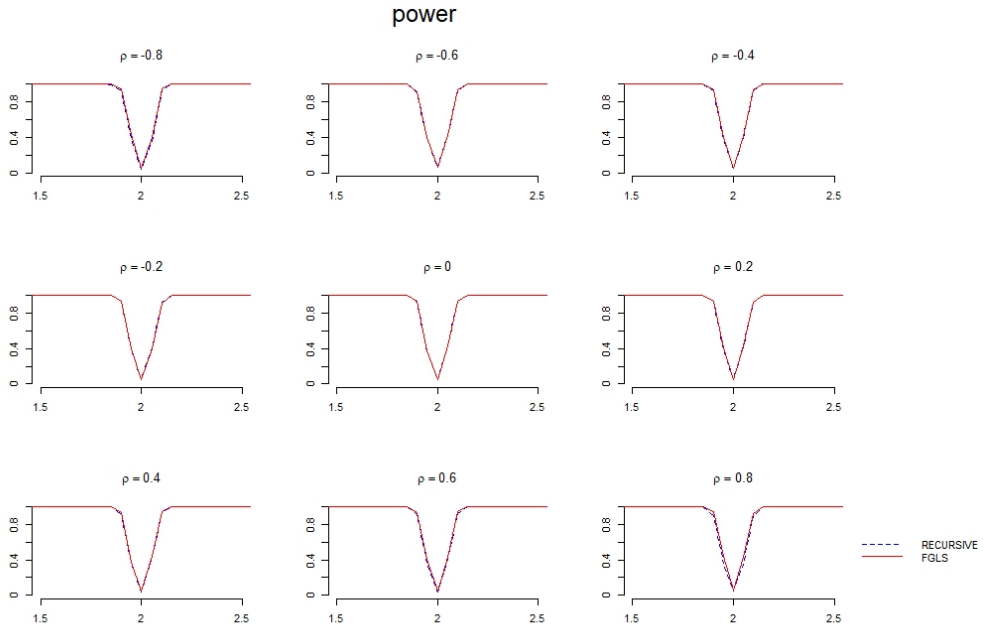

Figure 2.4: Comparison of the power of the test on β obtained with the recursive formulation and the Feasible GLS (homoskedastic case). The results refer to the last step of the streaming.

Since the general statistical performance of the recursive model is fully comparable with the standard method, we move on to examine the computational time involved with it and to compare it with the one required by the FGLS. For this purpose, we simulate a streaming of data where the sample size increases from 25,000 to 250,000 and, at each steps, 25,000 new observations are made available. Following a traditional approach, one would need to compute the spatial weighting matrix at each steps. The major advantage of the recursive approach resides mostly in the fact that W only needs to be calculated once at the beginning of the steaming process and then continuously updated as described in Section 2. In a spatial context, the computational time, in addition to the sample size, is strictly related to the density of W, particularly when using the GMM approach. In this chapter, to construct W we draw two sets of coordinates

from a uniform process over the interval (0;2). To maintain more or less constant the percentage of non-zero elements of the spatial weighing matrix (i.e., to 1.5%), we consider different cut-off levels for the distances. The distances were calculated using the euclidean distance.

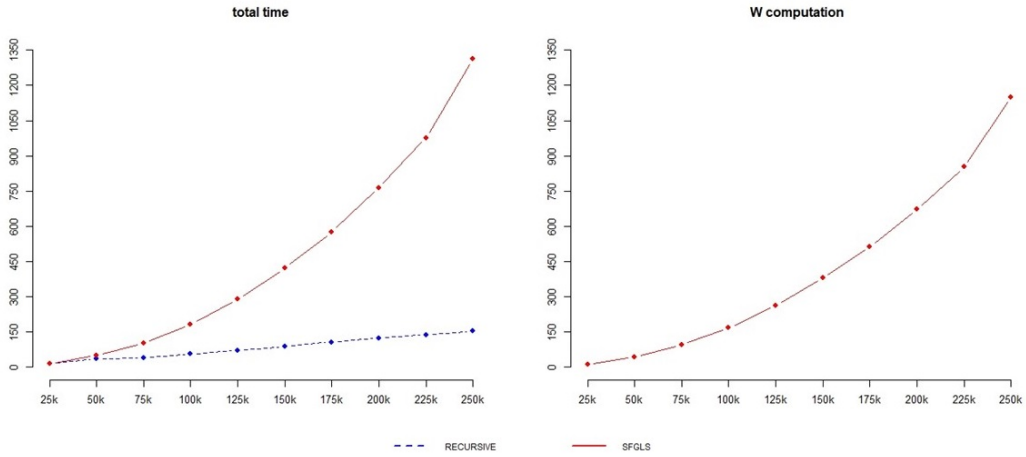

Figure 2.5: Computational time (measured in second) required from the recursive method and the Feasible Generalized Least Square to estimate the Spatial Error Model and to compute the weight matrix W.

Figure 2.5 is made up of two panels: the left hand side panel reports the comparison in computational time for the recursive and the FGLS while the right hand side panel highlights the time spent to compute W for each new streaming. Time is measured in seconds on the vertical axes. The computing time for the W matrix uses the function *dnearneigh* in the `spdep` package (Bivand and Piras; 2015; R Core Team; 2014), and then the auxiliary functions `nb2listw` (to transform the list of neighbors into an object of class `listw`) and `listw2dgCMatrix` (to transform the `listw` into a sparse matrix from the package `Matrix` Bates and Maechler (2019)). For estimation we used the function *spreg* in the `sphet` package (Piras; 2010; Bivand and Piras; 2015). Focusing on the left panel, one can immediately observe that for the initial sample size the timing of both procedures is exactly the same. This is not at all surprising because the streaming needs to be "initialized" and so the first step is the same. However, while the computational time for the recursive approach (the blue curve in the graph) is almost linear, the red line follows an exponential shape that, with the increase of sample size, gets more and more distant from the blue line. Interestingly, the right hand side panel shows that most of the difference between the two approaches is due to the computation of the spatial weights matrix.

2.4 Conclusions

In this chapter, we proposed a new approach, which allows for a recursive estimation of a spatial error regression model when spatial data are collected in streaming. This approach is based on an approximation and a “restriction".

The approximation is that the spatial correlation parameter is assumed to be stable and so it can be estimated only once and then kept fixed for the following streaming steps.

The restriction relates to the fact that the spatial weighting matrix should be row-standardized and based on of a cut-off distance. Based on the results of our Monte Carlo simulations, we can conclude that these two hypothesis do not deteriorate the performance of the estimated model and that the proposed methodology produces relevant advantages in terms of the computational effort.

Although limited by the specific hypotheses considered in this chapter, our results can be extremely useful in many practical instances when data are gathered continuously through time and data-based decisions need to be taken timely and effectively like, for instance, in situations related to health surveillance (e.g., in monitoring the geographical spread of an epidemic) or to environmental emergency (e.g., in motoring extreme events like floods or hurricane spatial diffusion).

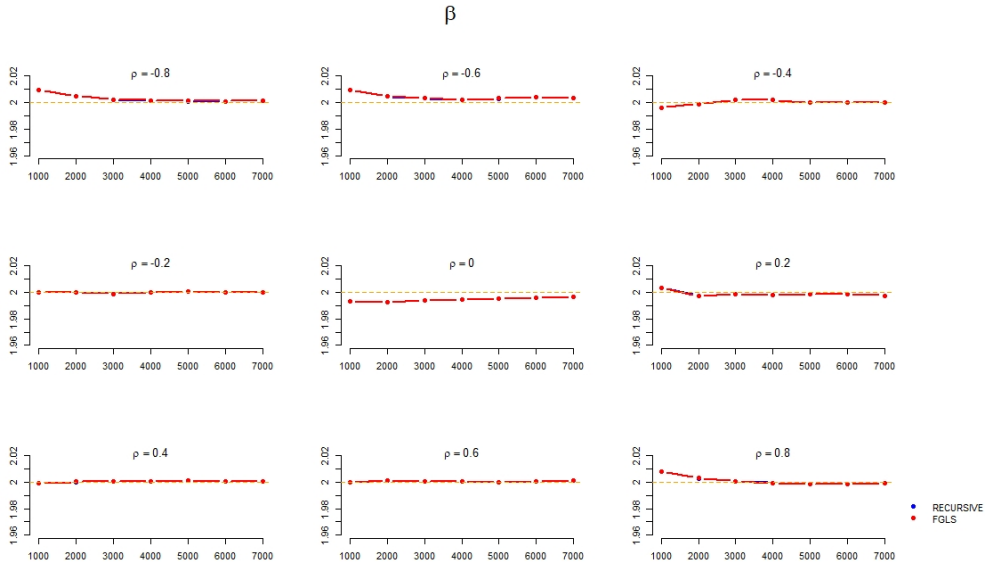

Figure 2.6: Bias of the slope parameter β obtained with the recursive formulation and the Feasible GLS (heteroskedastic case).

Appendix A

In this Appendix, Figures 2.6-2.9 report the simulation results for the heteroskedastic model.

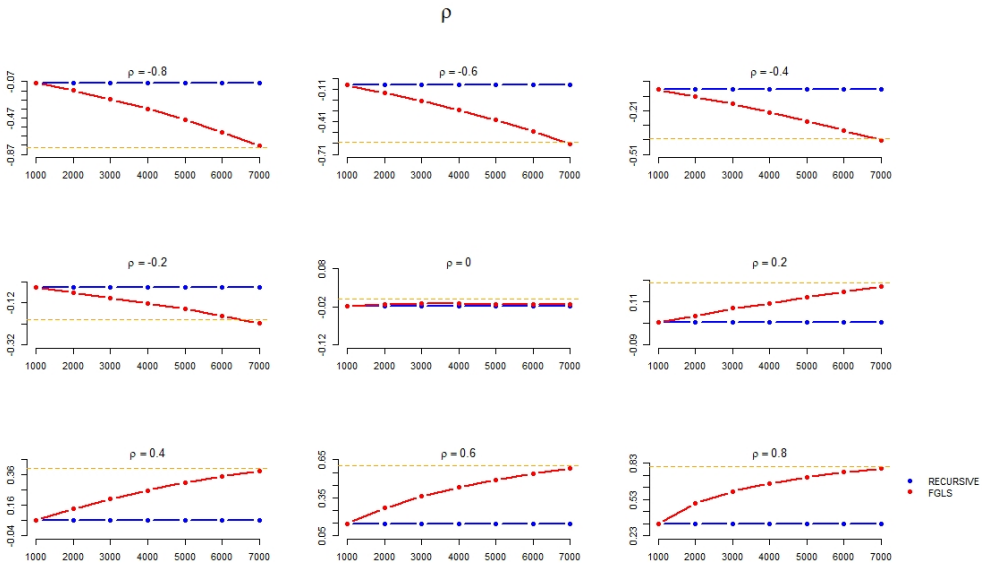

Figure 2.7: Bias of the spatial correlation parameter ρ obtained with the recursive formulation and the Feasible GLS (heteroskedastic case).

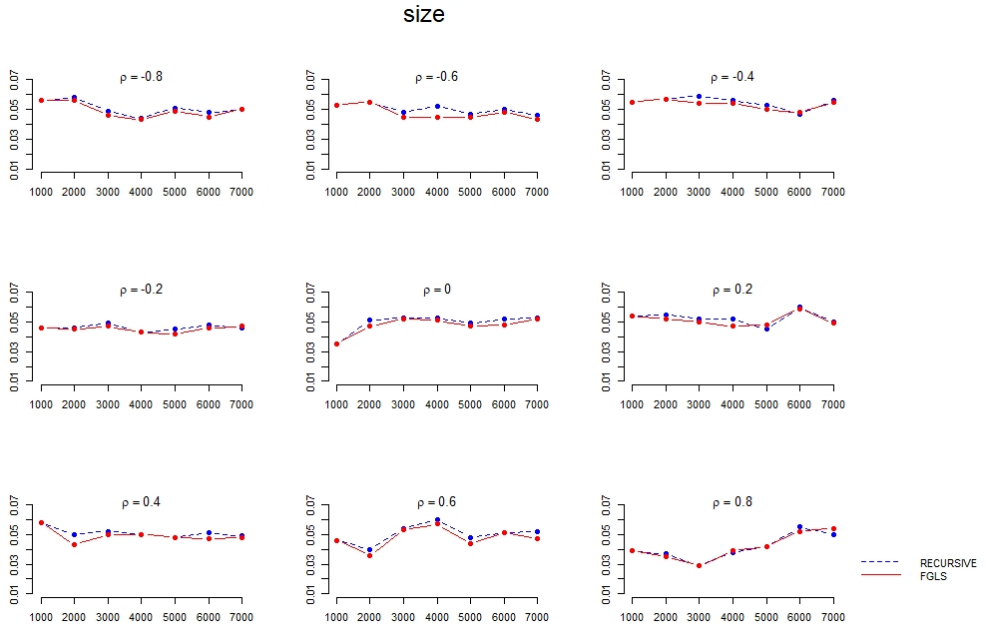

Figure 2.8: Comparison of the size of the test on β obtained with the recursive formulation and the Feasible GLS (heteroskedastic case).

