POLITECNICO DI MILANO

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

Corso di Laurea Magistrale in Ingegneria Matematica

TESI DI LAUREA MAGISTRALE

POLITECNICO
MILANO 1863

A Bayesian analysis of population density over time: how spatial correlation matters

Relatore: **Dott.ssa Ilenia Epifani**

Correlatore: **Prof.ssa Alessandra Guglielmi**

Candidato:
Chiara Ghiringhelli
Matr. 823392

Anno Accademico 2015 - 2016

Abstract

In this work we propose a dynamic Bayesian approach to modeling the population's density; predictors of different nature are used, e.g. economics and geographic indices. The model is applied to the evaluation of the location of population in the state of Massachusetts over a period of 50 years, from 1970 to 2010. The aim of this work is to introduce into the analysis both spatial and time correlation among data. We deal with AutoRegressive models, that provide the most common way to explore time dependence. In order to explore spatial correlation, we propose two different generalized regression mixed models: one with spatial independent random effects and one that includes spatial random effects evolving as a Conditionally AutoRegressive model (CAR). Both are compared with a baseline linear model. For the CAR model, we derive the analytical expression of the full conditional distributions necessary to build a MCMC algorithm efficiently coded in Julia language, and to sample from a posterior distribution. The implementation of the other two models were made in Stan.

Keywords: Areal data models; AutoRegressive model; Bayesian analysis; CAR; MCMC algorithm; Spatial random effects.

Sommario

In questo lavoro proponiamo un approccio dinamico bayesiano per modelizzare la densità di popolazione; vengono utilizzati predittori di diversa natura, per esempio indici economici e geografici. Il modello è applicato all'evoluzione dello stanziamento della popolazione nello stato del Massachusetts lungo un periodo di 50 anni, dal 1970 al 2010. Lo scopo del lavoro è di introdurre nell'analisi una correlazione spaziale e una temporale tra i dati. Utilizziamo un modello autoregressivo, che è uno degli strumenti fondamentali per esplorare la correlazione temporale. Per quanto riguarda la correlazione spaziale, proponiamo due modelli di regressione mista generalizzati: uno con effetti spaziali casuali independenti e uno che include effetti spaziali che evolvono come un modello Condizionatamente Autoregressivo (CAR). Entrambi sono confrontati con un modello di riferimento lineare. Per il modello CAR , calcoliamo l'espressione analitica delle distribuzioni full conditional necessarie per implementare un algoritmo MCMC efficiente e campionare dalla distribuzione a posteriori. Abbiamo implementato l'algoritmo nel linguaggio di programmazione Julia. Mentre l'implementazionde degli altri due modelli è stata effettuata in Stan.

Keywords:algoritmi MCMC; Analisi Bayesiana; Effetti spaziali casuali; Dati Spaziali; modelli AutoRegressivi; modelli CAR.

Contents

List of Figures

List of Tables

Introduction

This work investigates a Bayesian approach to the study of settlement of a population in the territory. The Bayesian approach to spatial problems involves several advantages: it makes the computation feasable, otherwise without using full conditional distributions it would require too much time for computation, especially for large dataset. In addition, the use of hierarchical levels allows to model dependence and correlation among data by defining *ad hoc* prior distributions. Spatial models are really flexible and allow a lot of different combinations, for this reason these kinds of models are applicable in a lot of different topics. In particular they frequently arise in economic, geographical and epidemiologic studies. In these contexts it is important the location of the elements and one wants to find a spatial pattern among the data. Therefore one looks for common features between an element and its neighbourings. These type of data are called *areal data*; datasets are usually available with a huge number of units, especially in an extended geographic area, that makes the computation hard to do.

We study the population distribution in the state of Massachusetts. We take a picture of the state in a period of 50 years, from 1970 to 2010. During the whole period there were no drammatical events, i.e. neither wars, nor revolutions, nor earthquakes, hence we try to describe the movement of population in normal conditions. Modeling individual choice is not immediate, because it is lead by subjective preferences, work requirements and so on; this kind of information are hard to classify. Anyway we try to determine some fundamental features (like distance from big cities, ethnic composition, natural amenities, education, house holds) that condition the population distribution. In this work, the new contribution is to explore spatial and time correlation between data simultaneously. Clearly, an individual prefers to settle down in a context that corresponds to his/her own features, in other words in a place of economic wellness, safety area and similar ethnic composition. We wonder whether the position, hence the neighbourhood, really influences individual choices, or whether some new areas with particular features arises over time, for exemple a ghetto. The second basic idea is to explore whether features of current time are correlated to information at past period. By the way, we try to determine if there is an effective "reputation effect": the individual choices are guided by the reputation of a city in the past.

Following Epifani and Nicolini (2013, 2015a, 2015b), in this thesis we apply three different dynamic Bayesian hierarchical regression lognormal models to population density at level of census tracts; these models involve the spatial correlation in different ways. A first model takes account for spatial correlation only at a county level by means of the introduction of a global county effect given by the "amenities". A second one includes independent random effects, one for each census tract. Finally, the last model is the most complex, it is a mixed generalized linear model, where the census tract random effects evolve according to a *Conditionally AutoRegressive Model* (CAR); this structure allows to include the neighbourhood's influence in the analysis. Since Besag (1974), there are in literature a huge amount of this kind of model see for example Cressie and Stern(1999) , Banerjee et Carlin (2003a) among the others.
Differently from the general theory of linear models, the regression coefficients are not a priori gaussian distributed. Instead, they have a dynamic structure ruled by an *Autoregressive Model* of order one (AR(1)), that allows to model the time dependence. The computational heaviness of the last model is due to the huge number of parameters, depending each others, whose are required to sample at each iteration: they are as many as the data. In order to overcome this problem and make the code as efficient as possible, the model has been implemented in Julia, an efficient language with fast performance.

This thesis is organized as follows: Chapter 1 presents areal data, formulation and main properties of the CAR models and a brief description of the Bayesian approach. In Chapter 2 the basic theory of time series is summarized, with a special attention to fundamental theorical results for AR models. Chapter 3 has been dedicated to the dataset of the census information in Massachusetts, with an exploring analysis of the variables. Afterwards, we set three different models to investigate spatial and time correlation and describe the calculation of the full conditionals and the sampling scheme. In Chapter 4 we present the results of the models and compare them. We also make a deeper analysis of some particular counties. Finally, in Appendix A we derive the analytic expression of the full conditionals of the model, whereas in Appendix B their implementation in Julia language and Stan is briefly described.

Chapter 1

Areal data models

Areal data are data collected for areal units: every element of the data set has a position and an associated area. Despite the idea of the existence of an influence for data among the space is quite old, the concept of spatial correlation was theorized only in the 1960s by Cliff and Ord. Since that time, spatial models have been applied in a lot of differet fields like econometrics, epidemiology, geography, statistic and so on. The concept of spatial correlation is really similar to the one of temporal correlation, developed in the 1950s by Durbin and Watson. In both cases the aim is to identify the outliers or a trend in the data, however spatial analysis studies are more complicated because we need to verify correlation in all directions, as opposed to the one way temporal direction. Areal models can be applied both in problems with areal units with an irregular shape, for example a geographic map, and in case of a regulare grid, like pixels in a photo.

In the context of areal units analysis the general inferential issue is to identify a spatial pattern. In other words one has to determine if the features of nearby areal units take similar values, while they are different from the ones of far units. If high values at one locality are associated with high values at neighbouring localities, then the spatial autocorrelation is positive. Instead if high and low values alternate between adjacent localities the spatial autocorrelation is negative. Defining a spacial pattern is not immediate because a unique definition does not exist. We can say that there is spacial dependence if the values of a variable in a mapped pattern deviate significantly from a pattern in which the values are assigned randomly (see Goodchild, 1987, Griffith, 1991).

If a spatial patter has been found, it is important to discover how much it is.
The response of the model is usually expressed by a regression on some available covariates. In order to introduce spatial correltion, our approach does not apply a spatial model directly to the data, but introduces spacial association by means of random effects; in this way we obtain a generalized linear mixed model.

1.1 Introduction of spacial correlation

One can introduce spatial correlation by a *proximity matrix* W. Let $Y_1, \dots, Y_n$ be n observation on a response y associated with $1, \dots, n$ areal units and W an $n \times n$ matrix, where each $w_{i,j}$ measures the "distance" between element i and element j. The concept of distance is really ambiguous, in fact there are lots of way to interpret if a point is near to another one. The most common method is the euclidean distance between the coordinates. Alternatively there are more general definitions, for example a binary determination where $w_{i,j} = 1$ if i and j are neighbours. Instead, if i and j are linked through infrastructures that allow to move fastly from one to the other, their distance can be defined as the time between the two places. The distance can also depend on the direction, that is the case of a regular lattice, like that in Figure 1.1.

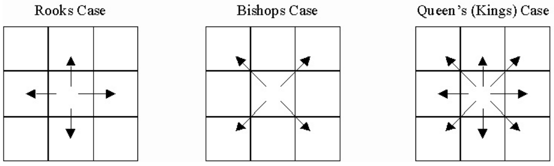

Figure 1.1: Example of different definition of distance for a regular grid.

In areal analysis with geographical elements the most common way is to set $w_{i,j} = 1$ if i and j are neighbors and 0 otherwise. Such a matrix $W = [w_{i,j}]$ is symmetric. In some applications, it can be useful to normalize W by dividing each element by the sum of its row:

$$a_{i,j} = \frac{w_{i,j}}{\sum\limits_{j=1}^{n} w_{i,j}} \quad . \tag{1.1}$$

We call A a *contingency matrix*. Unfortunately the new matrix A is not symmetric any more.

In this work we have decided to measure the distance by a contingency matrix A as defined in Equation (1.1).

1.2 Measures of spatial association

Before applying the model, in order to explore the presence of an effective spacial association, it is recommended to perform some statistical tests. All test statistics that measure the spatial autocorrelation have a common root given by the following matrix cross-product

$$\Gamma = \sum_{i,j} w_{i,j} c_{i,j} \tag{1.2}$$

where $W = [w_{i,j}]$ is a proximity matrix 0,1 and C represents a measure of the association between two elements. The general cross-product Γ is a statistic in the sense that the implied matrix is a sample of a number of possible matrices. The value of the cross-product can be compared to the range of values that might be produced if a number of maps with the same set of values were created by a complete random assignment of values to locations. There are $n!$ different possible maps that could be produced randomly if each of the original values were randomly assigned. Once Γ has been calculated, if we compute all the cross-product related to the $n!$ possible matrices, we have generated an empirical distribution for Γ. In this way we can estabilish if Γ is an outlier and so decide if there is spacial correlation among areal units. This procedure is computationally unfeasible, expecially if n is really big. It turns out to be more convenient to compute some indices like the Moran's I or Geary's C, that derive from the cross-product but can be asympotically approximated. In our Bayesian application we will use such indices only for an exploratory analysis; we do not interpret them as a frequentist test of spatial significance.