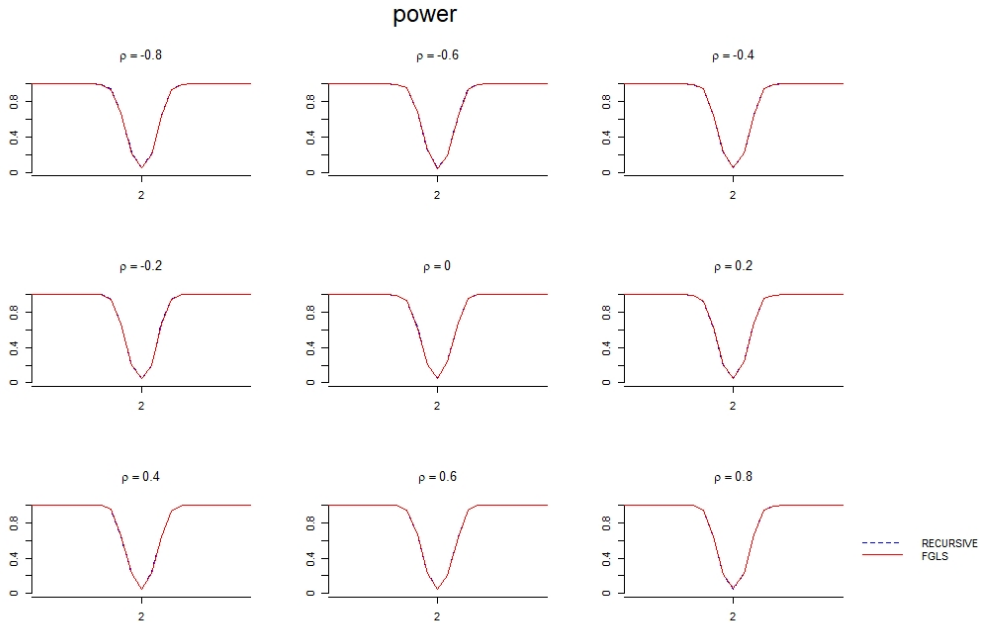

Figure 2.9: Comparison of the power of the test on β obtained with the recursive formulation and the Feasible GLS (heteroskedastic case). The results refer to the last step of the streaming.

Bibliography

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Boston: Kluwer Academic Publishers.

Arbia, G. (2014). *A primer for Spatial Econometrics*, Palgrave Macmillan.

Arbia, G., Solano-Hermosilla, G., Micale, F., Nardelli, V. and Genovese, G. (2018). Post-sampling crowdsourced data to allow reliable statistical inference : the case of food price indices in nigeria, *49th Conference of the Italian Statistical Association*.

Armstrong, M. P., Wang, S. and Zhang, Z. (2019). The internet of things and fast data streams: prospects for geospatial data science in emerging information ecosystems, *Cartography and Geographic Information Science* **46**(1): 39–56.
URL: *https://doi.org/10.1080/15230406.2018.1503973*

Bar-Yossef, Z., Jayram, T., Kumar, R. and Sivakumar, D. (2004). An information statistics

approach to data stream and communication complexity, *Journal of Computer and System Sciences* **68**: 702 – 732.

Barry, R. and Pace, R. (1999). Monte Carlo estimates of the log determinant of large sparse matrices, *Linear Algebra and its Applications* **289**(1-3): 41–54.

Bates, D. and Maechler, M. (2019). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-18.
URL: *https://CRAN.R-project.org/package=Matrix*

Bivand, R., Hauke, J. and Kossowski, T. (2013). Computing the jacobian in gaussian spatial autoregressive models: An illustrated comparison of available methods, *Geographical Analysis* **45**: 150–179.

Bivand, R. and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics, *Journal of Statistical Software* **63**: 1–36.

Créquit, P., Mansour, G., Benchoufi, M., Vivot, A. and Ravaud, P. (2018). Mapping of crowdsourcing in health: Systematic review, *J Med Internet Res* **20**: e187.

Drukker, D. M., Egger, P. and Prucha, I. R. (2013). On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors, *Econometric Reviews* **32**(5-6): 686–733.

Escobar, L. and Moser, E. (1993). A note on the updating of regression estimates, *The American Statistician* **47**: 192–194.

Felber, D. (2015). *Order statistics and variability in data stream*, Master's thesis, UCLA Electronic Theses and Dissertations.

Fox, G. C., Aktas, M. S., Aydin, G., Bulut, H., Pallickara, S., Pierce, M., Sayar, A., Wu, W. and Zhai, G. (2006). Real time streaming data grid applications, *in* F. Davoli, S. Palazzo and S. Zappatore (eds), *Distributed Cooperative Laboratories: Networking, Instrumentation, and Measurements*, Springer US, Boston, MA, pp. 253–267.

Galić, Z., Baranović, M., Križanović, K. and Mešković, E. (2014). Geospatial data streams: Formal framework and implementation, *Data Knowledge Engineering* **91**: 1 – 16.
URL: *http://www.sciencedirect.com/science/article/pii/S0169023X14000160*

Gertz, M., Hart, Q., Rueda, C., Singhal, S. and Zhang, J. (2006). A data and query model for streaming geospatial image data, *in* T. Grust, H. Höpfner, A. Illarramendi, S. Jablonski, M. Mesiti, S. Müller, P.-L. Patranjan, K.-U. Sattler, M. Spiliopoulou and J. Wijsen (eds), *Current Trends in Database Technology – EDBT 2006*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 687–699.

Griffith, D. (2004). Faster Maximum Likelihood Estimation of Very Large Spatial Autoregressive Models: An Extension of the Smirnov-Anselin Result, *Journal of Statistical Computation and Simulation* **74**(12): 855–866.

Hiriotappa, K., Thajchayapong, S., Chaovalit, P. and Pongnumkul, S. (2017). A streaming algorithm for online estimation of temporal and spatial extent of delays, *Journal of Advanced Transportation* **vol.2017**: Article ID 4018409.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural Computation* **9**: 1735–1780.

Kelejian, H. H. and Prucha, I. R. (2010). Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances, *Journal of Econometrics* **157**(1): 53–67.

Kelejian, H. and Prucha, I. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances, *The Journal of Real Estate Finance and Economics* **17**: 99–121.

Kramer, W. and Donninger, C. (1987). Spatial autocorrelation among errors and the relative efficiency of ols in the linear regression model, *Journal of the American Statistical Association* **82**: 577–579.

LeSage, J. and Pace, K. (2009). *Introduction to Spatial Econometrics*, CRC Press.

Morales, S. and Klette, R. (2013). Kalman-filter based spatio-temporal disparity integration, *Pattern Recognition Letters* **34**: 873–883.

Muckell, J., Hwang, J.-H., Lawson, C. T. and Ravi, S. S. (2010). Algorithms for compressing gps trajectory data: An empirical evaluation, *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, ACM, pp. 402–405.

Ord, K. (1975). Estimation methods for models of spatial interaction, *Journal of the American Statistical Association* **70**: 120–126.

Oud, J. H. L., Folmer, H., Patuelli, R. and Nijkamp, P. (2012). Continuous-time modeling with spatial dependence, *Geographical Analysis* **44**: 29–46.

Pace, R. and Barry, R. (1997). Quick computation of spatial autoregressive estimators, *Geographics Analysis* **29**(3): 232–247.

Pace, R. and LeSage, J. (2004). Chebyshev approximation of log-determinants of spatial weight matrices, *Computational Statistics & Data Analysis* **45**(2): 179–196.

Patroumpas, K. and Sellis, T. (2012). Event processing and real-time monitoring over streaming traffic data, *in* S. Di Martino, A. Peron and T. Tezuka (eds), *Web and Wireless Geographical Information Systems*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 116–133.

Piras, G. (2010). sphet: Spatial models with heteroskedastic innovations in R, *Journal of Statistical Software* **35**(1): 1–21.
URL: *http://www.jstatsoft.org/v35/i01/*

Potamias, M., Patroumpas, K. and Sellis, T. (2006). Sampling trajectory streams with spatiotemporal criteria, *18th International Conference on Scientific and Statistical Database Management (SSDBM'06)*, pp. 275–284.

Putatunda, S. (2017). *Streaming data: New models and methods with applications in the transportation industry*, PhD thesis.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Raman, B. S. and Ali, M. (2010). Spatial data streaming or streaming spatial data: Just stream it the way you like, COM.Geo '10, Association for Computing Machinery, New York, NY, USA.
URL: *https://doi.org/10.1145/1823854.1823856*

Sargan, J. D. and Bhargava, A. (1983). Testing residuals from least squares regression for being generated by the gaussian random walk, *Econometrica* **51**: 153–174.

Seidou, O. and Ouarda, T. B. M. J. (2007). Recursion-based multiple changepoint detection in multiple linear regression and application to river streamflows, *Water Resources Research* **43**.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples), *Biometrika* **52**: 591–611.

Smirnov, O. and Anselin, L. (2001). Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach, *Computational Statistics & Data Analysis* **35**(3): 301–319.

Smirnov, O. and Anselin, L. (2009). An O(N) parallel method of computing the Log-Jacobian of the variable transformation for models with spatial interaction on a lattice, *Computational Statistics & Data Analysis* **53**(8): 2980 – 2988.

Tang, L.-A., Zheng, Y., Yuan, J., Han, J., Leung, A., Peng, W.-C. and Porta, T. L. (2014). A framework of traveling companion discovery on trajectory data streams, *ACM Trans. Intell. Syst. Technol.* **5**: 3:1–3:34.

Thakur, G. S., Bhaduri, B. L., Piburn, J. O., Sims, K. M., Stewart, R. N. and Urban, M. L. (2015). Planetsense: A real-time streaming and spatio-temporal analytics platform for gathering geo-spatial intelligence from open source data, SIGSPATIAL '15, Association for Computing Machinery, New York, NY, USA.
URL: *https://doi.org/10.1145/2820783.2820882*

Welch, G. and Bishop, G. (2006). An introduction to the kalman filter, *Proc. Siggraph Course* **8**.

Wermuth, N. (1992). On block-recursive linear regression equations, *Brazilian Journal of Probability and Statistics* **6**: 1–32.

Zhang, J. (2006). Spatio-temporal aggregation over streaming geospatial data, *In Proceedings of the 10th International Conference on Extending Database Technology Ph.D. Workshop*.

Chapter 3

On the efficiency of estimators of the Spatial Error Model

3.1 Introduction

It is widely known that the presence of spatial correlation in the disturbance term causes problems on statistical inference. In particular, OLS estimation while still unbiased, will not be efficient (Cressie; 1993; Schabenberger and Gotway; 2004). As a consequence of this, many researchers have estimated a Spatial Error Model (SEM) by taking a Generalized Least Square (GLS) approach. In a first step, the spatial parameter is obtained either using Maximum Likelihood (Ord; 1975) or Methods of Moments (Kelejian and Prucha; 1999). Then, an OLS model is estimated on the spatially filtered data.

However, Kramer and Donninger (1987) argued that, under certain conditions the relative efficiency of the GLS and the OLS in a spatial context is close to one. To prove this point, they make use of assumptions that are difficult to verify in practice. In this note we compare the two methods within the framework of a Monte Carlo experiment. Our results show that the performance of the OLS and GLS are very similar if compared in terms of the size and the power of a test of hypothesis on the slope coefficients of the model.

The note is organized as follows: Section 3.2 reviews the algebraic conditions that relates to the structure of the spatial weighting matrix. In Section 3.3, we discuss the details of the Monte Carlo experiment and comment on the results obtained. Finally, Section 3.4 concludes the Chapter.

3.2 Algebraic conditions

A general linear model is defined as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a fixed design matrix, with full rank, and $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ the covariance matrix of the errors.

Let us now consider the column space of $\boldsymbol{X}$ and $\boldsymbol{VX}$, where $\boldsymbol{X} = [\boldsymbol{x_1}...\boldsymbol{x_p}]$ and $\boldsymbol{V}$ is a general covariance matrix, that is

$$\mathcal{M} = span(\boldsymbol{X}) = \{\boldsymbol{x} | \boldsymbol{x} = a_1\boldsymbol{x_1} + ... + a_p\boldsymbol{x_p}, a_i \in \mathbb{R}\}$$
$$\mathcal{N} = span(\boldsymbol{VX}) = \{\boldsymbol{y} | \boldsymbol{y} = b_1\boldsymbol{Vx_1} + ... + b_p\boldsymbol{Vx_p}, b_i \in \mathbb{R}\}.$$

In the following, we list some of the general conditions on $\boldsymbol{V}$ and $\boldsymbol{X}$ in order for the GLS and OLS to have the same efficiency (Zyskind; 1967; Puntanen and Styan; 1989): [1]

C1: A subset of k eigenvectors of $\boldsymbol{V}$ exists, forming a basis for $\mathcal{M}$ (Anderson; 1948).