If we set $c_{i,j} = (Y_i - \overline{Y})^2 (Y_j - \overline{Y})^2$ in Equation (1.2), we obtain the $Moran's\ Index$:

$$I = \frac{n \sum_{j=1}^{n} \sum_{i=1}^{n} a_{i,j} (Y_i - \overline{Y})(Y_j - \overline{Y})}{\sum_{i \neq j} a_{i,j} \sum_{i=1}^{n} (Y_i - \overline{Y})^2} \quad .$$

Moran demostrated that if $Y_1, \dots, Y_n$ are independent and equally distributed, then I is asymptoticaly normally distributed with the following law

$$I \sim \mathcal{N}\left(-\frac{1}{n-1}, \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2}\right)$$

where $S_0 = \sum\limits_{i \neq j}^{n} a_{i,j}$, $S_1 = \frac{1}{2}\sum\limits_{i \neq j}^{n}(a_{i,j} + a_{j,i})^2$ and $S_2 = \sum\limits_{k=1}^{n}(\sum\limits_{j=1}^{n} a_{k,j} + \sum\limits_{i=1}^{n} a_{i,k})^2$.
Index I belongs to $[-1, 1]$ and we can interpretate its value as follows:

$$I = \begin{cases} I < -\frac{1}{n-1} & \text{there is evidence for negative spatial association} \\ I = -\frac{1}{n-1} & \text{there is no evidence for spatial association} \\ I > -\frac{1}{n-1} & \text{there is evidence for positive spatial association} \quad . \end{cases}$$

The expression of $Geary's\ C$ takes form

$$C = \frac{(n-1)\sum\limits_{j=1}^{n}\sum\limits_{i=1}^{n} a_{i,j}(Y_i - Y_j)^2}{2\sum\limits_{i \neq j} a_{i,j} \sum\limits_{i=1}^{n}(Y_i - \overline{Y})^2} \quad .$$

If $Y_1, \dots, Y_n$ are independent and identically distributed then C is asymptotically normal distributed with law

$$C \sim \mathcal{N}\left(1, \frac{(2S_1 + S_2)(n-1) - 4S_0^2}{2(n+1)S_0^2}\right)$$

where the above notation still holds. Usually $0 < C < 2$, only in rare cases $C > 2$. The interpretation is the following:

$$C = \begin{cases} 0 < C < 1 & \text{there is evidence for positive spatial association} \\ C = 1 & \text{there is no evidence for spatial association} \\ C > 1 & \text{there is evidence fot negative spatial association} \quad . \end{cases}$$

Moran's I is a more global measurement and sensitive to extreme values of y, whereas Geary's C is more sensitive to differences in small neighborhoods.

1.3 Calculation of the joint distribution

Given our data $Y_1, \ldots, Y_n$, we need to calculate the joint distrtibution $p(y_1, \ldots, y_n)$. Bayesian methodology has existed for a long time, but only recently it approaches to estimation of these models, making them practically feasible. The computation approach known as Markov Chain Monte Carlo (MCMC) decomposes complicated estimation problems into simpler problems that rely on the lower-dimensional conditional distributions for each parameter in the model (Gelfand and Smith, 1990).

1.3.1 Bayesian method

In this subsection we shortly describe the basic ideas of the Bayesian analysis.

The most important aspect of the Bayesian methodology is the focus on distributions for the data as well as for the parameters. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be independent and identically distributed observations from a probability distribution f, conditionally to some unknown parameters $\boldsymbol{\theta}$. The basic difference between a Bayesian and frequentist approach is that in Bayesian perspective the parameters $\boldsymbol{\theta}$ are not constant, but they are random variables. We set a prior distribution $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$; that represents the knowledge that we have on the topic before the analysis. If our knowledge from the prior experience is very poor, then this distribution should represent a vague probabilistic statement, whereas a great deal of previous experience would lead to a very narrow distribution centered on some hyperparameters gained from previous experience. Datasets are usually large and prior information will tend to play a minor role in determining the character of the posterior distribution. Formally a Bayesian model is given by:

$$X_1, \ldots, X_n | \boldsymbol{\theta} \overset{\text{iid}}{\sim} f(\boldsymbol{x}, \boldsymbol{\theta}) \tag{1.3}$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) \quad . \tag{1.4}$$

Basically, the Bayesian inference relies on the Bayes' formula

$$P(A_i|E) = \frac{P(E|A_i)P(A_i)}{P(E)} = \frac{P(E|A_i)P(A_i)}{\sum\limits_{k=1}^{n} P(E|A_k)P(A_k)}$$

where $A_1, A_2 \ldots, A_n$ is a finite or infinite partition of the sample space $(\Omega, \mathcal{B})$ such that $P(A_j) > 0 \quad \forall j$ and $P(E) > 0\ \forall E \in \Omega$.

Given Equations (1.3) and (1.4), the aim is to calculate "a" posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ for the parameters $\boldsymbol{\theta}$ given data $\boldsymbol{X}$, this represents an update of $\pi(\boldsymbol{\theta})$ after conditioning on the sample data, i.e. the Bayes formula for density provides the posterior density

$$\pi(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{f(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{X})} \quad .$$

Since the marginal distribution of the data $f(\boldsymbol{X})$ is independent from the parameters $\boldsymbol{\theta}$, we can calculate the posterior density to less than a costant

$$\pi(\boldsymbol{\theta}|\boldsymbol{X}) \propto f(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \quad .$$

The posterior distribution forms the basis for all inference, since it contains all relevant information regarding the estimation problem. Relevant information includes both sample data information coming from the likelihood $f(\boldsymbol{X}|\boldsymbol{\theta})$, as well as past or subjective information embodied in the prior distributions of the parameters. The prior choice is determinant for the calculation. To semplify the procedure it is common to choose a prior conjugate to the model, this means that the posterior distribution belongs to the same family of the priors, but with updated parameters.

Tipically we are interested in the expected value of a function $h(\theta)$ of the parameters

$$E_\pi[h(\boldsymbol{\theta})|\boldsymbol{X}] = \int_\Theta h(\boldsymbol{\theta})\pi(\mathrm{d}\boldsymbol{\theta}|\boldsymbol{X}) \quad .$$

The integral could be difficult to compute, thus we approximate the result. Rather than working with the exact posterior density of the parameters, we simulate a large random sample from the posterior distribution. Under some hypothesis of regularity (Harris-recurrence and irreducibility), the invariant distribution of a Markov Chain $\theta_1, \ldots, \theta_M$ is given by the target distribution $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ when M goes to infinity. Applying the strong law of large numbers, choosen M large enough, we can approximate the mean with the ergodic mean of the chain $\theta_1, \ldots, \theta_M$:

$$E_\pi[h(\boldsymbol{\theta})|\boldsymbol{X}] = \frac{1}{M}\sum_{i=1}^{M} h(\theta_i) \quad .$$

From a computational point of view, the only problem is to generate a sample from the posterior when it is not in a closed form. Gibbs Sampler and Metropolis-Hastings algorithms will be useful for this purpose. For more details about Markov Chains, see for example Jackman (2009).

1.3.2 Existence and uniqueness of the joint distribution

In the context or areal data it is natural to calculate the joint distribution using the full conditional distributions $p(y_i|y_j, i \neq j)$ for $i = 1, \dots, n$, which usually have a simpler formula and direct interpretation. Given $p(y_1, \dots, y_n)$, the full conditional distributions are uniquely determined, but unfortunately the converse is not always true. In addition, using the full conditional distributions to determine a joint distribution could lead to an improper result, even if $p(y_i|y_j, i \neq j)$'s are proper for all i. To alleviate this problem we can apply the Brook's Lemma, that concerns the uniqueness of the joint distribution.

Brook's Lemma. *Let $\pi(x)$ be a density for $x \in \mathbb{R}^n$ and define*
$\Omega = \{x \in \mathbb{R}^n : \pi(x) > 0\}$. *Then for $x, x' \in \Omega$ the following holds*

$$\frac{\pi(x)}{\pi(x')} = \prod_{i=1}^{n} \frac{\pi(x_i|x_1, \dots, x_{i-1}, x'_{i+1}, \dots, x'_n)}{\pi(x'_i|x_1, \dots, x_{i-1}, x'_{i+1}, \dots, x'_n)} = \prod_{i=1}^{n} \frac{\pi(x_i|x'_1, \dots, x'_{i-1}, x_{i+1}, \dots, x_n)}{\pi(x'_i|x'_1, \dots, x'_{i-1}, x_{i+1}, \dots, x_n)} \quad .$$

Fixed a point $\boldsymbol{y^{(0)}} = (y_1^{(0)}, \dots, y_n^{(0)})$, applying iteratively the Brook's Lemma, we obtain :

$$p(y_1, \dots, y_n) = \frac{p(y_1|y_2, \dots, y_n)}{p(y_1^{(0)}|y_2, \dots, y_n)} \cdot \frac{p(y_2|y_1^{(0)}, y_3, \dots, y_n)}{p(y_2^{(0)}|y_1^{(0)}, y_3, \dots, y_n)} \cdot \dots$$
$$\dots \cdot \frac{p(y_n|y_1^{(0)}, \dots, y_{n-1}^{(0)})}{p(y_n^{(0)}|y_1^{(0)}, \dots, y_{n-1}^{(0)})} \cdot p(y_1^{(0)}, \dots, y_n^{(0)})$$

for all point $y = (y_1, \dots, y_n)$.

Thanks to this relation, instead of calculating the joint distribution, we can work with n unidimensional full conditional distributions. It is worth to notice that this is very useful for large n.

In a spacial models, referring to the proximity matrix W, we can imagine that the full conditional distribution only depends on the set of neighbours of i, namely ϱ_i. The full conditional distribution becomes

$$p(y_i|y_j, j \neq i) = p(y_i|y_j, j \in \varrho_i) \tag{1.5}$$

We want to be sure that using local specifications does not change the uniqueness and the stationarity of the joint distribution, this assumption is largely used in the theory of

Markov Random Field (MRF).
Intuitively we can represent a set of random variables with a graph: every node is associated with one variable and two nodes are connected only if the corresponding variables are correlated. For example, the graph in Figure 1.2 rapresents a problem where there are seven variables $X_1, .., X_7$, such that X_1 and X_3 are uncorrelated, while X_4 and X_7 are correlated.

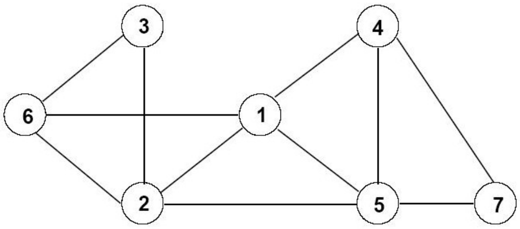

Figure 1.2: Example of a graph.

To better understand the developement of this dissertation, we have to mention some preliminary results on MRF.

Definition of Markov Random Field. *Given an undirected graph $G = (V, E)$, where V is the set of nodes and E of arches, and a set of random variables X_v, $v \in V$, then $X = \{X_v, v \in V\}$ forms a Markov Random Field with respect to G if $X = \{X_v, v \in V\}$ satisfy the local Markov properties:*

- *pairwise Markov property: if u and v are no adjacent variables, in other words $(u, v) \notin E$, then u and v are conditionally independent with respect to the other variables, i.e.*

$$X_u \perp X_v \mid X_{V \setminus \{u,v\}} \quad \forall (u, v) \notin E$$

- *local Markov property: a variable is conditionally independent from all other variables with respect to its neighbors, i.e.*

$$X_v \perp X_{V \setminus \varrho_v} \mid X_{\varrho_v} \quad \forall v \in V$$

- *global Markov property: given $A, B \in V$ and a separate subset $S \in V$ such that* *every path from node A to node B passes across S, then:*

$$X_A \perp X_B \mid X_S \quad .$$

In order to understand the above definition, an example is a stochastic process $X = (X_t)_{t>0}$ adapted with respect to the probability space $(\Omega, \mathcal{F}, \mathcal{F}_s, P)$ [1] . In this case the only neighbour of the variable X_t is X_{t-1}. A stochastic process is Markov if the three properties above hold, in other word X is a MRF if the conditional probability distribution of a future state depends only on the present state, i.e.

$$P(X_t \mid X_{t-1}, \dots, X_1) = P(X_t \mid X_{t-1}) \quad \forall t.$$

Since it can be difficult to verify all these properties for a generic graph, we focus on those graphs that can be factorized by cliques. The same procedure can be extended to all the graphs under other hypoteses.

Definition of Cliques. *Given an undirected graph $G = (V, E)$, a subset $C \in E$ is a clique if $(u, v) \in E \quad \forall u, v \in E$; $k = |C|$ is said size of the clique.*

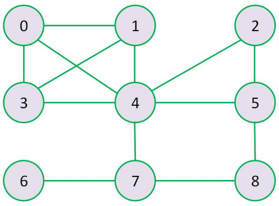

Figure 1.3: Example of clique: the subset $\{0, 1, 3, 4\}$ is a clique because all the nodes are connected.

If $k = 1$, i.e. none node has neighbours, then the model is independent; with $k \geq 2$ we start to introduce a spatial structure.

Definition of Potential function. *$f(x_1, x_2, \dots, x_k)$ is a potential function of order k if it is exchangeable with respect to its arguments, i.e.*

$$f(x_1, x_2, \dots, x_k) = f(s(x_1, x_2, \dots, x_k))$$

for any $s(x_1, x_2, \dots, x_k)$ permutation of $x_1, x_2, \dots, x_k$.

[1] $\mathcal{F}$ is the $\sigma - algebra$ that makes the whole the process measurable, $\mathcal{F}_s$ is the filtration, i.e. $\mathcal{F}_s = \sigma\{X_u, \forall u < s\}$ is the $\sigma - algebra$ that makes the process measurable since istant s .

Definition of Gibbs distribution. *$p(y_1,\dots,y_n)$ is a Gibbs distribution if it is a function of y_i only through potential function on clique:*

$$p(y_1,\dots,y_n) \propto exp\left\{\gamma \sum_k \sum_{\boldsymbol{\alpha}\in\mathcal{M}} \phi^{(k)}(y_{\alpha_1},\dots,y_{\alpha_k})\right\} \tag{1.6}$$

where $\phi^{(k)}$ is a potential of order k, $\mathcal{M}$ is the collection of all the cliques of size k from $\{1,\dots,n\}$ and $\gamma > 0$ is a parameter.

In order to prove that the full conditional distributions in (1.5) define a unique joint distribution, one can use the following fundamental theorem of random fields.

Hammersley-Clifford Theorem. *A probability distribution P with positive and continuous density f satisfies the pairwise Markov property with respect to an undirected graph G if and only if it factorizes according to the cliques of G.*

Applying the Hammersley-Clifford Theorem, we can deduce that (1.6) is a probability distribution on a MRF.
Now fix $k = 2$ and take the potential function as $\phi = (y_i - y_j)^2$, $j \in \varrho_i$, the Gibbs distribution becomes

$$p(y_1,\dots,y_n) \propto exp\left\{-\frac{1}{2\tau^2}\sum_{i,j}(y_i - y_j)^2\mathbb{1}(i\sim j)\right\} \tag{1.7}$$

and the respective full conditional distributions are

$$p(y_i|y_j, j\neq i) = \mathcal{N}\left(\sum_{j\in\varrho_i}\frac{y_i}{n_i}, \frac{\tau^2}{n_i}\right) \quad \forall i \tag{1.8}$$

where $n_i = \sum_{k=1}^{n} w_{i,k}$ is the number of the neighbors of unit i.
The relationship between (1.7) and (1.8) can be easily proved (see for example Carlin and Banerjee, 2003b). From Equation (1.7), let us write the joint distribution as

$$p(y_1,\dots,y_n) \propto exp\left\{-\frac{1}{2\tau^2}\sum_{i,j}(y_i - y_j)^2\mathbb{1}(i\sim j)\right\} = \prod_{j=1}^{n} exp\left\{-\frac{1}{2\tau^2}\sum_{i\neq j} w_{i,j}(y_i - y_j)^2\right\} =$$

$$= exp\left\{-\frac{1}{2\tau^2}\left[\sum_{i=1}^{n} n_i y_i^2 + \sum_{i=1}^{n}\sum_{j\neq i} w_{i,j}y_i y_j\right]\right\} \quad .$$

Hence, collecting only the terms involving y_i and keeping in mind that W is symmetric, one finds out that

$$p(y_i|y_j, j \neq i) \propto exp\left\{-\frac{1}{2}\left[n_i y_i^2 - 2y_i \sum_{j \neq i} w_{i,j} y_j\right]\right\} \quad .$$

The result in (1.8) now follows simply completing the square.

Distributions (1.8) are clearly in the form of distributions (1.5), so we have demostrated that, given the local full conditional distributions, we can find a unique joint distribution for $Y_1, \ldots, Y_n$.

1.4 Conditionally Autoregressive Models (CAR)

An intuitively way to define an areal model is setting the full conditional distributions. Different sets of full distributions identify different models. The Conditionally Autoregressive Model (CAR) is one of the most important; it was introduced by Besag in the 1970s and became very popular because of the simple form of its full conditional distributions. See Getis (2008).

In the following we deal with normal CAR model, but CAR can be generalized to the exponential family.

Given $Y_1, \ldots, Y_n$ areal data with contingency matrix A, let $Y_{-i} = (y_j, j \neq i)$ and set

$$Y_i|Y_{-i} = y_{-i} \sim \mathcal{N}\left(\sum_j a_{i,j} y_j, \tau_i^2\right) \qquad i = 1, \ldots, n \tag{1.9}$$

We can easily recognise a distribution of the form (1.7), so applying the previous results we can obtain the joint distribution

$$p(y_i, \ldots, y_n) \propto exp\left\{-\frac{1}{2}\boldsymbol{y}' D^{-1}(I - A)\boldsymbol{y}\right\}$$

where D is a diagonal matrix with $D_{i,i} = \tau_i^2$. It seems to be a multivariate normal distribution

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, (I - A)^{-1}D) \quad . \tag{1.10}$$

Actually we have to verify that

$$\Sigma = (I - A)^{-1}D$$

really represents a covariance matrix. If we set $\tau_i^2 = \tau^2/n_i$, it is immediate to verify the simmetry of

$$\Sigma^{-1} = D^{-1}(I - A) = (D_w - W)/\tau^2$$

where D_w is diagonal with $(D_w)_{i,i} = n_i$. Instead $(D_w - W)\mathbf{1} = \mathbf{0}$, so Σ^{-1} is singular. Hence the joint distribution of $\boldsymbol{Y}$ is improper. This is a problem if we want to use (1.6) as a model for data. One way to avoid this problem is to weight the mean of the neighbours by a suitable parameter $\rho \neq 1$, in such a way that Σ^{-1} becomes $\Sigma^{-1} = D_w - \rho W$. The full conditionals become

$$Y_i|\boldsymbol{Y}_{-i} = y_{-i} \sim \mathcal{N}\left(\rho \sum_j a_{i,j} y_j, \frac{\tau^2}{n_i}\right) \qquad i = 1, \ldots, n \tag{1.11}$$

The model described in (1.10) is named *proper CAR* . We have to choose a value of ρ that makes Σ^{-1} non singular. There are different approches that lead to different intervals of values. The first one is based on the

Gershgoring disk Theorem. *Given a symmetric matrix C, if* $c_{i,i} > 0$ *and* $\sum\limits_{i \neq j} |c_{i,i}| < c_{i,i}$ *for all i, then C is positive definite.*

We can apply this result to $D_w - \rho W$ that is symmetric and a sufficient condition that implies Gershgoring Theorem is $|\rho| < 1$.
The second one provides a more narrow interval: in literature it has been proved that $D_w - \rho A$ is non singular if $\frac{1}{\lambda_1} < \rho < \frac{1}{\lambda_n}$, where $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ are the eigenvalues of $D_w^{-\frac{1}{2}} W D_w^{-\frac{1}{2}}$. In a Bayesian context, a classical prior distribution for ρ is a uniform distribution in the selected interval, i.e.

$$\rho \sim \mathcal{U}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_n}\right).$$

Parameter ρ can be interpretated as a measure of the spatial correlation between the data. One can prove that λ_n is equal to 1 and that $\lambda_1 < 0$. There is a strong positive spatial correlation for $0.8 < \rho < 1$, instead if there is negative correlation, then ρ would be negative. We notice that ρ could be 0, this is equivalent to set up an independent model, without spatial correlation. It is important to underline that, despite the introduction of the variable ρ is necessary to obtain a proper joint distribution, it changes the mean of the conditional disribution (see 1.10) and can reduce the breadth of the spatial pattern. Referring to (1.10), we may re-write the system of random variables as

$$\begin{aligned} \boldsymbol{Y} &= \rho W \boldsymbol{Y} + \boldsymbol{\epsilon} \\ \boldsymbol{Y} &= (I - \rho A)^{-1} \boldsymbol{\epsilon} \quad . \end{aligned}$$

A fundamental feature of this model is that the vector of errors has the law

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, D(I - A)') \quad .$$

In other words the errors are not independent as in a general linear regression, this is because we are modeling spatial dependence.
It is possible to introduce a regression component into (1.10), without changing the idea of the model, but only adding a term in the mean structure as follows:

$$\boldsymbol{Y} | X, \boldsymbol{\beta}, \tau^2, \rho \sim \mathcal{N}(X\boldsymbol{\beta}, (I - \rho A)^{-1} D) \quad . \tag{1.12}$$

Finally note that by means of (1.7) an equivalent rapresentation of (1.11) in terms of full conditionals is provided by

$$Y_i | X, \boldsymbol{Y}_{-i}, \boldsymbol{\beta}, \tau^2, \rho \sim \mathcal{N}\left(\boldsymbol{X_i'}\boldsymbol{\beta} + \rho \boldsymbol{A}_i \boldsymbol{Y}, \frac{\tau^2}{n_i}\right) \quad i = 1, \ldots, n \quad . \tag{1.13}$$

1.4.1 Introduction of spatial random effects

As we previously said, we do not apply a CAR model to the data, but we prefer to introduce some spatial random effects. A model alternative to (1.12) can be obtained by substituting $\rho \boldsymbol{A}_i \boldsymbol{Y}$ in the mean structure with one spatial random effect ϕ_i for each i, such that $\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_n)$ evolves as a spatial CAR model. We introduce the notation

$$\boldsymbol{\Phi} \sim CAR(A, \tau^2, \rho) \tag{1.14}$$

that stands for

$$\boldsymbol{\Phi} \sim \mathcal{N}(\mathbf{0}, (I - \rho A)^{-1} D) \quad .$$

Furthermore $Y_1, \ldots, Y_n$ are assumed to be independent with

$$Y_i | \boldsymbol{\beta}, \tau^2, \rho \sim \mathcal{N}(\boldsymbol{X}_i \boldsymbol{\beta} + \phi, c_i) \tag{1.15}$$

where c_i is a generic notation for the variance of y_i. It is worth to underline that if we expect areal correlation among $Y_1, \dots, Y_n$ and we set a linear regression model

$$\boldsymbol{Y} \sim \mathcal{N}(X\boldsymbol{\beta}, \Sigma)$$

the covariance matrix Σ is not diagonal and hence $Y_1, \dots, Y_n$ are not independent. This represents a problem in a statistical analysis because we need to factorize the likelihood expression: for this reason it is recommended to set the model in the form (1.13) and (1.14).

We can find lots of exemples of this approach in literature; one can see frameworks like Stern and Cressie (1999), Banerjee and Carlin (2015), Epifani and Nicolini (2015b) among the others.

Chapter 2

Time series

A time series process $\{X_t, t \in \mathcal{T}\}$ ia a stochastic process or a collection of random variables X_t indexed in time. If $\mathcal{T} \subseteq \mathbb{N}$ then the process is discrete in time, if $\mathcal{T} \subseteq \mathbb{R}$ then it is a continuous time random process. We will indicate with X_t, $t = 1, \dots, T$, a collection of T equally spaced realization of a time series process.
The aim of time series analysis is to describe the dependence among a sequence of random variables, the hypotesis is that they can not be independent realizations from a unique distribution. If we are able to identify a trend, i.e. a stochastic process that has trajectories that describe the data, it is possible to make prevision for the value in the future.
Many time series models are based on the assumption of stationarity.

Dedinition of strong stationarity. *A time series process $\{X_t, t \in \mathcal{T}\}$ is strongly stationary if, for any sequence of time $t_1, \dots, t_n$ and any lag h, the probability distribution of the vector $(X_{t_1}, \dots, X_{t_n})$ is identical to the probability distribution of the vector $(X_{t_1+h}, \dots, X_{t_n+h})$.*

This mean that the realizations of the process do not depend on the starting time. However this definition is not operative, because it is difficult to verify for a generic process. We can apply the definition of weak stationarity.