C2: The matrix $\boldsymbol{V}$ can be written as $\boldsymbol{V} = [\tilde{\boldsymbol{X}}\ \boldsymbol{Z}]\boldsymbol{D}[\tilde{\boldsymbol{X}}\ \boldsymbol{Z}]^T$, where the matrix $[\tilde{\boldsymbol{X}}\ \boldsymbol{Z}]$ is orthogonal, $\tilde{\boldsymbol{X}}$ is any orthonormal basis of $\mathcal{M}$, $\boldsymbol{Z}$ is any orthonormal basis of $\mathcal{M}^{\perp}$, and $\boldsymbol{D}$ is any diagonal matrix with non negative elements.

C3: If $\boldsymbol{H}$ denotes the orthogonal matrix projection operator on $\mathcal{M}$, then $\boldsymbol{HV} = \boldsymbol{VH}$.

C4: $\mathcal{M}$= $\mathcal{N}$; that is $\boldsymbol{X}$ is $\boldsymbol{V}$ invariant. It follows if $\mathcal{N} \subseteq \mathcal{M}$.

In practice, verifying either of these rules is not trivial, especially for large dimensions of the data matrix. In fact, the examples provided in the literature are, for the most part, unrealistic and they only deal with very small matrices (Puntanen and Styan; 1989; Isotalo and Puntanen; 2009).

For an OLS model, $\boldsymbol{V} = \sigma^2\boldsymbol{I}$ and therefore is quite simple to show that the four conditions above are satisfied. The SEM error covariance is $\boldsymbol{V} = \sigma^2[(\boldsymbol{I} - \rho\boldsymbol{W})^T(\boldsymbol{I} - \rho\boldsymbol{W})]^{-1}$, where $\boldsymbol{I} \in \mathbb{R}^{n \times n}$ is an identity matrix, $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ is the spatial weighting matrix, and ρ is the spatial error autocorrelation coefficient.

We have to keep in mind that, due to the presence of ρ, $\boldsymbol{V}$ is unknown. Formulating any of the above conditions as a function of ρ would be extremely challenging. As suggested from

[1] Note that the four conditions below are equivalent in the sense that if one is verified, also the other three are.

Kramer and Donninger (1987), it is possible to operate with $\boldsymbol{W}$ and $\boldsymbol{W}^T$ instead of $\boldsymbol{V}$. A brief demonstration of this assertion is provided below:
if $\boldsymbol{W} \in \mathcal{M}$ then

i $(\boldsymbol{I} - \rho\boldsymbol{W}) \in \mathcal{M}$ because linear transformation of $\boldsymbol{W}$

ii $(\boldsymbol{I} - \rho\boldsymbol{W})^{-1} \in \mathcal{M}$ because if $span(\boldsymbol{I} - \rho\boldsymbol{W}) = \gamma_1\boldsymbol{v}_1 + ... + \gamma_p\boldsymbol{v}_p$, then $span((\boldsymbol{I} - \rho\boldsymbol{W})^{-1}) = \frac{1}{\gamma_1}\boldsymbol{v}_1 + ... + \frac{1}{\gamma_p}\boldsymbol{v}_p$, where $\gamma_1, ..., \gamma_p$ are the eigenvalues.

The same holds for $\boldsymbol{W}^T$.

Thus, C4 becomes

$$span(\boldsymbol{W}\boldsymbol{X}) \subseteq \mathcal{M} \qquad span(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X}) \subseteq \mathcal{M}.$$

Unfortunately, unless $\boldsymbol{W}$ is symmetric, checking these conditions remains prohibitive. Kramer and Donninger (1987) also demonstrated that the relative efficiency between OLS and GLS equals one as long as the eigenvector $\boldsymbol{w}_{max}$ related to the maximum eigenvalue λ_{max} of $\boldsymbol{W}$ belongs to $\mathcal{M}$. The trivial case is when $\boldsymbol{W}$ is row- standardized. For a row-standardized $\boldsymbol{W}$, the maximum eigenvalue λ_{max} is one and, therefore, $\boldsymbol{w}_{max}$ is a unit vector. In this case, if the intercept is included in the model GLS and OLS have the same efficiency. The challenging aspect would be to investigate the case when $\boldsymbol{w}_{max}$ is not included in $\mathcal{M}$. Related to this, an interesting question would be how far $\boldsymbol{w}_{max}$ must be from $\mathcal{M}$ in order to significantly affect the efficiency. Another non-trivial aspect is the presence of complex eigenvalues for certain specifications of $\boldsymbol{W}$. Since it is no possible to sort complex numbers, it is complicated to identify the maximum eigenvalue. For the above reasons, we propose an empirical analysis in order to verify the assertions.

3.3 Monte Carlo Simulation

In this Section, we propose a Monte Carlo simulation, which aims to verify empirically a comparison between the performance of OLS and GLS. In the following we refers to the coordinates of the well known Boston dataset (Bivand et al.; 2013). We investigated the effect of different $\boldsymbol{W}$ matrices, ρ values and number of neighbors. We computed the $\boldsymbol{W}$ matrix with the k-nearest neighbours criterion with 5, 10 and 15 neighbours. In the first considered case, $\boldsymbol{W}$ is row standardized. In the second, we set the weights different from zero as the inverse distance between the points. Then, the matrix is standardized with minmax rule (Kelejian and Prucha; 2010). Our aim is to compare the efficiency of GLS and OLS, focusing on the size and the power of

the test over β. The setting of our simulation concerns a variable $\boldsymbol{X}$ sampled from a gaussian distribution (with mean equal to 0 and standard deviation 10) , $\beta = 1$ and $\sigma^2 = 3.5$. With regard to ρ, the values considered ranges between 0.1 and 0.8 .

Figure 3.1 and 3.2 summarizes the results of 3000 Monte Carlo samples in the case of row standardized and minmax standardized $\boldsymbol{W}$ respectively. We reported the size (I line) and the power (II line). For the sake of succinctness, we decide to print the power for $\rho = 0.8$ which is the situation with higher difference between the two models. On the columns we show instead the effect of the increasing number of neighbors.

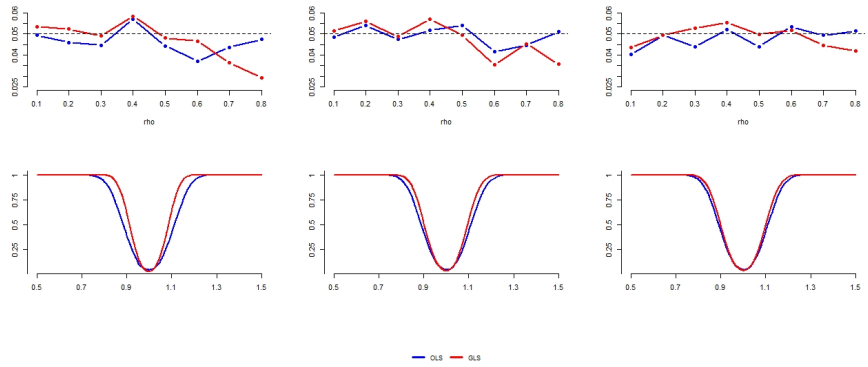

Figure 3.1: Results for row standardized $\boldsymbol{W}$: size (I row) and power for $\rho = 0.8$ (II row). The number of neighbours is respectively 5, 10 and 15.

As expected from theoretical properties, the performance of the models are totally comparable in the case of row standardized matrix (Figure 3.1). This is particularly true with the increase of the number of neighbors. Regarding the results for minmax standardized matrix (Figure 3.2), OLS and GLS seem to be also surprisingly similar.

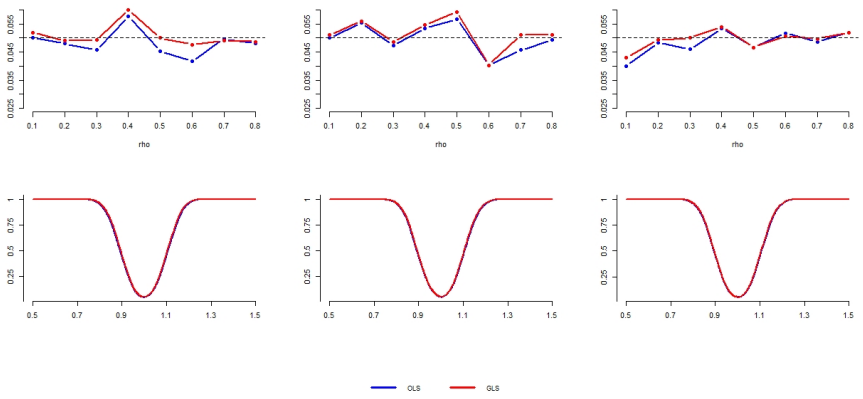

Figure 3.2: Results for minmax standardized $\boldsymbol{W}$: size (I row) and power for $\rho = 0.8$ (II row). The number of neighbours is respectively 5, 10 and 15.

To further clarify the results obtained, we also check if $\boldsymbol{w}_{max} \in \mathcal{M}$. Since some eigenvalues belong to $\mathbb{C}$, we decide to order them by the modulus in order to find the maximum one and the corresponding eigenvector $\boldsymbol{w}_{max}$. Then, we compute the projection $\boldsymbol{w}_{max}^{P}$ on $\mathcal{M}$. The distance between the two vector is evaluated as

$$d = \frac{\sum_i |\boldsymbol{w}_{max,i}^{P} - \boldsymbol{w}_{max,i}|}{n}.$$

Since the distance between $\boldsymbol{w}_{max}$ and $\mathcal{M}$ is greater the 0 (Table 3.1), $\boldsymbol{w}_{max}$ does not belong to the subspace. However, the empirical results suggest that is not far enough to lead to a worsening in the OLS method.

	5 neigh	10 neigh	15 neigh
λ_{max}	0.7415	0.6858	0.7007
d	0.0095	0.0192	0.0225

Table 3.1: Maximum eigenvalue for minmax standardized $\boldsymbol{W}$ and distance between the relative eigenvector $\boldsymbol{w}_{max}$ and the column space of $\boldsymbol{X}$.

Since OLS is a particular case of GLS when ρ is equal to zero, we report the result of a further experiment which consists in estimating a GLS by imposing a different ρ from the one used to generate the data. We sample data from a SEM with a value of ρ equal to 0.7. Then we estimate GLS filtering $\boldsymbol{Y}$ and $\boldsymbol{X}$ imposing wrong value of ρ from 0 to 0.8.

Figure 3.3 displays the comparison of the size of the estimation of GLS with the true parameter and with a wrong one referring to 3000 Monte Carlo samples. As we expected, we observe how a misestimated value of ρ seems to not affect in a significant manner the performance of the model.

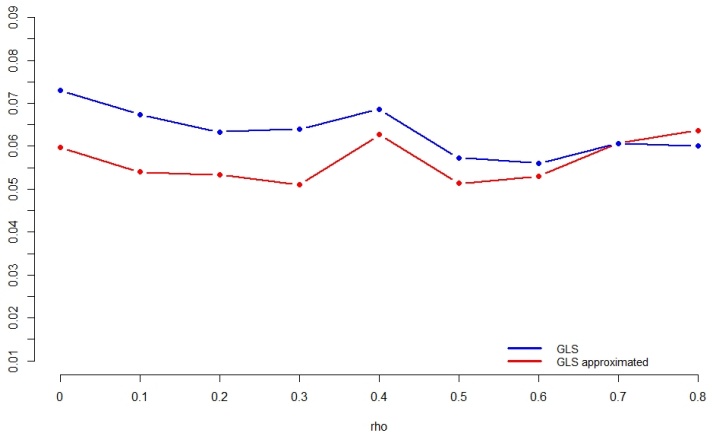

Figure 3.3: Size of the estimation of a GLS with correct and wrong values of the spatial parameter ρ.

3.4 Conclusions

In this note we reviewed the main properties, which guarantee same efficiency in OLS and GLS results. In the trivial of row-standardized weight matrices, the relative efficiency is one. In addition, we observe that also in other specification of $\boldsymbol{W}$, when the properties are almost true, the difference between the two methods seem to not appear significant.

Bibliography

Anderson, T. (1948). On the theory of testing serial correlation, *Scandinavian Actuarial Journal* **1948**: 88–116.

Bivand, R. S., Pebesma, E. and Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition*, Springer, NY.

Cressie, N. (1993). *Statistics for Spatial Data*, John Wiley & Sons, Inc.

Isotalo, J. and Puntanen, S. (2009). A note on the equality of the olse and the blue of the parametric function in the general gauss–markov model, *Statistical Papers* **50**: 185–193.

Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model, *International Economic Review* **40**: 509–533.

Kelejian, H. H. and Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances, *Journal of Econometrics* **157**: 53 – 67. Nonlinear and Nonparametric Methods in Econometrics.

Kramer, W. and Donninger, C. (1987). Spatial autocorrelation among errors and the relative efficiency of ols in the linear regression model, *Journal of the American Statistical Association* **82**: 577–579.