Definition of weak stationarity. *A time series process $\{X_t, t \in \mathcal{T}\}$ is weakly stationary if, for any sequence of time $t_1, \dots, t_n$ and any lag h, the first and the second moments of* *the vector* $(x_{t_1}, \ldots, x_{t_n})$ *are identical to the first and the second moments of the vector* $(x_{t_1+h}, \ldots, x_{t_n+h})$.

In other words, the mean and the variance of the process are constant over time, and the covariance function γ between two different realizations depends only on del lag of the time, i.e.

$$\begin{aligned} E[X_t] =& \mu \qquad \forall t \in \mathcal{T} \\ Var[X_t] =& v \qquad \forall t \in \mathcal{T} \\ Cov(X_t, y_s) =& \gamma(t-s) \qquad \forall s < t \in \mathcal{T}. \end{aligned}$$

Intuitively, a stochastic process is stationary if its probabilistic structure is constant over time, so that the process is easyer to analyse.

In an arbitrary model we can set X_t as a function of the past values, of some parameters $\boldsymbol{\theta}$ and of the estimation error ϵ, i.e.

$$x_t = f(x_0, x_1, \ldots, x_{t-1}, \boldsymbol{\theta}, \epsilon) \qquad \forall t.$$

From a Bayesian point of view, once set a prior distribution on the parameters $\pi(\boldsymbol{\theta})$, we can apply the Bayes' Theorem, and get the posterior distribution

$$p(\boldsymbol{\theta}|x_0, \ldots, x_t) \propto \prod_{n=1}^{t} p(x_t|x_{-t}, \boldsymbol{\theta}) p(x_0|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \quad .$$

In a more complex case it can happen that also the parameters depend on the time $(\boldsymbol{\theta}_t)_{t \in \mathcal{T}}$ and we have to introduce a dynamic for the behavior of the parameters

$$\begin{aligned} x_t &= f(x_0, x_1, \ldots, x_{t-1}, \boldsymbol{\theta}_t, \epsilon) \qquad \forall t \\ \boldsymbol{\theta}_t &= g(\theta_0, \theta_1, \ldots, \theta_{t-1}, \Phi, \nu) \qquad \forall t. \end{aligned}$$

In this case it becomes more difficult to obtain the posterior distribution for $\boldsymbol{\theta}_t$ because the calculus depends on the specific case.

A useful method to verify the dependence between the data is by the autocorrelation function (ACF) ρ defined as

$$\rho(x_s, x_t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}} \quad , \forall s, t \in \mathcal{T}.$$

If the model is stationary and $h = |t - s|$, we can write the ACF function in the form

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} \quad .$$

If $\rho(h)$ is different from zero there is an effective correlation among the data.

2.1 Autoregressive Models

Among the several different models for time series, let us discuss the properties of the Autoregressive (AR) models, because they are the simplest class of empirical models for exploring dependences over time. So they are the basis for more complex models.
In AR models the random variable X_t is function of the past values.

Definition of AR models. *X_t is an Autoregressive model of order p (AR(p)) if*

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \epsilon_t \qquad t = 1, 2, \ldots$$

where ϵ_t is the error at time t.

We assume that the errors are independent and normally distributed:

$$\epsilon_t \sim \mathcal{N}(0, v) \qquad \forall t.$$

The sequential nature of the model implies a sequential structure of the data distribution

$$p(x_1, \ldots, x_T | \boldsymbol{\phi}, \boldsymbol{\epsilon}) = p(x_1, \ldots, x_p | \boldsymbol{\phi}, \boldsymbol{\epsilon}) \prod_{t=p+1}^{T} p(x_t | x_{t-1}, \ldots, x_{t-p}, \boldsymbol{\phi}, \epsilon_t) \quad .$$

We assume to know the initial values $x_1, \ldots, x_p$, so we obtain

$$p(x_1, \ldots, x_T | x_1, \ldots, x_p, \boldsymbol{\phi}, \boldsymbol{\epsilon}) = \prod_{t=p+1}^{T} p(x_t | x_{t-1}, \ldots, x_{t-p}, \boldsymbol{\phi}, \epsilon_t) = \mathcal{N}(F' \boldsymbol{\phi}, vI) \tag{2.1}$$

where $F = [\boldsymbol{f}_T, \ldots, \boldsymbol{f}_{p+1}]$ and $\boldsymbol{f}_t = (x_{t-1}, \ldots, x_{t-p})'$. The above distribution is very generic, we can introduce some extensions like a nonzero mean for the variable X_t, a variance of the error that changes over time, etc.

2.1.1 Stationarity for AR models

Let us now introduce some criteria in order to guarantee AR models stationarity.

Definition of causality. *An AR(p) process is causal if it can be written as a linear process dependent on all the past events*

$$X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \quad .$$

In order to verify this condition, let us write the model using a backshift operator s, i.e. $x_{t-1} = sx_t$ such that

$$\begin{aligned}
&x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \epsilon_t \\
&x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} = \epsilon_t \\
&x_t(1 - \phi_1 s - \cdots - \phi_p s^p) = \epsilon_t \\
&x_t = \Phi^{-1}(u)\epsilon_t
\end{aligned}$$

$\Phi(u)$ is called *autoregressive characteristic polynomial*, and x_t is causal if the roots u of $\Phi^{-1}(u) = 0$ satisfy $|u| > 1$. This causality condition implies stationarity. Alternatively, we can write the characteristic polynomial as $\Phi(u) = \sum_{j=1}^{p}(1 - \alpha_j u)$, in this case $u_j = \frac{1}{\alpha_j}$ and the causality condition becomes $|\alpha_j| < 1$.

2.1.2 Autocorrelation structure and Partial Autocorrelation Function

The autocorrelation structure of an AR(p) is given in terms of the solution of the equation

$$\rho(h) - \phi_1 \rho(h-1) - \cdots - \phi_p \rho(h-p) = 0 \qquad h \geq p \tag{2.2}$$

Let us call $m_1 \ldots, m_r$ the multiplicy of the roots of $\Phi(u)$. Then the general solution of (2.2) is

$$\rho(h) = \alpha_1^h p_1(h) + \alpha_2^h p_2(h) + \cdots + \alpha_r^h p_r(h) \qquad h \geq p$$

where $p_j(h)$ is a polynomial of degree $m_j - 1$.

Another important quantity to better understand the correlation between the data is the *Partial Autocorrelation Function* (PACF). The PACF is a function defined by the coefficients at lag h, each coefficient $\phi(h, h)$ is a function of the best linear predictor of

x_t given $x_{t-1}, \ldots, x_1$. The predictor is $x_h^{h-1} = \beta_1 x_{h-1} + \cdots + \beta_{h-1} x_1$, where $\beta_1, \ldots, \beta_{h-1}$ are chosen by minimizing the mean square linear prediction error $E(x_h - x_h^{h-1})^2$. If the process is stationary, it is known that $x_0^{h-1} = \beta_1 x_1 + \cdots + \beta_{h-1} x_{h-1}$ with the same coefficients of the expression of x_h^{h-1} and

$$\phi(h,h) = \begin{cases} \rho(x_1, x_0) & h = 1 \\ \rho(x_h - x_h^{h-1}, x_0 - x_0^{h-1}) & h > 1 \end{cases} .$$

The values of $\phi(h,h)$ help to estimate the correct order p of AR: if $(x_t)_{t \geq 0}$ follows an AR(p), then $\phi(h,h) = 0$ for $h > p$.

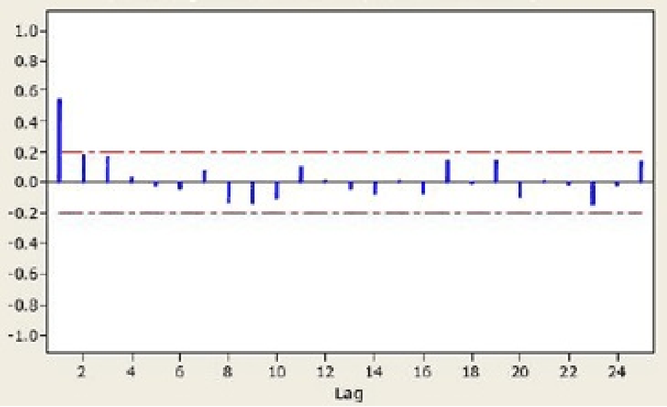

Figure 2.1: Example of PACF function: the value of the coefficients are non influential after the first time, so the correct order for this data set is p=1.

2.1.3 Bayesian inference for AR models

We use a general analysis from a Bayesian point of view: the results obviously depend on the prior choice, in this framework we present only a general case, we will not discuss all the possible cases, for a more detailed dissertation you can see for example Prado and West (2014).

The parameters of a generic AR(p) model are $\boldsymbol{\phi}, v$. Once we obtain an estimation of $\boldsymbol{\phi}$ we can also estimate the roots of the caracteristic polynomial $\boldsymbol{\alpha}$ in order to verify the stationarity of the model. In the previous paragraphs we deal with the order of the model p as a constant. For a deeper analysis we can treat it as a parameter itself. An easy

way to determine the correct value of p is to repeat the analysis with an increasing value of it and compare the results according to some criterion like AIC or BIC for example. The only precaution is to use a value of p small enough with respect to T, otherwise problems like overfitting or collinearity can occur. In a Bayesian context we assume a prior distribution over p.
The posterior distribution is

$$\pi(\boldsymbol{\phi}, v, p|\boldsymbol{x}) = p(x_T, \ldots, x_{p+1}|\boldsymbol{\phi}, v, p, x_p, \ldots, x_1)\pi(x_p, \ldots, x_1|\boldsymbol{\phi}, v, p)\pi(\boldsymbol{\phi}, v|p)p(p)$$

In this section we do not really care about the prior distribution on p, we limit our consideration on the prior distributions $\pi(\boldsymbol{\phi}, v|p)$ and statistical model $p(x_p, \ldots, x_1|\boldsymbol{\phi}, v, p)$ and consider p as a fixed number.
There are two different ways to select a prior for the initial values: the first one is to choose a distribution independent from the parameters, just to initialize the time series, in this case the analysis is independent by the initial values. The second way is to set a prior distribution that depends on the parameters; under this hypothesis the final result could dependent on the initial values: the effect of the initial values is fixed, if the time serie is "long", i.e. T$\gg$ p, then $p(x_p, \ldots, x_1|\boldsymbol{\phi}, v, p)$ is negligible with respect to $p(x_T, \ldots, x_{p+1}|\boldsymbol{\phi}, v, p, x_p, \ldots, x_1)$, otherwise the analysis will depend on the value of $x_1, \ldots, x_p$.
The choice for $\pi(\boldsymbol{\phi}, v|p)$ is more relevant. First of all it is important to remind that the prior for $\boldsymbol{\phi}$ and v should be concentrated only in the stationarity region and be zero outside. In a simulation approach there is no way to verify stationarity condition before having estimated $\boldsymbol{\alpha}$, so as we have said previously, firstly we have to estimate $\boldsymbol{\phi}$, then calculate $\boldsymbol{\alpha}$. The simplest method is to proceed in an unconstrained analysis and then reject the values of $\boldsymbol{\phi}$ that stay outside the stationarity region. If the model is really stationary, the rejection rate will be low and the analysis reliable. But it can happen that the rejection rate is high; in such a case probably the stationarity assumption does not work for the model, the analysis is not efficient and hence other methods are needed.
The prior $\pi(\boldsymbol{\phi}, v|p)$ depends strongly on p, to avoid problems in calculation it is suggest to assume an improper prior

$$\pi(\boldsymbol{\phi}, v|p) \propto \frac{1}{v} \quad .$$

Under stationarity assumption, the problem has a multinormal likelihood

$$\begin{aligned} &p(x_p, \ldots, x_1 | \boldsymbol{\phi}, v, p) \sim \mathcal{N}(0, vA_\phi) \\ &p(x_T, \ldots, x_{p+1} | \boldsymbol{\phi}, v, p, x_p, \ldots, x_1) \sim \mathcal{N}(F'\boldsymbol{\phi}, vI) \end{aligned}$$

so we obtain

$$\begin{aligned} p(x_T, \ldots, x_1 | \boldsymbol{\phi}, v, p) \propto \frac{1}{v^{T/2}|A_\phi|^{1/2}} \times \\ \times\, exp \left\{ \frac{(x_T, \ldots, x_{p+1})'(F'\boldsymbol{\phi})^{-1}(x_T, \ldots, x_{p+1}) + (x_1, \ldots, x_p)' A_\phi^{-1}(x_1, \ldots, x_p)}{2v} \right\} \end{aligned}$$

In order to obtain a conjugate model, the best prior choice for the parameters is the Jeffreys' prior $\pi(\boldsymbol{\theta}) \propto \sqrt{|\mathcal{I}(\boldsymbol{\theta}|\boldsymbol{x})|}$, where $\mathcal{I}(\boldsymbol{\theta}|\boldsymbol{x})_{i,j} = -E[\partial^2 log(p(\boldsymbol{x}|\boldsymbol{\theta}))/\partial\theta_i\partial\theta_j]$. In literature (see Box, Jenkins and Reinsel, 2008) it is known that Jeffreys' prior for this specific problem is approximately $\pi(\boldsymbol{\phi}, v) \propto |A_\phi|^{1/2} v^{-1/2}$. With this prior we can proceed with a Gibbs algorithm, because the full conditionals are known "popular" distribution: inverse-gamma for v and multinormal distribution for $\boldsymbol{\phi}$.

Chapter 3

A model for the population density

In this chapter, we introduce the data set and the general goal of the analysis, then we compare different models in order to interpret the problem. After describing the model, the calculation of the full conditionals and the sampling scheme are illustrated.

3.1 The data

The analysis is applied to the state of Massachusetts. The data set comes from the NHGIS project[1]; it contains the census tract data for the period between 1970 and 2010. As the census are repeated every 10 years, then the data are affected by the important changes in the census tract composition across time, anyway it is possible to compare them over different periods thanks to the fundamental re-elaboration made by the NHIGIS project, that grants compatibility across time. The most difference is between the first and second time (1970 and 1980) with the other periods (1990-2010). During the analysis it becomes clear that the result obtained in 1990-2010 are more accurate.

First of all it is important to explain what a census tract is: according to the definition provided by the US census, a census tract is defined as a spatial area whose population size ranges between 1500 and 8000 inhabitants, with an optimal size of 4000.

Massachusetts has 14 counties.

[1] Minnesota Population Center. National Historical Geographic Information System: Version 2.0. Minneapolis, MN: University of Minnesota 2011.

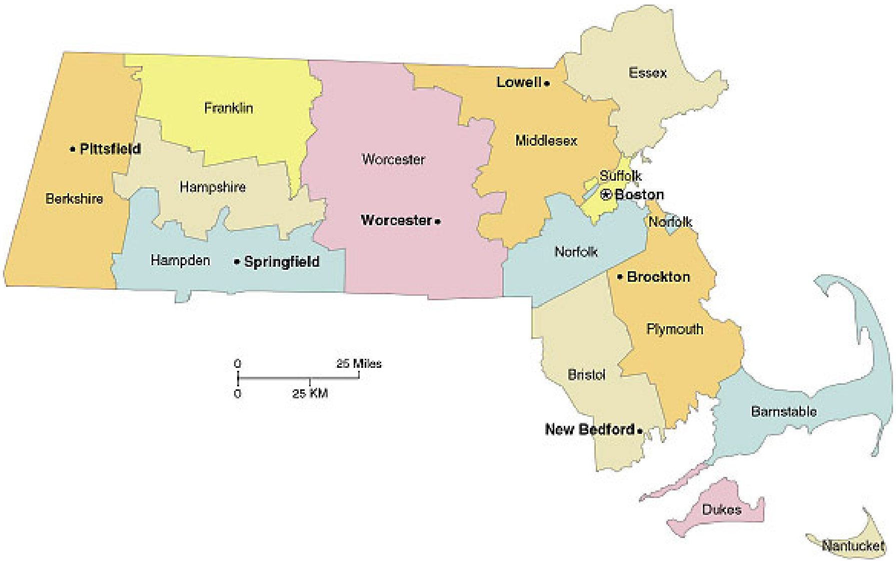

Figure 3.1: Massachussettes state.

Each county is divided in census tracts, each of them has a centre in which the information about the population are collected. The peculiarity of the data set is that the whole number of census tract changes every year: when a census tract becomes too populated, then it is splitted, while in the opposite case it is absorbed by another one. This means that the census tract evolution is not easily traceable over time, even if they are all identified by a sequential string code: so we cannot describe the evolution of a single census tract over time. But we can study the dynamic of the distribution of the population density across each county. Our observations are the census tracts density y computed as the total population divided by the area of the census tract.

As we mentioned above, the goal of the analysis is to determine which "global" features lead the individual choice, in addition to the personal preferences. The most relevant variable is the distance between Boston and each census tract, calculated as the euclidean distance between the geographic coordinates of Boston and the centroid of the census tract. Previous study (see Epifani and Nicolini, 2013) states that the distance from Boston is the key element for modelling population distribution. In order to focus the contest, we have to introduce the concept of Central Business District (CBD): it is a selected point which includes services, leisure activities, professional and economics centres, infrastructure elements that guarantee mobility.

Once a CBD has been identified, it turns out to be the centre of the density population distribution (see Helsey and Strange, 2007). Economic literature selected Boston as the most attractive city in Massachusetts (see Glaeser, 2005) and it maintained its attractivness across time, since the census tract with highest density have always been in Boston's county. It is important to remind that distance can not be considered costant over time because the census tract changes every time.

Despite the distance from Boston is determinant, basing the analysis only on this predictor would be too restrictive, in fact there are other factors that can have a relevant role, like for example natural amenities and etnic composition (see Topa and Zenou, 2015). The presence of natural amenities can have an important role in the decision process, because they contribute to create space for leasure time. The water is taken as a prototype of amenities because it is the most attractive natural element. Hence it is available a variable *zeta* that summarizes the proportion of water's area in the county; Epifani and Nicolini (2015) has argued that *zeta* can be considered constant over time.

We should also focus on further variables that allow to describe "the reputation" of a zone, these variables are the ethnic composition, the income and the education level. It is easy to understand that an individual prefers to settle close to individuals who share the same level of income and education, but especially who belong to the same ethnic group. As an indicator of the ethnic composition we take the proportion *mix* of white individuals over the total population. As for the income we introduce two distinct local predictors: an indicator of the level of the income per-capita (*income*) per-census tract, and a measure of the dispersion and inequality of income measured by the Gini index (*gini*). The variable *income* is not collected in 1970, so we do not use it in the regression for the first period. The census provides us the joint data on the income's distribution, there are 4 classes for each census tract: income less than 10000 dollars, income between 10000 and 15000 dollars, income between 15000 and 25000 dollars, and finally income more than 25000 dollars. We compute the frequencies on the classes and from that the Gini index with the formula

$$gini = \frac{\sum_{i=1}^{n-1} (P_i - Q_i)}{\sum_{i=1}^{n-1} P_i}$$

where Q_i are the cumulative percentage of the income and P_i are the cumulative percentage if the income would be equidistributed. The Gini index varies between 0 and 1; its value is equal to 0 when there are no inequalities, and increases with inequality, to reach 1 when one individual earns the entire income of the census tract. Instead, in order to propose a comprehensive indicator of the distribution of the education in each census tract we elaborate a synthetic measure of the degree of education (*education*) by ranking all of the census-tracts according to the level of education of its population. First, we rank them according to the relative frequency of citizens having a high degree of education, and then according to the relative frequency of persons having a low degree of education. Hence, for each census tract unit we subtract the second value of ranking from the first. This type of index presents the extent to which a census tract unit may emerge as a highly educated in respect to the rest of the census tract units in Massachusetts. Let us now summarize all the covariates; for each time $t = 1, \dots, 5$, for each census tract $j = 1, \dots, J[t]$ and for each county $i = 1, \dots, 14$ we have the following information:

- $dist_j^{(t)}$: euclidean distance of census tract j from Boston;
- $mix_j^{(t)}$: proportion of white individuals;
- $education_j^{(t)}$: education level;
- $gini_j^{(t)}$: Gini index;
- $income_j^{(t)}$: income pre-capita;
- $cc_j^{(t)}$: county of the j-th census tract;
- z_i : amenities in county i;
- $y_j^{(t)}$: population density of j-th census tract.

3.2 Descriptive statistic

In this section we present an exploratory analysis of the data set.

First of all, let us standardize the predictors substracting the mean and dividing by the standard deviation for year. Furthermore, we transform the population density to a log-function $Y_j^{(t)} = log(Y_j^{(t)})\ \forall j, \forall t$.

We can assert that there are no drammatical change in the population during the whole period, Figure 3.3 rapresents the mean of the logarithm of the population's density for every county and we can see that it is almost constant. We can deduce that the analysis is photographing the evolution of a population's density in normal conditions.

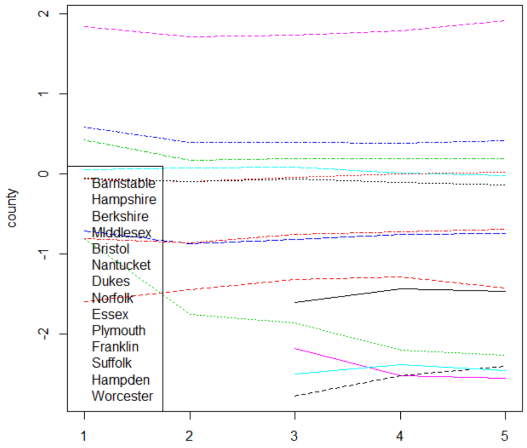

Figure 3.2: Evolution of the mean of ppopulation log-density for every county during the year 1970 (t=1), 1980 (t=2), 1990 (t=3), 2000 (t=4), 2010 (t=5). Notice that data in Barnstable, Franklin, Dukes, Nantucket are not collected in 1970 and 1980 (see Table 3.1) .

It is worth to do some consideration on the census tracts. Referring to Figure 3.4 and Table 3.2, it is clear that the most the population lives in Suffolk, where there is the CBD of Boston, and in the nearby counties. In Suffolk there are lots of census tracts, they are really small so we expected that the population density is very high (as one can see in Table 3.4). In faraway counties, like for example Franklin or Barnstable, there is a little number of census tracts, they are more extended and have a lower density.

The mean of population density is shown in Table 3.3. It is quite clear that we expect that the distance from Boston would be relevant in the descriprion of the evolution of population distribution, there is a strong negative correlation between the two variables (see the scatterplot in Figure 3.6).

	1970	1980	1990	2000	2010
Barnstable	0	0	50	50	56
Hampshire	27	25	30	31	35
Berkshire	15	32	34	41	39
Middlesex	249	271	277	297	317
Bristol	102	105	106	116	125
Nantucket	0	0	4	5	5
Dukes	0	0	4	4	4
Norfolk	101	103	117	121	130
Essex	112	136	146	156	162
Plymouth	46	84	90	90	99
Franklin	0	0	15	16	18
Suffolk	168	177	183	175	193
Hampden	71	83	87	92	103
Worcester	158	157	159	163	171

Table 3.1: Number of census tracts for each county per year .

Regarding the other covariates, we observe a negative correlation between density population and mix of the population and a positive one between population's density and Gini index. So we expect that people prefer to move to more comfortable place with economic wellness and a high percentage of white people.