Ord, K. (1975). Estimation methods for models of spatial interaction, *Journal of the American Statistical Association* **70**: 120–126.

Puntanen, S. and Styan, G. P. H. (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator, *The American Statistician* **43**: 153–161.

Schabenberger, O. and Gotway, C. (2004). *Statistical Methods for Spatial Data Analysis*, Chapman and Hall/CRC.

Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models, *Ann. Math. Statist.* **38**: 1092–1109.

Chapter 4

Modelling Non-Stationary Spatial Lag Model with Hidden Markov Random Fields

4.1 Introduction

In spatial statistics and econometrics, it is common practice to model relationships among variables using Cliff-Ord-type linear regression models (Arbia; 2006; LeSage and Pace; 2010). One basic assumption which is rarely questioned in the literature, is second-order stationarity, which in particular can assume the form of homogeneity (i.e. invariance to rigid motion of the generating process) or isotropy (i.e. invariance to rotations) (Christakos; 1992; Arbia; 2006). If a random field is second-order stationary, the associated parameters will be invariant to the statistical units position. In particular, in a linear spatial regression model both the spatial correlation and the regression coefficients are constant over space implying that a spatial relationship depends only on the distance between each pair of spatial units irrespective of other possible directional effects.

However, such an assumption is often unrealistic when using a spatial random field to model socio-economic and epidemiological phenomena. In many empirical situations, indeed, it is unreasonable to believe that the relationship between variables could be a realization of a unique stationary process and that the model's parameters remain unchanged in different portions of the study area and it is, indeed, more sensible to conjecture that the parameters vary in space

according to some regular pattern. For instance, if we are dealing with a large sample size observed on a large geographical area (e.g. a large country or a continent), it is more sensible to build a more flexible regression mechanism where the relationship between variables is allowed to change smoothly from one location to the other. In this Chapter we are interested, in particular, to model different spatial behaviors that are not directly observable, as they are due to hidden characteristics that changes smoothly over space.

In the literature, there are several attempts to approach the problem of spatially varying coefficients. A first approach refers to the concept of scan statistics (Fisher; 1959; Glaz et al.; 2001). Following such an approach, rather than considering all available observations together, we select subsamples that are geographically close and computations are performed within these subsamples. Arbia (1990) proposes the use of scan statistics to study first- and the second-order spatial non-stationarity, Ott and Hoh (2012) provide an interesting example of their use to genome screening, while Páez et al. (2008) show an application in the modelling of hedonic house prices. A second approach is based on the *Locally Weighted Regression* (LWR) (Cleveland and Devlin; 1988). LWR is a non-parametric technique, which can be compared to scan statistic used to estimate a regression around a point of interest using only a limited number of training points (McMillen and McDonald; 2004). A particular case of LWR is the so-called *Geographically Weighted Regression* (GWR) (Brunsdon et al.; 1996, 1998; Fotheringham et al.; 1998; Brunsdon et al.; 2007; Fotheringham et al.; 2002; McMillen and McDonald; 2004), that makes use of the geographical space as a selection criterion. Following a different line of though, Gelfand et al. (2003) propose to impose a Bayesian structure to individual regression coefficients that are spatially correlated, with a covariance matrix that takes into account the distances between them. Another interesting approach is proposed by Li and Sang (2019) who suggest to estimate point-specific regressor coefficients by penalizing pairs of neighbors with different parameter values. The methodology also includes the possibility of a varying spatial correlation coefficient.

Some more recent improvements have been developed in the field of anisotropy suggesting to estimate the spatial correlation coefficient as a function of one or more anisotropy-generating variables (Deng; 2008) or modelling directly anisotropy both in the space (Arbia et al.; 2018) and in the frequency domain (Santi et al.; 2017).

Our approach tackles the problem of parameters non-stationarity trying to identify groups of observations that share similar behaviours, through the modelling of a latent process. More specifically, the latent process evolves as a Hidden Markov Random Field (Besag; 1986). This assumption allows us to consider the spatial structure typical of the models commonly used in

the spatial literature and, in this way, to capture similarities in the neighbourhood. Notable articles in this area (Charras-Garrido et al.; 2011; Green and Richardson; 2002; Spezia et al.; 2018) offer good examples of possible applications in disease mapping and species distribution.

In this spirit, we introduce a latent process evolving according to a Potts model (Okabayashi et al.; 2011), which drives the value of the coefficients in a Cliff-Ord-type spatial autoregressive linear model. Since it is reasonable that close-by groups of observations behave in a similar fashion rather than being different in each point, our choice appears to be realistic and it may greatly improve the interpretation of the model output. It is also worth recalling that the Potts model definition allows us to consider similarities between neighbouring points based only on the value of the latent process in the neighbourhood. This is a crucial characteristic because, in this way, it does not require the definition of a priori possible causes of non-stationarity. From this point of view, our approach can be seen as an evolution of the so-called wombling problem (Lu and Carlin; 2005). In wombling, the interest focuses only in finding borders, i.e. adjacent points with dramatic differences in the response variable or in some of the covariates (Lee; 2013). In this sense, the criterion we suggest to detect different groups is a generalization of the wombling problem. However, while wombling methods are based on dissimilarity matrices defined on the basis of available covariates, the strength of our approach is the possibility to obtain groups identification with no need of any further information.

The layout of the Chapter is the following. In Section 2 we specify the model and we describe the Bayesian strategy for the estimation. Section 3 reports the results of our Monte Carlo simulations. In Section 4 we apply our method to the well known Boston dataset. In the end, conclusion follows in Section 5.

4.2 A spatially varying regression model

In this section, we provide an exhaustive explanation of the model and of the estimation scheme. A criterion to select the best number of groups k is also provided.

4.2.1 Model assumptions

In this section we will illustrate our approach by making reference to the Spatial Lag Model (SLM) which can be formulated as follows:

$$Y_i = X_i\beta + \rho \boldsymbol{w}_i \boldsymbol{Y} + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

where the values X_i and Y_i are collected in the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively, with $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^n$. The vector $\boldsymbol{w_i}$ is the $i-th$ row of $\boldsymbol{W} \in \mathbb{R}^{n \times n}$. $\boldsymbol{W}$ is the classical weight matrix which describes the topology of spatial system. For further details about SLM, among others, we refer to Arbia (2014) and Banerjee et al. (2014). Within this framework, assuming a more flexible definition of the SLM we can relax the hypothesis of stationariety and we assume the following:

$$Y_i = X_i\beta_i + \rho_i \boldsymbol{w}_i \boldsymbol{Y} + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_i^2), \tag{4.1}$$

where now β_i and ρ_i are observation-specific parameters. Equation (1) expresses non-stationarity in its extreme formulation by assuming a different parameter value for each location. However, in practice it is more reasonable to postulate an intermediate formulation with groups of observations which behave in a similar way. For this reason, we will assume heterogeneity that can be successfully modelled by the introduction of a discrete latent process capturing the structure of the unknown groups. Observations in the same group share the same value of the latent process and, hence, of the parameters. As already mentioned, the factors giving rise to non-stationariety could be unknown and, in general, they are not captured in an exhaustive way by the covariates. For this reason we introduce a spatial latent process evolving as an Hidden Markov Random Field. The main feature of this structure is that the value of the latent variable value depends only on the value of the variables associated to its neighbours. This seems to be particularly suitable for the kind of problems at issue.

In our model, we propose a latent process $\boldsymbol{u}$ evolving according to a Potts model (Besag; 1986). We adapt the structure to a K possible groups among the data, thus, we include a multinomial latent variables $u_i \in \{1, ..., K\}, i = 1, ..., n$. It follows that

$$\beta_i = \sum_{k=1}^{K} \gamma_k \mathbb{1}(u_i = k), \quad \rho_i = \sum_{k=1}^{K} \psi_k \mathbb{1}(u_i = k), \quad \sigma_i^2 = \sum_{k=1}^{K} \nu_k \mathbb{1}(u_i = k).$$

To introduce our approach, let us define the diagonal matrices of parameters $\boldsymbol{R} = \text{diag}(\rho_1, ..., \rho_n)$, $\boldsymbol{B} = \text{diag}(\beta_1, ..., \beta_n)$ and $\boldsymbol{S} = \text{diag}(\sigma_1^2, ..., \sigma_n^2)$ and let us also define the matrix $\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{RW})^{-1}$. Adopting this notation, the likelihood of model (1) can be written as:

$$\boldsymbol{Y}|\boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{MBX}, \boldsymbol{MSM}^T).$$

The hierarchical structure of the latent process $\boldsymbol{u}$ is

$$u_i|\boldsymbol{u}_{-i}, \boldsymbol{\theta} \sim Multinom(p_{1,i}(\boldsymbol{u}_{-i}), ..., p_{K,i}(\boldsymbol{u}_{-i}))$$

where $\boldsymbol{u}_{-i}$ indicates the vector $\boldsymbol{u}$ without the $i-th$ element. Given z_k as the number of neighbours of i within group k, the success probability is given by:

$$p_{k,i}(\boldsymbol{u}_{-i}) = P(u_i = k|\boldsymbol{u}_{-i}, \boldsymbol{\theta}) = \begin{cases} \frac{1}{1+\sum_{k=2}^{K} e^{-(\theta_{1,k}-\sum_{j=2}^{K}\theta_{j,k}z_k)}} & k = 1 \\ \frac{e^{-(\theta_{1,k}-\sum_{j=2}^{K}\theta_{j,k}z_k)}}{1+\sum_{k=2}^{K} e^{-(\theta_{1,k}-\sum_{j=2}^{K}\theta_{j,k}z_k)}} & k = 2, ..., K. \end{cases}$$

It is worth recalling that the parameter $\theta_{1,k}$ models the baseline effect, while $\theta_{j,k}$ is a penalization parameter that discourages the membership to different groups of adjacent elements.

The priors distribution of the model parameters are set in a standard non-informative way (Gelman; 2006; Banerjee et al.; 2014; Lingren and Rue; 2015) as follows:

$$\begin{aligned} &\gamma_k \overset{ind}{\sim} \mathcal{N}(0, \sigma^2_\gamma) && k = 1, ..., K, \\ &\text{logit}(\psi_k) \overset{ind}{\sim} \mathcal{N}(0, \sigma^2_\psi) && k = 1, ..., K, \\ &\nu_k \overset{ind}{\sim} inv\Gamma(a, b) && k = 1, ..., K, \\ &\theta_{j,k} \overset{ind}{\sim} \mathcal{N}(0, \sigma^2_\theta) && k = 2, ..., K \quad j = 1, ..., K, \end{aligned}$$

with hyperparameters $\sigma^2_\gamma = \sigma^2_\theta = 100$, $\sigma^2_\psi = 10$, and $a = b = 0.001$.
Henceforth, we will refer to our model as Non Stationary Spatial Lag Model (NS-SLM). In specifying of our model using the Hidden Markov Random Field framework, we encounter the so-called label switching problem (Jasra et al.; 2005; Bartolucci et al.; 2012), which consists in a permutation of the clusters labels that can occur during the Markov chain Monte Carlo (MCMC) estimation process. As a consequence of this phenomenon, even if the observations are correctly clustered, it would be impossible to properly identify groups from the posterior. Following Bertarelli et al. (2018), we tackle this issue by ordering the MCMC output at every iteration with respect to the estimated values of ϕ. In particular, we assign the label 1 to the group with smallest ϕ, the label 2 to the group with the second smallest ϕ and so on.