	1970	1980	1990	2000	2010
Barnstable	NA	NA	-1.60	-1.43	-1.47
Hampshire	-1.59	-1.44	-1.31	0.-1.28	-1.43
Berkshire	-0.81	-1.76	-1.86	-2.20	-2.27
Middlesex	0.58	0.40	0.39	0.39	0.41
Bristol	0.05	0.007	0.009	0.01	-0.02
Nantucket	NA	NA	-2.18	-2.52	-2.55
Dukes	NA	NA	-2.77	-2.52	-2.40
Norfolk	-0.07	-0.09	-0.04	-0.001	0.02
Essex	0.43	0.17	0.19	0.19	0.19
Plymouth	-0.71	- 0.87	- 0.82	-0.75	-0.75
Franklin	NA	NA	-2.49	-2.38	-2.45
Suffolk	1.84	1.71	1.73	1.79	1.92
Hampden	-0.05	-0.09	-0.06	-0.10	-0.14
Worcester	-0.80	-0.87	-0.76	-0.72	-0.69

Table 3.2: Population's log-density mean for each county per year .

	1970	1980	1990	2000	2010
Distance	0.067	-0.288	-0.486	-0.520	-0.537
Mix	-0.281	-0.391	-0.509	-0.588	-0.632
Gini	-0.185	0.531	0.489	0.483	0.505
Education	0.066	-0.300	-0.316	-0.345	-0.322
Amenities	0.406	0.400	0.319	0.327	0.350

Table 3.3: Correlation between the covariates and log-density per year.

By the preliminary analysis we find out that for the first two years we do not have data for four counties, i.e. Barnstable, Nantucket, Dukes and Franklin, because in 1970 and 1980 there were not a structure of city and population distribution that allow to define a census tract. Anyway this do not affect the analysis.

Since the goal of this work is looking for a geographical dependence among data, we

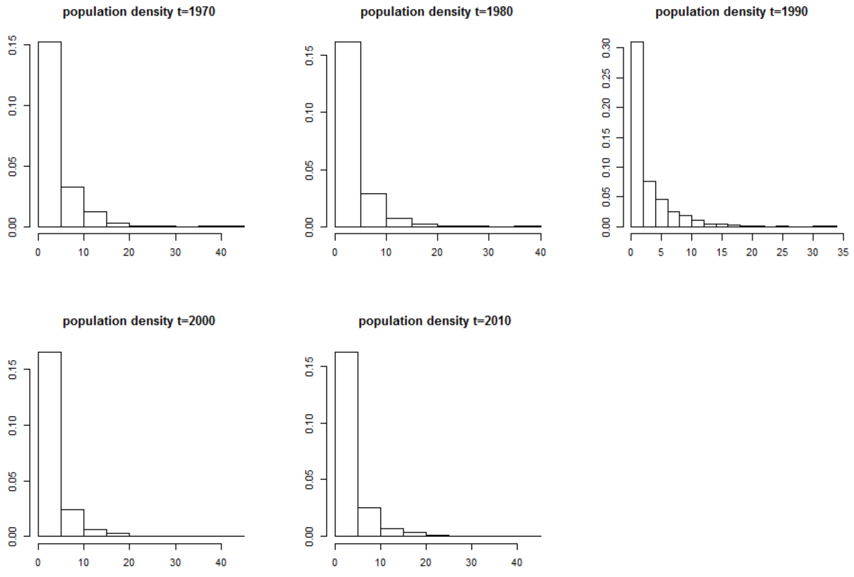

Figure 3.3: Distribution of population density per year: in general the density is quite small over time.

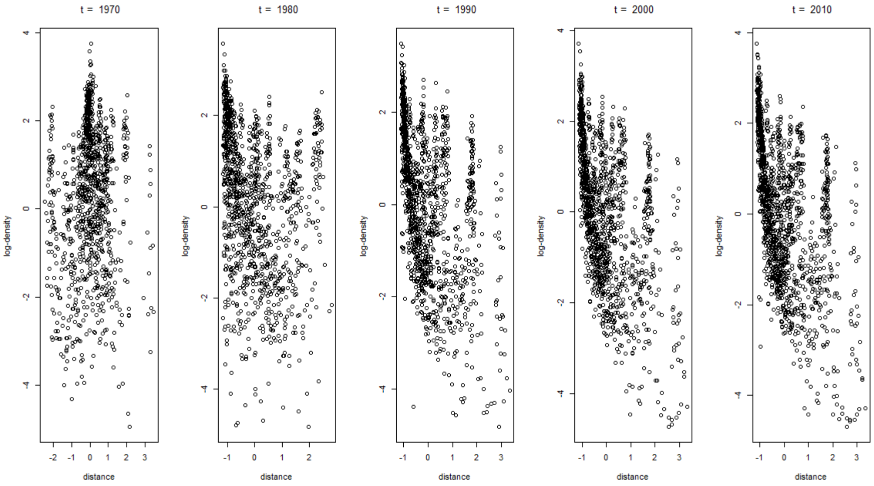

Figure 3.4: Graph of population's log-density and distance per year.

delete those census tracts that have no neighbours or whose population's density is equal to zero. We interpret this occurrence as an error in the collection of data. We end up with a data set composed by a different number of census tract $J[t]$ for every year. As it is reported in Table 3.4, the total number of census tracts grow up over time, so there is an increment of population. Comparing Figure 3.5 and Figure 3.6, we can clearly see

	1970	1980	1990	2000	2010
J	1049	1173	1302	1357	1457

Table 3.4: Total number of census tracts in Massachusetts for every year.

the evolution of the census tracts. In 1970 the population density is low, the most part of the population settles near to the biggest cities while in the countryside there are even spaces without census tracts, i.e. the withe space in the map; in such place we do not have informaion. In 50 years the population has grown: the mean of the density is almost constant, but the number of census tacts has increased, especially near to the cities and in Suffolk (see Table 3.1 and 3.2). Even if the cities remain the residencial centre because of services, aniway the population is spread all over the state, in fact there is not lack of census tracts any more, this means that there are people and structures enough to define a census tract.

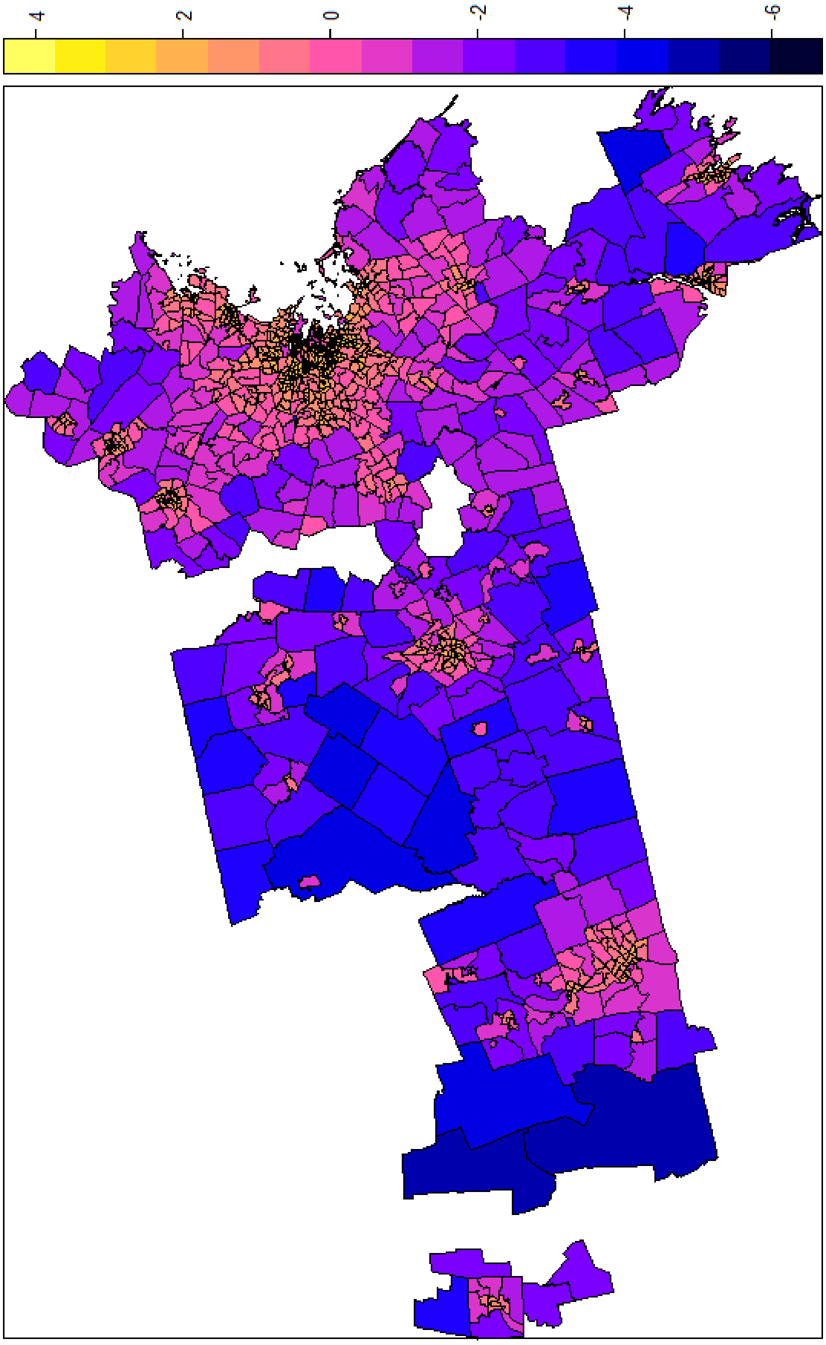

Figure 3.5: Map of census tracts in Massachusettes in 1970.

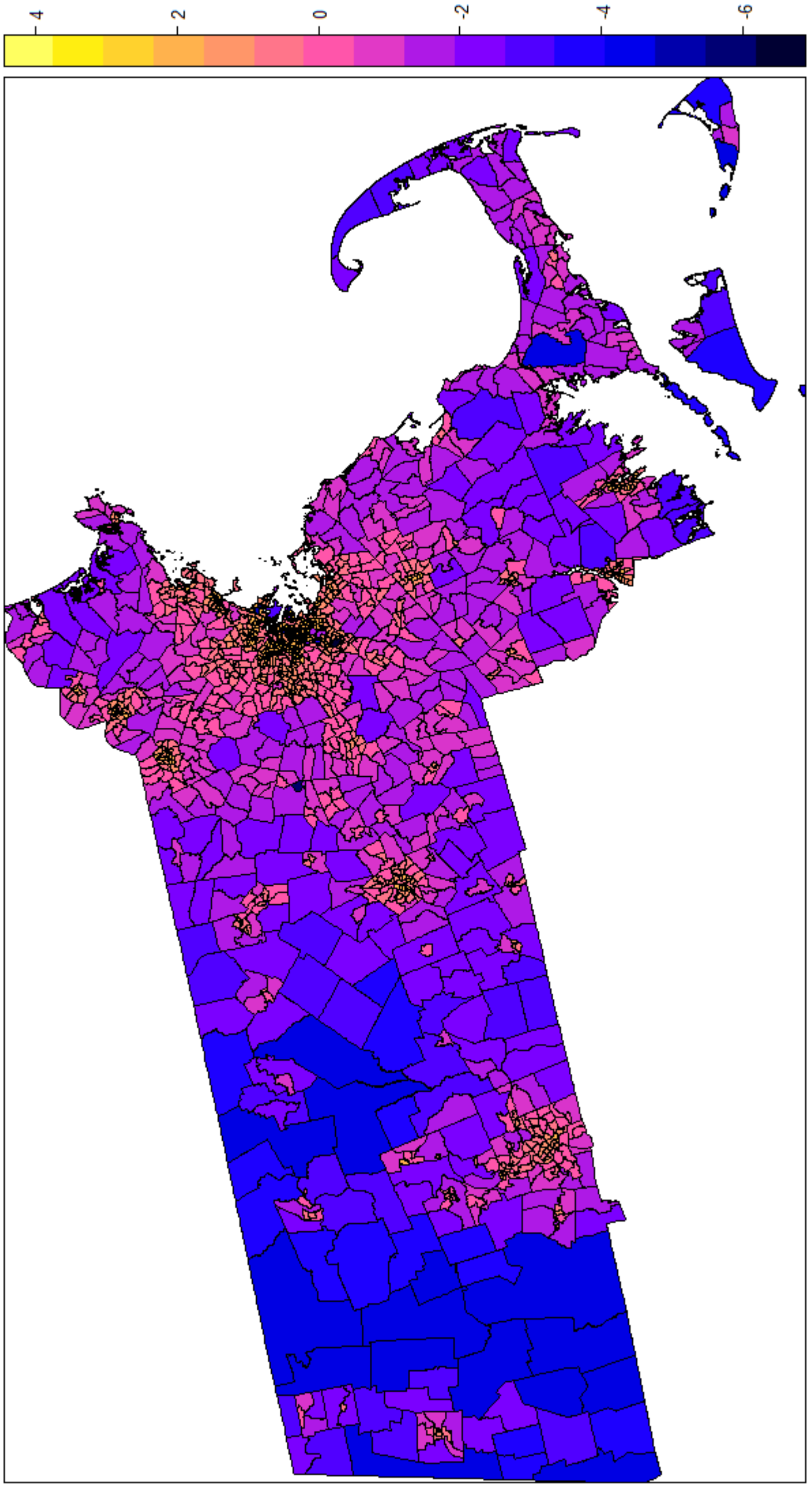

Figure 3.6: Map of census tracts in Massachusettes in 2010.