4.2.2 Bayesian Inference

Following the Bayesian paradigm, we propose to estimate our proposed NS-SLM using a MCMC algorithm. From the prior specifications given in Section 2.1, it follows that the posterior distribution of the model parameters is given by:

$$p(\boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{u}|\boldsymbol{Y}) \propto L(\boldsymbol{Y}|\boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{u}, \sigma^2)\pi(\boldsymbol{\gamma})\pi(\boldsymbol{\psi})p(\boldsymbol{u}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\pi(\sigma^2). \tag{4.2}$$

Equation (2) suggests that $\boldsymbol{u}, \boldsymbol{\gamma}$ and σ^2 can be updated by Gibbs steps because the full conditionals are available analytically and are easy to sample from. Let us define $\tilde{\boldsymbol{Y}} = \boldsymbol{M}^{-1}\boldsymbol{Y}$

and let us further divide the data into the vectors $\tilde{\boldsymbol{Y}}_k, \boldsymbol{X}_k$ depending on the value of the underlying latent process $\boldsymbol{u}$; then the full-conditional distributions are expressed as:

$$\gamma_k|\boldsymbol{Y}, u = k, \psi_k, \nu_k \sim \mathcal{N}\left(\left(\frac{1}{\sigma_\gamma^2} + \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{\nu_k}\right)^{-1}\frac{\boldsymbol{X}_k^T\tilde{\boldsymbol{Y}}_k}{\nu_k}, \left(\frac{1}{\sigma_\gamma^2} + \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{\nu_k}\right)^{-1}\right),$$

$$\nu_k|\boldsymbol{Y}, \boldsymbol{u}, \boldsymbol{\psi}, \boldsymbol{\gamma} \sim inv\Gamma\left(a + \frac{n_k}{2}, b + \frac{[\tilde{\boldsymbol{Y}}_k - \gamma_k\boldsymbol{X}_k]^T[\tilde{\boldsymbol{Y}}_k - \gamma_k\boldsymbol{X}_k]}{2}\right),$$

$$u_i|\boldsymbol{Y}, \boldsymbol{u}_{-i}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\theta} \sim Multinom(q_{1,i}(\boldsymbol{u}_{-i}), ..., q_{K,i}(\boldsymbol{u}_{-i})),$$

where

$$q_{k,i}(\boldsymbol{u}_{-i}) = \frac{L(Y_i|\boldsymbol{Y}_{-i}, u_i = k, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\nu}) \cdot p_{k,i}(\boldsymbol{u}_{-i})}{\sum_{j=0}^{K-1} L(Y_i|\boldsymbol{Y}_{-i}, u_i = j, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\nu}) \cdot p_{j,i}(\boldsymbol{u}_{-i})}.$$

The details of the computation of the full-conditional distributions are reported in Appendix A. The remaining parameters $(\boldsymbol{\psi}, \boldsymbol{\theta})$ are updated by a Metropolis-Hasting step where adopt a Gaussian proposal for both parameters. Notice that, since the calculation of the distribution of $\boldsymbol{u}|\boldsymbol{\theta}$ is computationally expensive, we approximate such distribution by the pseudo-likelihood (Besag; 1986):

$$P(\boldsymbol{u}|\boldsymbol{\theta}) \approx \prod_i P(u_i|\boldsymbol{u}_{-i}, \boldsymbol{\theta}) = \prod_i P(u_i|\boldsymbol{u}_{i\sim j}, \boldsymbol{\theta}).$$

4.2.3 Model comparison

An important aspect of our proposed model is constituted by the identification of the number of groups. However, such identification in an optimal way, in general, is not trivial. Our proposal is fixing the maximum number of groups, say K, and to estimate models K different models, say $m_1, \cdots, m_K$, respectively, with $1, \cdots, K$ groups and then to compare them in order to choose a posteriori the best one. We operate a model selection by using the following approximation of the probability of model m_i involving the integrated likelihood:

$$\frac{p(m_i|i, \boldsymbol{Y})}{p(m_j|j, \boldsymbol{Y})} = \frac{p(\boldsymbol{Y}|m_i)}{p(\boldsymbol{Y}|m_j)} \times \frac{p(m_i)}{p(m_j)} \tag{4.3}$$

where $p(m_i)$ is the prior probability of model m_i with i groups. Following Zucchini and MacDonald (2009), we approximate numerically the integrated likelihood $p(\boldsymbol{Y}|m_i)$. Let B be the number of samples in the MCMC algorithm, for the sake of brevity we indicate with $\boldsymbol{\Theta}_j$ all the parameters of model m_j. It has been proved (Congdon; 2006) that

$$p(m_j|j, \boldsymbol{Y}) \approx B^{-1}\sum_{i=1}^{B} p(m_j|j, \boldsymbol{Y}, \boldsymbol{\Theta}_j^{(i)}) = B^{-1}\sum_{i=1}^{B} \frac{p(\boldsymbol{Y}|\boldsymbol{\Theta}_j^{(i)}, m_j)p(\boldsymbol{\Theta}_j^{(i)}|m_j)p(m_j)}{\sum_{j=1}^{K} p(\boldsymbol{Y}|\boldsymbol{\Theta}_j^{(i)}, m_j)p(\boldsymbol{\Theta}_j^{(i)}|m_j)p(m_j)}.$$

4.3 A simulation exercise for the non-stationary SLM

In this section we will present the result of a simulation study aiming at showing the performances of our proposed model in detecting the correct number of groups and the sensitivity to different parameters values. As a first step, we perform a simulation study with only two groups. We sample a spatial dataset of dimension $n = 200$ based on model (1). The observations are distributed on a square lattice of dimension 10 obeying to a Complete Spatial Randomness pattern (Diggle; 1983) with coordinates generated by two independent uniform processes. The $\boldsymbol{W}$ matrix is defined using a k-nearest neighbour criterion with four neighbours. Such coordinates are kept as fixed in all the simulation runs. In order to explore the effects of the groups composition, we considered three different Scenarios (see Figure 1). In Scenario I all the observations belong to the same group. In Scenario II points are divided into two groups with an horizontal boundary in the middle of the lattice. Finally, Scenario III proposes a more complex structure where points belonging to the same groups are separated.

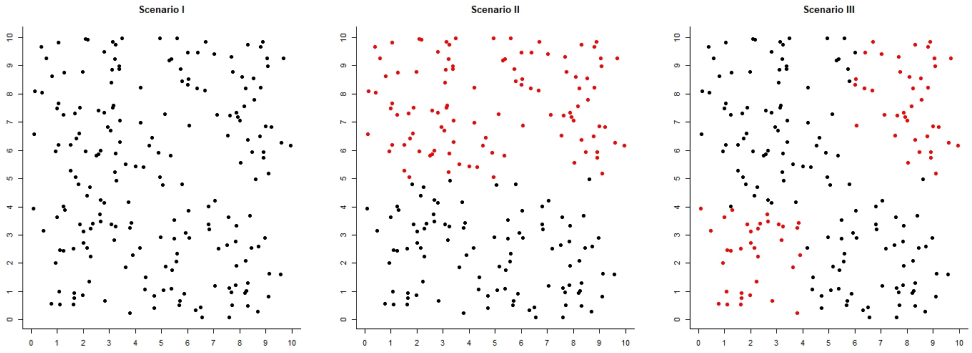

Figure 4.1: Representation of the three simulation Scenarios. The two groups of observations are coloured differently: Group 1 (red), Group 2 (black).

Consistently with this two-groups specification, the latent process $\boldsymbol{u}$ can assume the values 0 or 1, based on the group membership.

Then, we sample the covariate $\boldsymbol{X}$ from a standard Gaussian distribution, which is also considered fixed in all simulation runs. Once all the components have been defined, the dataset is sampled as:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{I}),$$
$$\boldsymbol{Y} = \boldsymbol{M}(\boldsymbol{B}\boldsymbol{X} + \boldsymbol{\varepsilon}).$$

Table 1 summarizes the combinations of parameters considered for each Scenario.

	Scenario I	Scenario II				Scenario III			
		A	B	C	D	A	B	C	D
γ_0	-	2	2	1	1	2	2	1	1
γ_1	2	2	2	3	3	2	2	3	3
ψ_0	-	0.2	0.1	0.2	0.5	0.2	0.1	0.2	0.5
ψ_1	0.75	0.7	0.9	0.7	0.5	0.7	0.9	0.7	0.5

Table 4.1: Combinations of parameters considered in the three simulation Scenarios.

Our aim is to investigate a wide casuistry of plausible situations. For this reason, we consider those Scenarios where only the spatial correlation (II A, II B, III A and III B) or where the linear dependence (II D and III D) changes and where both change their values (II D and III D). The situation with all observations within the same group is the one included in Scenario I.

The goodness of the estimations is evaluated through the Median Absolute Error (MAE), which is the median of the absolute difference between the real value and the obtained estimation. When considering the precision of the classification, we apply the Adjusted Rand Index (ARI) (Santos and Embrechts; 2009). Table 2 shows the MAE for all the parameters and the ARI for each simulations Scenario. Our results refer to the average computed on 100 Monte Carlo experiments. In each experiment, the Markov chain contains 50,000 samples, including a burn-in of 10,000 samples.

	Scenario I	Scenario II				Scenario III			
		A	B	C	D	A	B	C	D
γ_0	0.176	0.228	0.085	0.055	0.985	0.148	0.076	0.062	0.419
γ_1	0.229	0.168	0.067	0.093	1.037	0.121	0.081	0.118	0.416
ψ_0	0.381	0.155	0.044	0.061	0.058	0.073	0.038	0.061	0.052
ψ_1	0.097	0.026	0.013	0.017	0.050	0.032	0.016	0.019	0.055
ARI	0.022	0.303	0.671	0.870	0.789	0.214	0.478	0.851	0.754

Table 4.2: Median Absolute Error and Adjusted Rand Index: average values computed on 100 Monte Carlo experiments.

First of all, we can observe that a more complex topology, such as the one presented in Scenario III, overall entails more challenges in groups detection. Indeed, for the same values

of the parameters, ARI in Scenario III is always slightly lower, even though comparable. From Table 2, we can also deduce that NS-SLM seems to struggle to identify different groups when only spatial correlation changes (II A, II B and IIIA. IIIB). As expected, the difficulties are higher the more the values of ψ_0 and ψ_1 are closer. It seems that for similar parameters values, there is no statistical differences in their estimations, thus the model is not able to really distinguish the two groups. It is also interesting to observe how the difference among ϕ in distinct groups, which is required in order to properly identify the latent structure is quite high, especially if compared to limited range of ψ.

The information provided by the change of γ_0 and γ_1 results to be more informative for the groups detection, also when ψ_0 and ψ_1 are equal. While in the cases where both parameters change, the classification, hence the estimations, are more precise.

In addition, in order to emphasise the improvement in the estimations in a framework with non-stationarity of the NS-SLM, we compare our results with a standard SLM estimations. Following the notation introduced in Section 2.3, we refer to NS-SLM as m_2, due to the presence of 2 distinct groups, and to the standard SLM as m_1, which considers all observations in one single group. We perform model comparison as suggested in Equation 3, by adopting a uniform prior on the model space ($p(m_1) = p(m_2) = 0.5$). Table 3 displays the mean of the posterior probability $p(m_2|2, \boldsymbol{Y})$ and $p(m_1|1, \boldsymbol{Y})$.

	Scenario I	Scenario II				Scenario III			
		A	B	C	D	A	B	C	D
$p(m_1\|1, \boldsymbol{Y})$	0.95	0.95	0.35	0.02	0.05	0.92	0.51	0.06	0.26
$p(m_2\|2, \boldsymbol{Y})$	0.05	0.05	0.65	0.98	0.95	0.08	0.49	0.94	0.73

Table 4.3: Model comparison: mean of the posterior probability of Spatial Lag Model (m_1) and Non-Stationary Spatial Lag Model (m_2).

Our previous interpretation of the results is reinforced by the comparison with the SLM. In fact, we see that SLM is preferable not only in Scenario I, but also in the case IIA and IIIA. Therefore, in these Scenarios, the estimated optimal number of clusters is equal to be one. As we argued before, the group structure is not sensible to small spatial correlation parameter changes. In these cases, there is no statistical evidence to prefer a more complex model to the classical one. This issue is clearly overcome in the other Scenarios where there is a clear evidence of the presence of a group structure which partitions the points into two groups.

It is also interesting to show how mis-classified points are mostly concentrated near to the boundary of the true groups (see Figure 2 and 3).

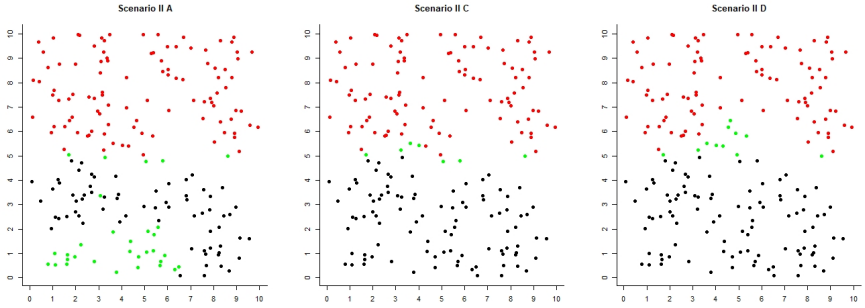

Figure 4.2: Mis-classified observations (green points) in different Scenarios. From left to right Scenario II A,C and D.

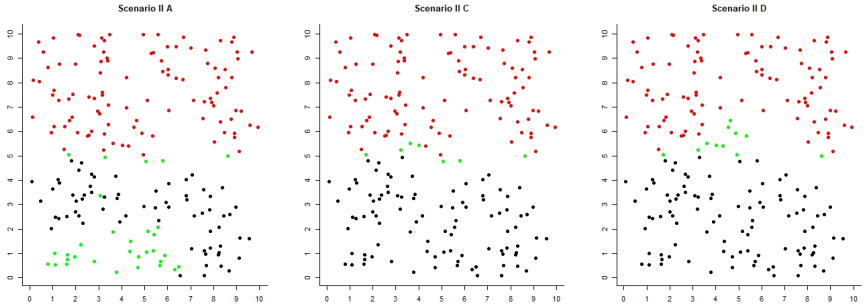

Figure 4.3: Mis-classified observations (green points) in different Scenarios. From left to right Scenario III A,C and D.