In order to investigate the spatial correlation, we calculated $Moran's\ I$ and $Geary's\ C$ indices; their values are reported in Table 3.5.

	1970	1980	1990	2000	2010
$Moran's\ I$	0.56	0.61	0.62	0.61	0.64
$Geary's\ C$	0.32	0.28	0.28	0.27	0.21

Table 3.5: Spatial indices.

The $Moran's\ I$ is positive and its value is in the middle between zero and one, hence we expected a global dependence, even if it is not very strong. The local correlation seems to be more pronunced, since the $Geary's\ C$ takes value quite close to zero. Both indices are quite constant, both point out a little increment of the spatial correlation over time.

3.3 The model

A Bayesian approach to the problem is usefull because, by defining a prior structure *ad hoc*, it allows to model time and spatial dipendences.
In order to study the spatial correlation, we propose a comparison between three models, that describe different aspects of the spatial dependence. The first model, assumed as baseline reference, exploits the segmentation of the data within the geographic areas, given by the counties, to assess the population density; the regressors are: distance, mix, education, Gini index, income and amenities z. The second model is the first attempt to outline behavior unexplained; it is obtained by adding independent census tracts random effects. In the third model the random effects are accounted for in a correlated way by means of a CAR model.
In the next the three models are presented in details.

Baseline model

First of all, we address the problem setting a baseline model without any spatial random effects. We implement a regression using the amenities z with the following individual covariates $V_{1,j}^{(t)}, V_{2,j}^{(t)}, V_{3,j}^{(t)}, V_{4,j}^{(t)}, V_{5,j}^{(t)}$ and an iteration term $V_{6,j}^{(t)}$:

$V_{1,j}^{(t)}$	$V_{2,j}^{(t)}$	$V_{3,j}^{(t)}$	$V_{4,j}^{(t)}$	$V_{5,j}^{(t)}$	$V_{6,j}^{(t)}$
Distance	*Mix*	*Gini*	*Educ*	*Income*	$Income * Dist$

The statitical model is

$$\begin{aligned}\log(Y_j^{(t)})|V_j^{(t)}, z_{i[j]}, b_0^{(t)}, b_1^{(t)}, \boldsymbol{\beta}^{(t)}, \nu_{i[j]}, \sigma^2, \Sigma_\beta \sim \mathcal{N}(b_0^{(t)} + b_1^{(t)} z_{i[j]} + V_j^{(t)}\boldsymbol{\beta}^{(t)}, \sigma^2 \nu_{i[j]}) \\ \text{independent} \quad \forall j = 1, \dots, J[t], \forall t = 1, \dots, 5 \qquad .\end{aligned} \tag{3.1}$$

In Equation 3.1 the notation $i[j]$ specifies the county i of the census tract j.
The time dipendence is caugth by the regression coefficients. For $b_0^{(t)}, b_1^{(t)}, \beta_1^{(t)}, \dots, \beta_6^{(t)}$ we assume an $AR(1)$ structure. We suppose that the initial values $b_0^{(0)}, b_1^{(0)}, \beta_1^{(0)}, \dots, \beta_6^{(0)}$ are independent from the hyperparameters. In this way our hypotesis is that the distribution of the population at time t in some way depends on the distribution at time $t-1$. The evolution of the regression coefficients is

$$\begin{aligned}&\boldsymbol{B}^{(t)} = (b_0^{(t)}, b_1^{(t)}, \beta_1^{(t)}, \dots, \beta_6^{(t)}) \\ &\boldsymbol{B}^{(0)} \sim \mathcal{N}(\mathbf{0}, 10^2 I) \\ &\boldsymbol{B}^{(t)}|\boldsymbol{B}^{(t-1)}, \Sigma_B \sim \mathcal{N}(\boldsymbol{B}^{(t-1)}, \Sigma_B) \quad \forall t = 1, \dots, 5 \\ &\Sigma_B = diag\{\sigma_{b0}^2, \sigma_{b1}^2, \sigma_{\beta_1}^2, \dots, \sigma_{\beta_6}^2\}\end{aligned}$$

The dynamic structure of the coefficients is the same in all the models that we propose. Another hypothesis is that there is a county's effects, in other words, census tracts belonging to the same county have same shared features not included within the other covariates. For this reason, we add z as a covariate in the model. We remind that z is a variable relative to the counties and represents the amenities of the county.
We also use a particular variance structure that can accommodate heteroscedastic disturbances or outliers. This particular form for the variance was introduced for linear regression by Geweke (1993). A set of variance scalars $(v_1, v_2, ..., v_{14})$ represents unknown parameters that allow us to assume the error $\epsilon \sim \mathcal{N}(0, \sigma^2 V)$, where V is a diagonal matrix containing parameters $(v_1, v_2, ..., v_{14})$. All the v_i are *iid* and their priors take the form of a set of independent $IG\,(r/2, r/2)$ distributions. This allows us to estimate

the additional 14 variance scaling parameters v_i by adding only a single parameter r to our model. We decide to adopt this structure for variance: in our model the idea is to set one variable ν_i for each county, so we can capture heteroschedasticy among different counties by comparing the value of ν. From a pracical point of view, in literature the hyperparameter r has been set equal to 4, for more details one can see LeSage and Pace (2009). While the variable σ^2 is constant for every census tract, σ^2 wants to capture the intrnisec variability that is common to the log density of population of all census tracts. Therefore the variance prior is

$$\begin{aligned}&\nu_i \sim IG\left(\frac{r}{2}, \frac{r}{2}\right), \quad \forall i = 1, \dots, 14\\ &r = 4\\ &\sigma^2 \sim IG(0.001, 0.001) \quad .\end{aligned}$$

In the baseline model (3.1), the prior distributions for the variances of the regression parameters are choosen in a standard way to be not informative, i.e.

$$\sigma_{b_0}, \sigma_{b_1}, \sigma_{\beta_1}, \dots, \sigma_{\beta_6} \overset{\text{iid}}{\sim} U(0, 10) \quad .$$

It is worth to underline that all the variances' structures are independent on the time. This is a strong assumption because we adfirm that the variability of the phenomenon is constant over time.

Indipendent random effect model

The basic idea for the second model is to verify if, in addition to heteroschedasticy among counties, there is heterogenity even among the census tracts. For this reason we add a random effect $\phi_j^{(t)}$ for each census tract j. The random effects are all independent each other; they aim at capturing outliers and census tract's behavior, that is particularly different from the others. The resulting normal model is:

$$\begin{aligned}\log(Y_j^{(t)})|V_j^{(t)}, z_{i[j]}, (b_0^{(t)}+b_1^{(t)}, \boldsymbol{\beta}^{(t)}, \nu_i, \sigma^2, \Sigma_\beta, \phi_j^{(t)}, \rho^{(t)}, \tau^{(t)} \sim \mathcal{N}(\phi_j+b_0^{(t)}+b_1^{(t)}z_{i[j]}+V_j^{(t)}\boldsymbol{\beta}^{(t)}, \sigma^2\nu_{i[j]})\\ \text{independent} \qquad \forall j = 1, \dots, J[t], \forall t = 1, \dots, 5\end{aligned} \tag{3.2}$$

whereas, the corresponding prior structure is:

$$\begin{aligned}
&\nu_i \sim IG\left(\frac{r}{2},\frac{r}{2}\right), \quad \forall i = 1,\dots,14\\
&\sigma^2 \sim IG(0.001, 0.001)\\
&\boldsymbol{B}^{(t)} = (b_0^{(t)}, b_1^{(t)}\beta_1^{(t)},\dots,\beta_6^{(t)})\\
&\boldsymbol{B}^{(0)} \sim \mathcal{N}(\mathbf{0}, 10^2 I)\\
&\boldsymbol{B}^{(t)}|\boldsymbol{B}^{(t-1)},\Sigma_B \sim \mathcal{N}(\boldsymbol{B}^{(t-1)},\Sigma_B) \quad \forall t = 1,\dots,5\\
&\Sigma_B = diag\{\sigma^2_{b_0},\sigma^2_{b_1},\sigma^2_{\beta_1},\dots,\sigma^2_{\beta_6}\}\\
&\sigma_{b_0},\sigma_{b_1},\sigma_{\beta_1},\dots,\sigma_{\beta_6} \overset{\text{iid}}{\sim} U(0,10)\\
&\phi_j^{(t)}|\lambda^{(t)} \sim \mathcal{N}\left(0,\lambda^{(t)}\right) \quad \forall j = 1,\dots,J[t], \forall t = 1,\dots,5\\
&\lambda^{(t)} \overset{\text{iid}}{\sim} IG(0.001, 0.001) \quad \forall t = 1,\dots,5
\end{aligned}$$

We fixed the same regressors and the same structure for the regression coefficients and the variance.

Proper CAR model

The innovative contribution of this work is the study of neighbours influence at a census tract's level. In order to explore the spatial correlation, we introduce the contingency matrix $A^{(t)}$ such that

$$A^{(t)}_{i,j} = \begin{cases} \frac{1}{n_i^{(t)}} & \text{if the census tract } i \text{ and } j \text{ are closed}\\ 0 & \text{otherwise}\end{cases}$$

where $n_i^{(t)}$ is the number of neighbours of the $i-th$ census tract at time t. We introduce a random effect for each census tract, which evolves as a CAR model. The resulting Bayesian proper CAR model is the following:

$$\log(Y_j^{(t)})|V_j^{(t)},\boldsymbol{\beta}^{(t)},\nu_{i[j]},\sigma^2,\Sigma_\beta,\phi_j^{(t)},\rho^{(t)},\tau^{(t)} \sim \mathcal{N}(\phi_j + b_0^{(t)} + b_1^{(t)} z_{i[j]} + V_j^{(t)}\boldsymbol{\beta}^{(t)}, \sigma^2\nu_{i[j]})$$
$$\text{independent} \quad \forall j = 1,\dots,J[t], \forall t = 1,\dots,5 \tag{3.3}$$

$$\begin{aligned}
&\nu_i \sim IG\left(\frac{r}{2},\frac{r}{2}\right), \quad \forall i = 1,\dots,14\\
&\sigma^2 \sim IG(0.001, 0.001)\\
&\boldsymbol{B}^{(t)} = (b_0^{(t)}, b_1^{(t)}, \beta_1^{(t)}, \dots, \beta_6^{(t)})\\
&\boldsymbol{B}^{(0)} \sim \mathcal{N}(\mathbf{0}, 10^2 I)\\
&\boldsymbol{B}^{(t)}|\boldsymbol{B}^{(t-1)}, \Sigma_B \sim \mathcal{N}(\boldsymbol{B}^{(t-1)}, \Sigma_B) \quad \forall t = 1,\dots,5\\
&\Sigma_B = diag\{\sigma^2_{b_0}, \sigma^2_{b_1}, \sigma^2_{\beta_1}, \dots, \sigma^2_{\beta_6}\}\\
&\phi_j^{(t)}|\mathbf{\Phi}_{-j}^{(t)}, \tau^{(t)}, \rho^{(t)} \sim \mathcal{N}\left(\rho^{(t)}\boldsymbol{A}_j^{(t)}\mathbf{\Phi}^t, \frac{\tau^{(t)}}{n_j^{(t)}}\right) \quad \forall t = 1,\dots,5, \forall j = 1,\dots,J[t]
\end{aligned}$$