To complete our simulation exercise, we test the performance of the NS-SLM in the context of a higher number of groups. In particular, we considered an experiment where points are grouped in three distinct clusters (see in Figure 4). The combination of parameters that generate the last experiment is reported in Table 4.

	γ_0	γ_1	γ_2	ψ_0	ψ_1	ψ_2
real value	7	1	3	0	0.4	0.9
MAE	0.347	1.28	0.140	0.013	0.081	0.023
ARI	0.71					

Table 4.4: Values of the parameters, Adjust Rand Index and Median Absolute Error of the estimations in a three clusters case (average over 100 Monte Carlo experiments).

In this second experiment, our model is able to successfully detect the three different groups, even if the estimation error (Table 4) are higher than those observed in the two clusters case.

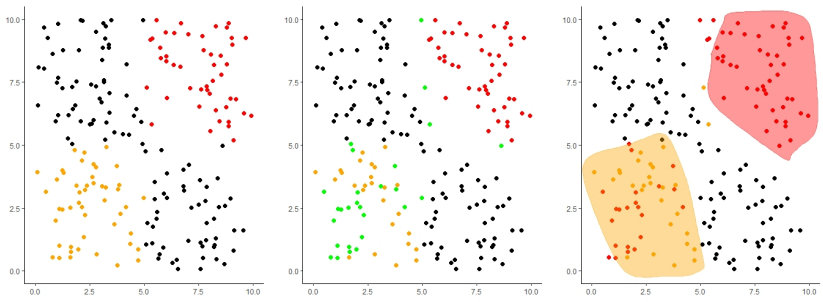

Figure 4.4: Three cluster case: from the left to the right I) spatial distribution of the cluster, II) Mis-classified observations (green points), III) Groups identified groups by Non-Stationary Spatial Lag Model.

Figure 4 also illustrates the mis-classified observations and the border of the groups identified by NS-SLM. Intuitively, the identification of hidden clusters among the observations leads to the definition of the borders where the parameters of the model change (Figure 4). A simply method to identify these borders is to compute the convex hull which contains all the points assigned to the same cluster.

4.4 A case study: Hedonic house price distribution in Boston

In this section, we describe the results obtained applying NS-SLM to the well known Boston dataset collected by Harrison and Rubinfeld (1978) and used in many spatial statistical studies. The dataset contains information about 506 house prices in Boston together with several explanatory variables which describe the house characteristics (age of the buildings and number

of rooms), the topology of the area (distance from the city center, accessibility to major roads, retail business area proportion) the quality of the air and the social composition (proportion of black population and proportion of lower status population). In our study we aim to identify groups of houses which displays a different dependence structure between the house price and the social features of a neighbourhood. More specifically we considered the SLM model as in Equation (1) and where the house price is explained by three predictors, namely: the distance from the city center, the crime rate and the black ratio. The results reported here refers to an MCMC procedure with a 150,000 sample chain, including the burn-in of 75,000 and a thin-in of 15 samples. Convergence analysis has been carried out for all the parameters, and convergence was achieved in all cases. In particular, for the latent parameters, which are discrete variables, we follow the procedure proposed by Deonovic and Smith (2017).

The two groups obtained in our analysis are mapped in Figure 4.5 and Table which 4.5 reports the expected values of the parameters and the standard deviations for both clusters.

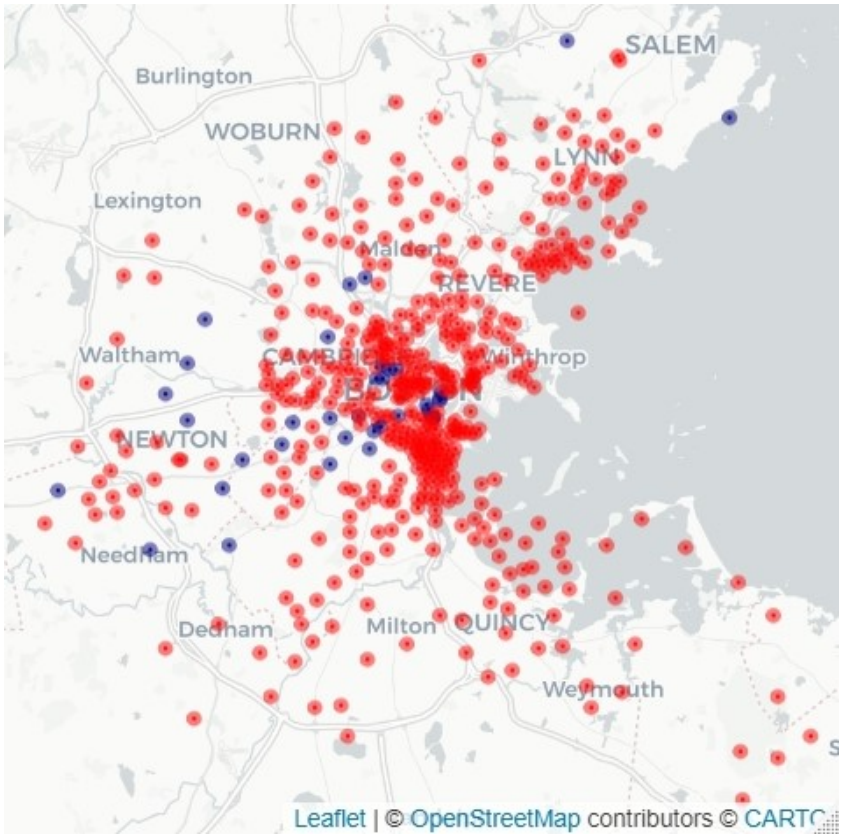

Figure 4.5: Map of the groups obtains with Non-Stationary Spatial Lag Model using the Boston dataset.

Although the standard deviations in the first cluster are quite high due to the small number of units included (35 statistical units over a total of 506 data points), there is a relevant difference in the estimations of the coefficients among the two groups. To better interpret these differences, we also refer to Table 4.6, which summarizes the distributions of the variables considered within the groups. We can observe that NL-SLM automatically divides the most expensive houses from the others. Indeed, the house price is extremely high in the first group. The spatial distribution of the houses, described by the distance from the city centre is similar in both groups, as well as the ethnic composition. The second main difference is the crime rate, which is lower in the first group.

	intercept	distance	crime	black	ρ	σ^2
Cluster 1	0.381	0.287	**0.119**	**−1.223**	**0.117**	15.205
	(8.127)	(0.343)	(0.025)	(0.529)	(0.106)	(5.310)
Cluster 2	**3.901**	**−0.104**	**0.007**	**0.165**	**0.645**	13.014
	(0.873)	(0.023)	(0.002)	(0.101)	(0.027)	(1.048)

Table 4.5: Expected values of the estimated coefficients. The standard deviations of the parameters are reported in brackets. Bold numbers correspond to those significantly different from zero.

As expected, the most expensive houses are located in safer neighbourhoods where the ethnic composition also appears to be relevant. It is worth to observe the difference between the spatial correlation parameter ψ in the two cases. In fact, in the second cluster the spatial correlation assumes a significant positive value, which denotes a greater effect of the perception of the surroundings. In contrast, in the first group it appears underdeveloped. Referring to the spatial distribution of the statistical units, this result does not have to surprise because the houses belonging to this group are less closer to each other.

	Cluster 1					Cluster 2				
	0.1	0.25	0.50	0.75	0.90	0.1	0.25	0.50	0.75	0.90
house price	39.74	43.30	48.50	50.00	50.00	12.70	16.45	20.60	24.10	30.10
distance	1.342	2.006	2.894	5.118	6.310	1.648	2.105	3.216	5.215	6.819
crime	0.017	0.046	0.520	1.491	6.191	0.042	0.083	0.245	3.736	11.108
black ratio	367.41	377.29	387.31	392.71	395.46	272.21	374.71	391.98	396.90	396.90

Table 4.6: Quantiles of the dependent variable and predictors included in the model.

4.5 Conclusions

In this Chapter, we propose the specification of a Spatial Error Model with heterogeneous coefficients. We assume the existence of a finite number of distinct groups among the observations with different groups corresponding to different parameters values. Heterogeneity is incorporated in the model through a discrete latent process evolving as an Hidden Markov Random Fields. In such a way, the structure of the groups is able to account for the spatial dependence in the data. Our Monte Carlo simulations shows that our proposed model can correctly identify the heterogeneous structure of the data. In addition, in the case of actual heterogeneity, our non-

stationary specification outperforms the classical stationary model. As expected, our model is sensitive to differences in the values of the parameters in the group. In order to properly classify the observations, small differences in the regressors coefficients are enough, while, for the spatial dependence coefficients greater differences are necessary to distinguish the group membership.

Appendix A

Update of $\boldsymbol{\gamma}$

The parameters γ_k are updated with a Gibbs sampler. Let us define define $\tilde{\boldsymbol{Y}} = \boldsymbol{M}^{-1}\boldsymbol{Y}$ and divide the data with respect to the value of the latent process, hence we obtain $\tilde{\boldsymbol{Y}}_k, \boldsymbol{X}_k$. The full-conditional distributions are computed as:

$$
\begin{aligned}
p(\boldsymbol{\gamma}|\boldsymbol{Y},\boldsymbol{u},\boldsymbol{\psi},\boldsymbol{\nu}) &\propto L(\boldsymbol{Y}|\boldsymbol{u},\boldsymbol{\psi},\boldsymbol{\nu},\boldsymbol{\gamma})\prod_{k=1}^{K}\pi(\gamma_k) \\
&\propto \exp\left\{-\frac{1}{2}[\tilde{\boldsymbol{Y}}-\boldsymbol{B}\boldsymbol{X}]^T\boldsymbol{S}^{-1}[\tilde{\boldsymbol{Y}}-\boldsymbol{B}\boldsymbol{X}]\right\}\prod_{k=1}^{K}\exp\left\{-\frac{1}{2\sigma_\gamma^2}\gamma_k^2\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(-2\sum_{i=1}^{n}\frac{\tilde{Y}_i X_i\beta_i}{\sigma_i^2}+\sum_{i=1}^{n}\frac{X_i^2\beta_i^2}{\sigma^2}\right)\right\}\prod_{k=1}^{K}\exp\left\{-\frac{1}{2\sigma_\gamma^2}\gamma_k^2\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(-2\sum_{k=1}^{K}\frac{\gamma_k}{\nu_k}\sum_{i=1}^{n_k}\tilde{Y}_{k,i}X_{k,i}+\sum_{k=1}^{K}\frac{\gamma_k^2}{\nu_k}\sum_{i=1}^{n_k}X_{k,i}^2\right)\right\}\times \\
&\quad\times\prod_{k=1}^{K}\exp\left\{-\frac{1}{2\sigma_\gamma^2}\gamma_k^2\right\} \\
&\propto \prod_{k=1}^{K}\exp\left\{-\frac{\gamma_k^2\boldsymbol{X}_k^T\boldsymbol{X}_k}{2\nu_k}+\frac{2\gamma_k\boldsymbol{X}_k^T\tilde{\boldsymbol{Y}}_k}{2\nu_k}\right\}\exp\left\{-\frac{1}{2\sigma_\gamma^2}\gamma_k^2\right\}.
\end{aligned}
$$

Since the distribution is factorisable, γ_k can be separately updated

$$
\begin{aligned}
p(\gamma_k|\boldsymbol{Y},\boldsymbol{u},\boldsymbol{\psi},\boldsymbol{\nu}) &\propto L(\boldsymbol{Y}_k|u=k,\psi_k,\nu_k,\gamma_k)\times\pi(\gamma_k) \\
&\propto \exp\left\{-\frac{1}{2\nu_k}[\tilde{\boldsymbol{Y}}_k-\boldsymbol{X}_k\gamma_k]^T[\tilde{\boldsymbol{Y}}_k-\boldsymbol{X}_k\gamma_k]\right\}\exp\left\{-\frac{1}{2\sigma_\gamma^2}\gamma_k^2\right\} \\
&\propto \exp\left\{-\frac{\gamma_k^2}{2\sigma_\gamma^2}-\frac{\gamma_k^2\boldsymbol{X}_k^T\boldsymbol{X}_k}{2\nu_k}+\frac{2\gamma_k\boldsymbol{X}_k^T\tilde{\boldsymbol{Y}}_k}{2\nu_k}\right\}.
\end{aligned}
$$

It follows that

$$
\gamma_k|\boldsymbol{Y},u=k,\psi_k,\nu_k \sim \mathcal{N}\left(\left(\frac{1}{\sigma_\gamma^2}+\frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{\nu_k}\right)^{-1}\frac{\boldsymbol{X}_k^T\tilde{\boldsymbol{Y}}_k}{\nu_k},\left(\frac{1}{\sigma_\gamma^2}+\frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{\nu_k}\right)^{-1}\right),
$$

Update of $\boldsymbol{u}$

The full conditional distribution of u_i is

$$\begin{aligned} p(u_i|\boldsymbol{Y}, \boldsymbol{u}_{-i}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{\nu}) &= L(Y_i|\boldsymbol{Y}_{-i}, u_i, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\nu}) L(\boldsymbol{Y}_{-i}|\boldsymbol{u}_{-i}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\nu}) p(u_i|\boldsymbol{u}_{-i}, \boldsymbol{\theta}) p(\boldsymbol{u}_{-i}|\boldsymbol{\theta}) \\ &\propto L(Y_i|\boldsymbol{Y}_{-i}, u_i, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\nu}) p(u_i|\boldsymbol{u}_{i\sim j}, \boldsymbol{\theta}). \end{aligned}$$

$$p(u_i = l|\boldsymbol{Y}, \boldsymbol{u}_{i\sim j}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{\nu}) = \frac{L(Y_i|\boldsymbol{Y}_{-i}, u_i = l, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\nu}) \cdot p_{l,i}(\boldsymbol{u}_{-i})}{\sum_k L(Y_i|\boldsymbol{Y}_{-i}, u_i = k, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\nu}) \cdot p_{k,i}(\boldsymbol{u}_{-i})} = q_{k,i}(\boldsymbol{u}_{-i})$$

Hence

$$u_i|\boldsymbol{Y}, \boldsymbol{u}_{-i}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{\nu} \sim Multinom(q_{1,i}(\boldsymbol{u}_{-i}), ..., q_{K,i}(\boldsymbol{u}_{-i})).$$

Update of $\boldsymbol{\nu}$

The parameters $\boldsymbol{\nu}$ are also updated with a Gibbs step:

$$\begin{aligned} p(\boldsymbol{\nu}|\boldsymbol{Y}, \boldsymbol{u}, \boldsymbol{\psi}, \boldsymbol{\gamma}) &\propto L(Y|\sigma^2, \boldsymbol{u}, \boldsymbol{\psi}, \boldsymbol{\gamma}) \prod_{k=1}^{K} \pi(\nu_k) \\ &\propto \prod_{k=1}^{K} (\nu_k)^{-n_k/2} \exp\left\{-\frac{1}{2}[\tilde{\boldsymbol{Y}} - \boldsymbol{B}\boldsymbol{X}]^T \boldsymbol{S}^{-1} [\tilde{\boldsymbol{Y}} - \boldsymbol{B}\boldsymbol{X}]\right\} \times \\ &\prod_{k=1}^{K} (\nu_k)^{-(a+1)} exp\left\{-\frac{b}{\nu_k}\right\} I_{[0,\infty]}(\nu_k) \end{aligned}$$

It follows that

$$\nu_k|\boldsymbol{Y}, u = k, \psi_k, \gamma_k \sim inv\Gamma\left(a + \frac{n_k}{2}, b + \frac{[\tilde{\boldsymbol{Y}}_k - \gamma_k \boldsymbol{X}_k]^T [\tilde{\boldsymbol{Y}}_k - \gamma_k \boldsymbol{X}_k]}{2}\right).$$

Appendix B

Convergence results of the study of Hedonic house price (Section 4). For each parameter we reported the traceplot, the autocorrelation and the Geweke plot (see Figures 4.6,4.7,4.8,4.9 and 4.10).

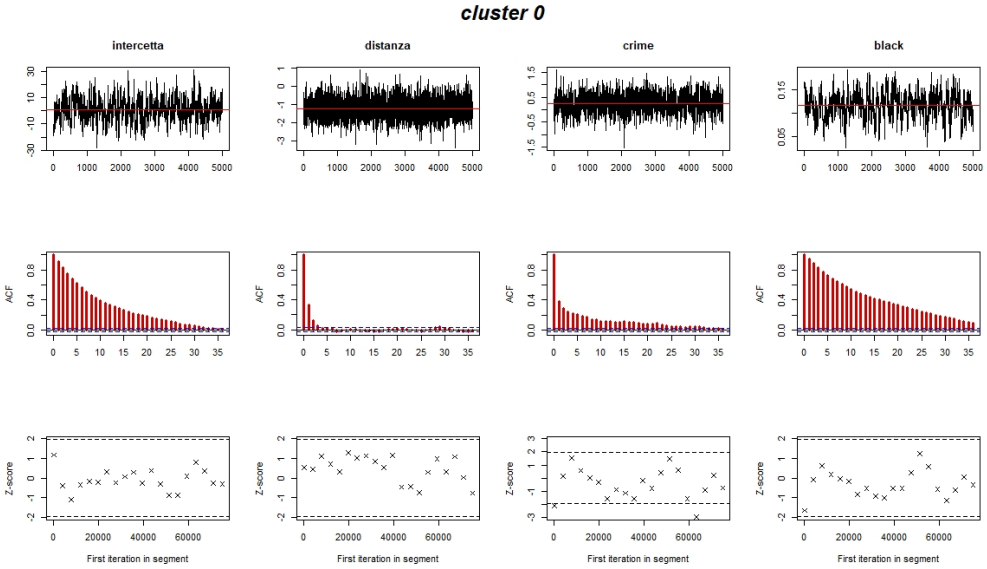

Figure 4.6: Traceplot, autocorrelation and the Geweke plot for the regressor parameters of the first cluster.

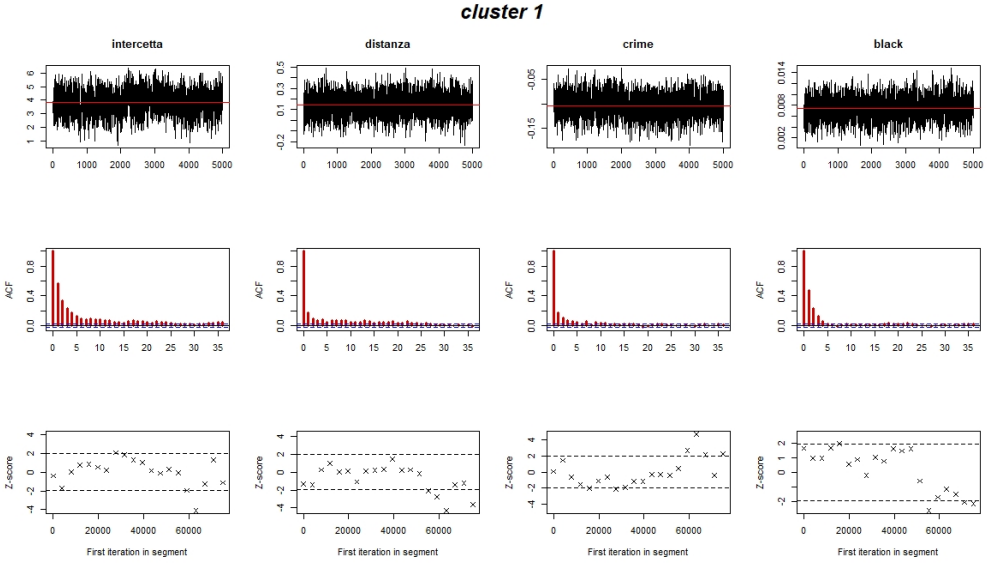

Figure 4.7: Traceplot, autocorrelation and the Geweke plot for the regressor parameters of the second cluster.

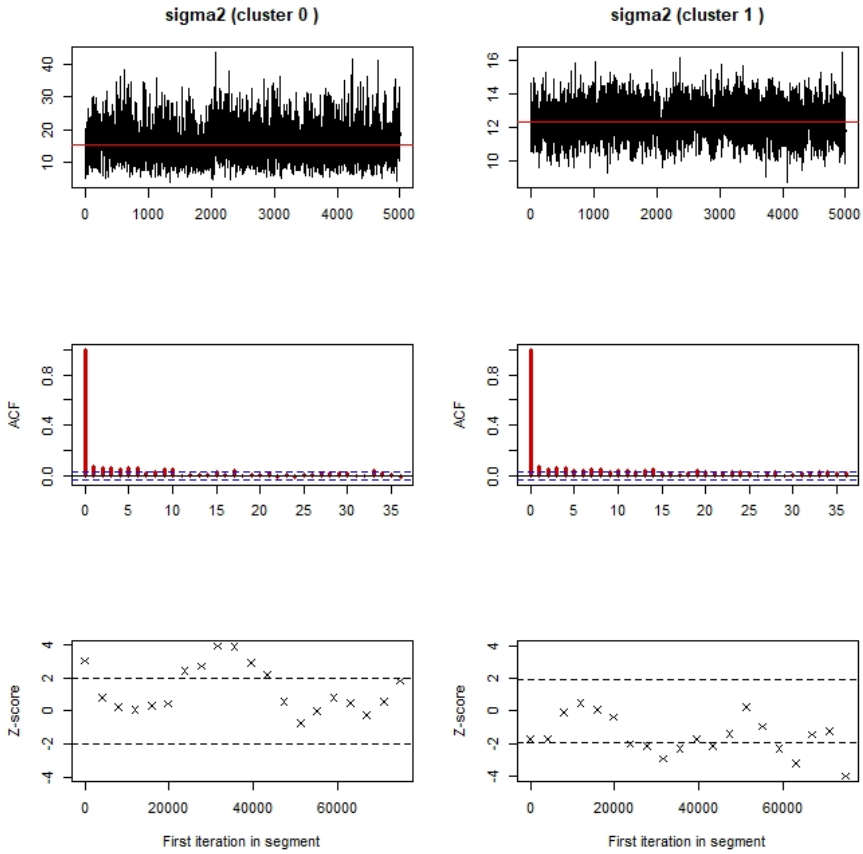

Figure 4.8: Traceplot, autocorrelation and the Geweke plot for the variance of the errors term.

Based on the results reported, we can state that convergence is reached for all the parameters.

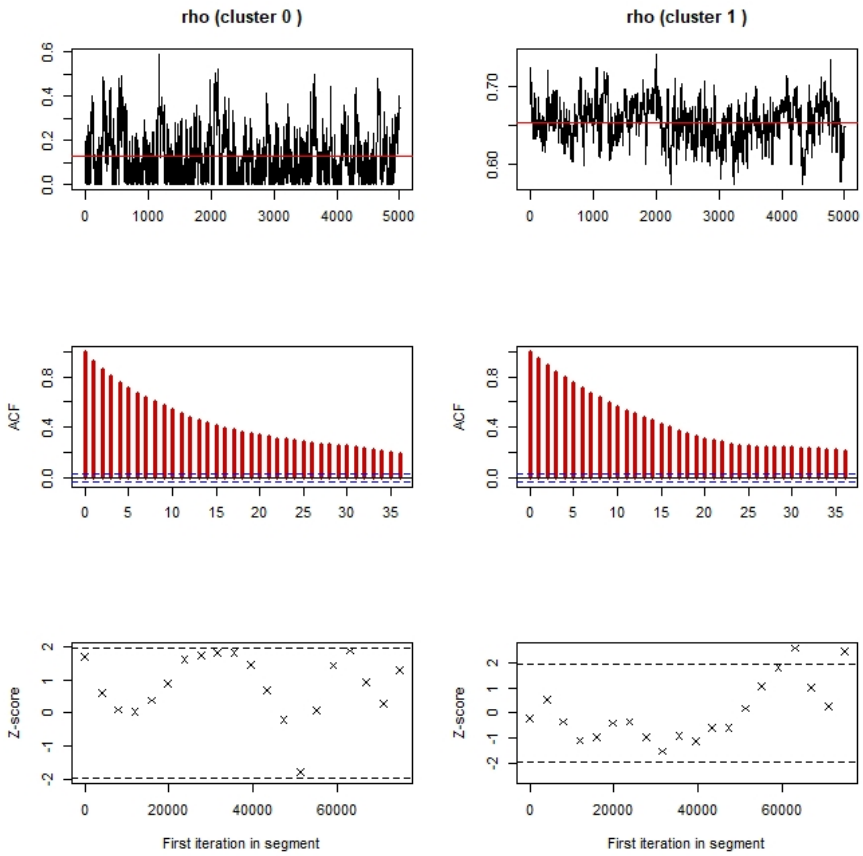

Figure 4.9: Traceplot, autocorrelation and the Geweke plot for spatial regression parameters.

The convergence for the latent process is checked using the *Mamba* package in Julia (Bezanson et al.; 2017) proposed by Deonovic and Smith (2017). The obtained p-values (all lower than 0.05) suggest the convergence of the discrete variables for the all observations.

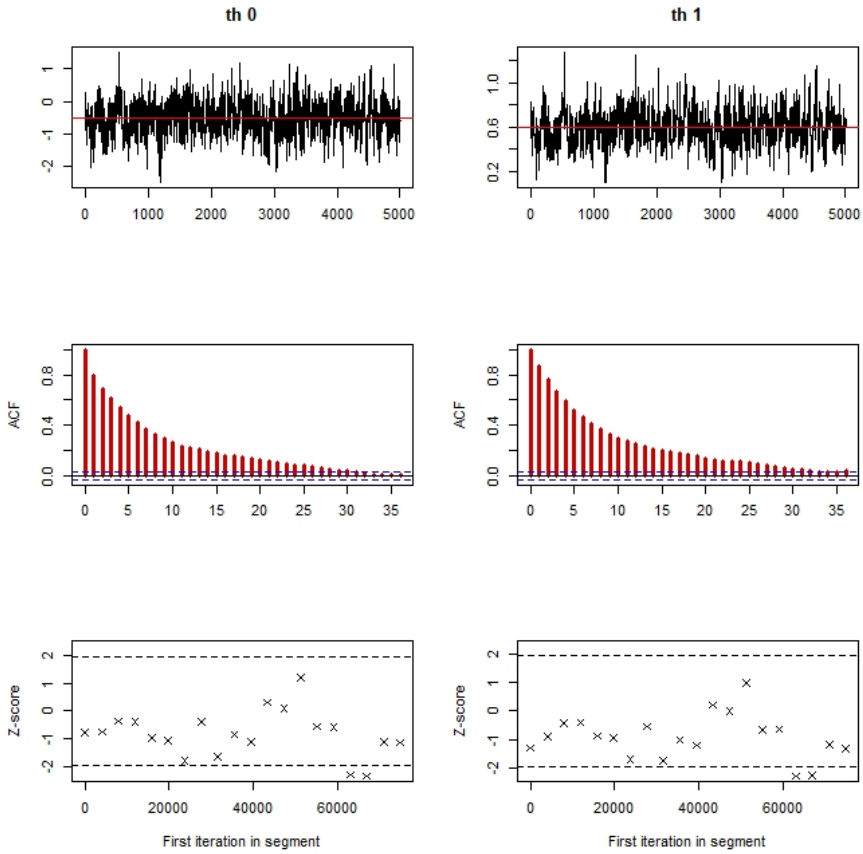

Figure 4.10: Traceplot, autocorrelation and the Geweke plot for $\boldsymbol{\theta}$.

Since in some case we notice that the autocorrelation in the chain decreases slowly, we also check the correlation in the parameters chains (Table 4.7) but we do not detect evidences for a strong correlation a posteriori between the parameters.

	σ_1^2	σ_2^2	ψ_1	ψ_2	τ_0	τ_1	int_1	$dist_1$	$crime_1$	b_1	int_2	$dist_2$	$crime_2$	b_2
σ_1^2	1.00	-0.29	0.21	-0.09	0.03	-0.07	0.06	-0.21	-0.05	-0.13	0.07	-0.11	-0.01	0.00
σ_2^2	-0.29	1.00	-0.23	-0.08	0.01	0.12	-0.09	0.18	0.07	0.16	0.02	0.18	-0.02	0.01
ψ_1	0.21	-0.23	1.00	0.00	-0.14	-0.22	0.27	0.04	-0.36	-0.64	-0.02	-0.12	0.01	0.03
ψ_2	-0.09	-0.08	0.00	1.00	-0.02	0.01	-0.05	-0.04	-0.04	0.06	-0.43	-0.28	0.26	-0.20
τ_0	0.03	0.01	-0.14	-0.02	1.00	0.91	-0.13	-0.11	0.04	0.18	0.05	-0.02	-0.01	-0.02
τ_1	-0.07	0.12	-0.22	0.01	0.91	1.00	-0.18	-0.04	0.07	0.25	0.03	0.03	0.00	-0.04
int_1	0.06	-0.09	0.27	-0.05	-0.13	-0.18	1.00	-0.15	-0.54	-0.89	-0.01	-0.05	-0.02	0.06
$dist_1$	-0.21	0.18	0.04	-0.04	-0.11	-0.04	-0.15	1.00	0.44	-0.08	-0.05	0.27	0.04	-0.02
$crime_1$	-0.05	0.07	-0.36	-0.04	0.04	0.07	-0.54	0.44	1.00	0.47	0.00	0.13	0.02	-0.03
b_1	-0.13	0.16	-0.64	0.06	0.18	0.25	-0.89	-0.08	0.47	1.00	0.02	0.05	0.00	-0.06
int_2	0.07	0.02	-0.02	-0.43	0.05	0.03	-0.01	-0.05	0.00	0.02	1.00	-0.16	-0.55	-0.66
$dist_2$	-0.11	0.18	-0.12	-0.28	-0.02	0.03	-0.05	0.27	0.13	0.05	-0.16	1.00	0.22	-0.10
$crime_2$	-0.01	-0.02	0.01	0.26	-0.01	0.00	-0.02	0.04	0.02	0.00	-0.55	0.22	1.00	0.22
b_2	0.00	0.01	0.03	-0.20	-0.02	-0.04	0.06	-0.02	-0.03	-0.06	-0.66	-0.10	0.22	1.00

Table 4.7: Posterior correlation matrix between all parameters.

Figures 4.11 and 4.12 illustrate the Credibility Intervals (level 0.89) which allows to determine if there are statistical differences between the parameters estimations among the two groups. As state in Section 4, the intervals for the parameters of the first clusters are wider due to the lower number of observations identified in that group. However, we can detect a significant different in the estimations of the spatial correlation parameter and the coefficients of the distance and the concentration of black people.

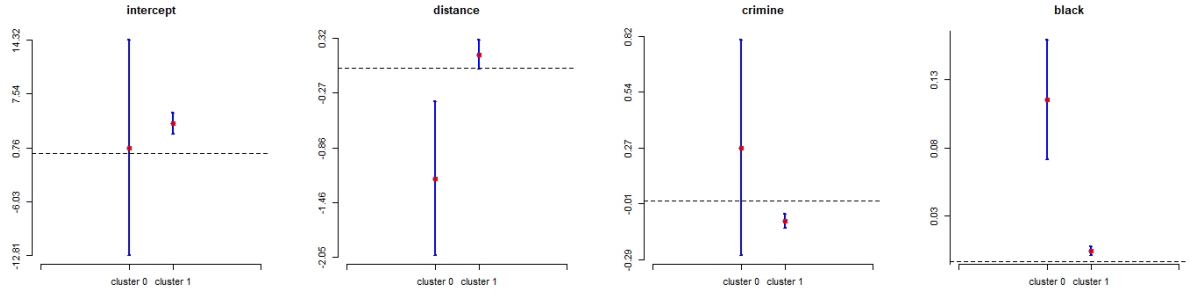

Figure 4.11: Credibility Intervals (level 0.89) for the regressors coefficients.

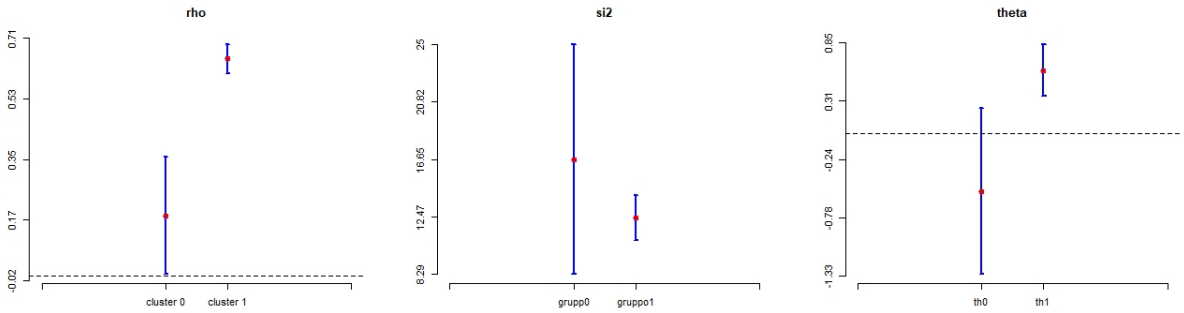

Figure 4.12: Credibility Intervals (level 0.89) for $\boldsymbol{\sigma}^2$,$\boldsymbol{\rho}$ and $\boldsymbol{\theta}$.

Bibliography

Arbia, G. (1990). On second-order non-stationarity in two dimensional lattice processes, *Computational Statistics & Data Analysis* **9**: 147 – 160.

Arbia, G. (2006). *Spatial Econometrics: Statistical foundations and applications to regional convergence*, Springer-Verlag.

Arbia, G. (2014). *A primer for Spatial Econometrics*, Palgrave Macmillan.

Arbia, G., Bee, M., Espa, G. and Santi, F. (2018). Fitting spatial regressions to large datasets using unilateral approximations, *Communications in Statistics - Theory and Methods* **47**: 222–238.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC.

Bartolucci, F., Farcomeni, A. and Pennoni, F. (2012). *Latent Markov Model for Longitudinal Data*, Chapman and Hall/CRC.

Bertarelli, G., Ranalli, M., Bartolucci, F., DAl, M. and Solari, F. (2018). Small area estimation for unemployment using latent markov models, *Survey Methodology* **44**.

Besag, J. (1986). On the statistical analysis of dirty pictures, *Journal of the Royal Statistical Society. Series B* **48**: 259–302.

Bezanson, J., Edelman, A., Karpinski, S. and Shah, V. B. (2017). Julia: A fresh approach to numerical computing, *SIAM review* **59**(1): 65–98.
URL: *https://doi.org/10.1137/141000671*

Brunsdon, C., Fotheringham, S. and Charlton, M. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity, *Geographical Analysis* **28**: 281–298.

Brunsdon, C., Fotheringham, S. and Charlton, M. (1998). Spatial nonstationarity and autoregressive models, *Environment and Planning A* **30**: 957–973.

Brunsdon, C., Fotheringham, S. and Charlton, M. (2007). Geographically weighted discriminant analysis, *Geographical Analysis* **39**: 376–396.

Charras-Garrido, M., Abrial, D., Goër, J. D., Dachian, S. and Peyrard, N. (2011). Classification method for disease risk mapping based on discrete hidden Markov random fields, *Biostatistics* **13**: 241–255.

Christakos, G. (1992). *Random field models in earth sciences*, Academic Press.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting, *Journal of the American Statistical Association* **83**: 596–610.

Congdon, P. (2006). Bayesian model choice based on monte carlo estimates of posterior model probabilities, *Computational Statistics & Data Analysis* **50**: 346–357.

Deng, M. (2008). An anisotropic model for spatial processes, *Geographical Analysis* **40**: 26–51.

Deonovic, B. E. and Smith, B. J. (2017). Convergence diagnostics for mcmc draws of a categorical variable, *arXiv: Methodology* .

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns.*, London: Academic Press.

Fisher, R. A. (1959). *Statistical methods and statistical inference*, Statistical methods and statistical inference.

Fotheringham, S., Brunsdon, C. and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Wiley.

Fotheringham, S., Charlton, M. and Brunsdon, C. (1998). Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis, *Environment and Planning A* **30**: 1905–1927.

Gelfand, A. E., Kim, H.-J., Sirmans, C. F. and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes, *Journal of the American Statistical Association* **98**: 387–396.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper), *Bayesian Anal.* **1**: 515–534.

Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan statistics*, Springer series in statistics.

Green, P. J. and Richardson, S. (2002). Hidden markov models and disease mapping, *Journal of the American Statistical Association* **97**: 1055–1070.

Harrison, D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air, *Journal of Environmental Economics and Management* **5**: 81–102.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling, *Statist. Sci.* **20**: 50–67.

Lee, D. (2013). CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors, *Journal of Statistical Software* **55**: 241–255.

LeSage, J. P. and Pace, R. K. (2010). *Spatial Econometric Models*, Springer Berlin Heidelberg, pp. 355–376.

Li, F. and Sang, H. (2019). Spatial homogeneity pursuit of regression coefficients for large datasets, *Journal of the American Statistical Association* **114**: 1050–1062.

Lingren, F. and Rue, H. (2015). Bayesian spatial modelling with r-inla, *Journal of Statistical Software* **63**.

Lu, H. and Carlin, B. P. (2005). Bayesian areal wombling for geographical boundary analysis, *Geographical Analysis* **37**: 265–285.

McMillen, D. P. and McDonald, J. F. (2004). Locally weighted maximum likelihood estimation: Monte carlo evidence and an application.

Okabayashi, S., Johnson, L. and Geyer, C. J. (2011). Extending pseudo-likelihood for potts models, *Statistica Sinica* **21**: 331–347.

Ott, J. and Hoh, J. (2012). Scan statistics in human gene mapping, *Am J Hum Genet* p. 91(5): 970.

Páez, A., Long, F. and Farber, S. (2008). Moving window approaches for hedonic price estimation: An empirical comparison of modelling techniques, *Urban Studies* **45**: 1565–1581.

Santi, F., Arbia, G., Bee, M. and Espa, G. (2017). A frequency domain test for isotropy in spatial data models, *Spatial Statistics* **21**: 262 – 278.

Santos, J. M. and Embrechts, M. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification, *in* C. Alippi, M. Polycarpou, C. Panayiotou and G. Ellinas (eds), *Artificial Neural Networks – ICANN 2009*, Springer Berlin Heidelberg, pp. 175–184.

Spezia, L., Friel, N. and Gimona, A. (2018). Spatial hidden markov models and species distributions, *Journal of Applied Statistics* **45**: 1595–1615.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*, CRC Press.