Moreover the specification for the hyperparameters is

$$\begin{aligned}
&\sigma^2_{\sigma^2_{b_0},\sigma^2_{b_1},\beta_1}, \dots, \sigma^2_{\beta_6} \overset{\text{iid}}{\sim} IG(0.001, 0.001)\\
&\rho^{(t)} \overset{\text{iid}}{\sim} U(0,1) \quad \forall t = 1,\dots,5\\
&\tau^{(t)} \overset{\text{iid}}{\sim} IG(0.5, 0.005) \quad \forall t = 1,\dots,5 \quad .
\end{aligned}$$

This choice is made in order to specify vague prior distribution and to have distributions coniugate to the model. We have uninformative and constant over time prior for all the variances, exept for $\tau^{(t)}$. As regard to the particolar choice of the prior of $\tau^{(t)}$, it is important to assess whether the prior allows for all reasonable levels of variability, in particular small values should not be excluded. As pointed out by Kelsall and Wakefield (1999), a prior $IG(0.001, 0.001)$ for the precision parameter of the spatial random effects in a CAR model, tends to place most of the prior mass away from zero (on the scale of the random effects standard deviation), and so in situations when the true spatial dependence between areas is negligible (i.e. standard deviation close to zero) this may induce artefactual spatial structure in the posterior. For this reason, following Kelsall and Wakefileld (1999), we set an $IG(0.5, 0.005)$ distribution for $\tau^{(t)}$. This expresses the prior belief that the random effects standard deviation is centred around 0.5 with a 1% prior probability of being smaller than 0.01 or larger than 2.5 (see Manual of GeoBUGS, 2014). Since we need to calculate the eigenvalues of the matrix $I - \rho^{(t)}A^{(t)}$ (see in the Appendix A the calculus for the posterior distribution of $\rho^{(t)}$), then we try to prove a relation with the eigenvalues of the matrix A, such that we do not need to calculate them in every iteration. Let $\lambda_j^{(t)}$ for $j = 1, \ldots, J[t]$ denote the eigenvalues of the matrix A. They are obtained by solving the equation

$$det(A^{(t)} - \lambda I) = 0 \quad .$$

In order to compute the eigenvalues $\mu_j^{(t)}$ of the matrix $I - \rho^{(t)} A(t)$ we set

$$\begin{aligned}
&det(I - \rho^{(t)} A^{(t)} - \mu I) = 0 \\
&det(-\rho^{(t)} A^{(t)} - (\mu - 1) I) = 0 \\
&(-\rho^{(t)})^{J[t]} \left(A^{(t)} - \frac{\mu - 1}{-\rho^{(t)}} I \right) = 0 \quad .
\end{aligned}$$

If $\rho^{(t)}$ is different from zero we can semplify and obtain the following relation between $\lambda_j^{(t)}$ and $\mu_j^{(t)}$:

$$\mu_j^{(t)} = 1 - \rho^{(t)} \lambda_j^{(t)} \quad \forall j = 1, \ldots, J[t] \quad .$$

By the way, to avoid to sample a zero value for $\rho^{(t)}$, we take in the interval of the prior distribution with respect to the one shown in the theorical dissertation. This fact does not imply any restriction in the analysis because, from the spatial indices (Moran's I and Geary's C), we expect a positive and low spatial correlation.

3.4 Computational strategy

In this section, we illustrate the MCMC strategy used to sample from the posterior distribution of the parameters. We refer in the following to the *proper CAR model*.

3.4.1 Gibbs Sampler

The posterior distribution of the *proper CAR model* is

$$\begin{aligned}
&\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(5)}, \sigma^2, \boldsymbol{\nu}, \boldsymbol{\Phi}^{(1)}, \dots, \boldsymbol{\Phi}^{(5)}, \Sigma_\beta, \tau^{(1)}, \dots, \tau(5), \dots \\
&\qquad\qquad\qquad\qquad \dots \rho^{(1)}, \dots, \rho^{(5)} | \boldsymbol{Y}^{(1)}, \dots, \boldsymbol{Y}^{(5)}, V^{(1)}, \dots, V^{(5)} \propto \\
&\propto \prod_{t=1}^{5} [\pi(\boldsymbol{Y}^{(t)} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\nu}, \sigma^2, \boldsymbol{\Phi}^{(t)}, V^{(t)})] \pi(\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(5)} | \boldsymbol{\beta}^{(0)}, \Sigma_\beta) \pi(\boldsymbol{\beta}^{(0)}) \pi(\Sigma_\beta) \pi(\sigma^2) \pi(\boldsymbol{\nu}) \times \\
&\quad \times \prod_{t=1}^{5} [\pi(\boldsymbol{\Phi}^{(1)}, \dots, \boldsymbol{\Phi}^{(5)} | \rho^{(t)}, \tau^{(t)}) \pi(\rho^{(t)}) \pi(\tau^t)] \propto \\
&\propto \prod_{t=1}^{5} [\pi(\boldsymbol{Y}^{(t)} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\nu}, \sigma^2, \boldsymbol{\Phi}^{(t)}, V^{(t)}) \pi(\boldsymbol{\Phi}^{(1)}, \dots, \boldsymbol{\Phi}^{(5)} | \rho^{(t)}, \tau^{(t)}) \pi(\rho^{(t)}) \pi(\tau^t) \pi(\boldsymbol{\beta}^{(t)} | \boldsymbol{\beta}^{(t-1)}, \Sigma_\beta)] \times \\
&\quad \times \pi(\boldsymbol{\nu}) \pi(\Sigma_\beta) \pi(\sigma^2) \pi(\boldsymbol{\beta}^{(0)}) \pi(\rho^{(t)}) \pi(\tau^{(t)}) \quad .
\end{aligned}$$

Because of the high number of parameters, it is quite complex to use standard R library like Stan or Jags, because they are very slow. We apply the Gibbs Sampler algorithm, that allows us to sample from the posterior distribution.
The basic idea of a Gibbs Sampler is to sample not directly from the posterior distribution, that usually has a complicated and unknown form, but to sample from the full conditional distributions of the parameters and update them one by one, using at each step the last sampled parameters. This type of algorithm results particularly efficient for this model because, as we said before, it is the best way to approach to a CAR model. In addition almost all the priors have a conjugated form at the full conditionals level, so, despite the huge number of parameters, the calculation of the full conditional distributions is feasible.

We illustate a general scheme of the algorithm, then we give the details of the full conditional distributions.

Algorithm 1 Gibbs sampler algorithm

```
Initialise the coefficients $\boldsymbol{\beta}^{(0)}$ randomly sampling from $\mathcal{N}(0,100)$
Initialise the coefficients $\boldsymbol{\beta_h} = \mathbf{0} \quad \forall h = 1,\dots,6$
Initialise the variance $\sigma^2_{\beta_h}$ randomly sampling from $\mathcal{U}(0,3) \quad \forall h = 1,\dots,6$
Initialise the variance $\sigma^2$ randomly sampling from $\mathcal{U}(0,3)$
Initialise the variance $\nu_i$ randomly sampling from $\chi^2(r) \quad \forall i = 1,\dots,14$
Initialise the random effect $\mathbf{\Phi}^{(t)} = \mathbf{1} \quad \forall t = 1,\dots,5;$
Initialise the variance $\tau^{(t)}$ randomly sampling from $\mathcal{U}(0,3) \quad \forall t = 1,\dots,5$
Initialise $\rho^{(t)}$ randomly sampling from $\mathcal{U}(0,1) \quad \forall t = 1,\dots,5$
update $\nu_i^{new} \leftarrow$ nu_upd($\boldsymbol{\beta}, \sigma^2, \mathbf{\Phi}, rest$) using (3.6) ;
update $(\sigma^2)^{new} \leftarrow$ sigma_upd($\boldsymbol{\beta}, \boldsymbol{\nu}^{new}, \mathbf{\Phi}, rest$) using (3.5) ;
update $(\sigma^2_{\beta_h})^{new} \leftarrow$ sigmabeta_upd($\boldsymbol{\beta_h}, rest$) using (3.4) ;
update $(\boldsymbol{\beta}_h^{(0)}, \boldsymbol{\beta}_h)^{new} \leftarrow$ beta_upd($\boldsymbol{\nu}^{new}, (\sigma^2)^{new}, (\sigma_{\beta_h})^{new}, rest$) using (3.1),(3.2) and (3.3);
update $(\tau^{(t)})^{new} \leftarrow$ tau _upd ($\mathbf{\Phi}^{(t)}, \rho^{(t)}, rest$) using (3.7) ;
update $(\rho^{(t)})^{new} \leftarrow$ rho_upd($\mathbf{\Phi}^{(t)}, (\tau^{(t)})^{new}, rest$) using (3.8);
update $(\mathbf{\Phi}^{(t)})^{new} \leftarrow$ phi_upd($\mathbf{\Phi}^{(t)}, (\tau^{(t)})^{new}, (\rho^{(t)})^{new}, rest$) using(3.9);
```

We now describe the full conditional distributions. For the calculation and more details, one can see Appendix A.

Update regression coefficients

The full conditional distributions of the $\boldsymbol{\beta}^{(t)}$ coefficients depend on the time. For $t = 0$, there is no dependence from the hyperparameters, so the full distribution is simply

$$\boldsymbol{\beta}^{(0)}|\boldsymbol{\beta}^{(1)}, \Sigma_\beta \propto \pi(\boldsymbol{\beta}^{(1)}|\boldsymbol{\beta}^{(0)}, \Sigma_\beta)\pi(\boldsymbol{\beta}^{(0)}) \propto$$
$$\propto exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)})'\Sigma_\beta^{-1}(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)})\right\} exp\left\{-\frac{1}{200}(\boldsymbol{\beta}^{(0)})'\boldsymbol{\beta}^{(0)}\right\} \quad .$$

Since the covariance matrix is diagonal, the components of the multinormal vector are independent and we can factorize the distribution. In this way we can sample from an onedimensional normal distribution. This happens at all time t, so in the next cases we will not repeat the argumentation. Hence the full conditional distribution of each $\beta_h^{(0)}$ for $h = 1, \ldots, 6$ takes form

$$\beta_h^{(0)} | \beta_h^{(1)}, \sigma^2_{\beta_h} \sim \mathcal{N}\left(\frac{100}{100 + \sigma^2_{\beta_h}} \beta_h^{(1)}, \frac{100\sigma^2_{\beta_h}}{100 + \sigma^2_{\beta_h}}\right) \quad h = 1, \ldots, 6. \tag{3.4}$$

For $t = 1, \ldots, 4$, we set

$$a = \frac{1}{\sigma^2} \sum_{j=1}^{J[t]} \frac{V_{i,j}^{(t)}(y_j^{(t)} - \sum\limits_{k \neq h} V_k^{(t)} \beta_k^{(t)} - \phi_j^{(t)})}{\nu_{i[j]}} + \frac{1}{\sigma^2_{\beta_h}}(\beta_h^{(t+1)} + \beta_h^{(t-1)})$$

$$b = \frac{1}{\sigma^2} \sum_{j=1}^{J[t]} \frac{(V_{i,j}^{(t)})^2}{\nu_{i[j]}} + \frac{2}{\sigma^2_{\beta_h}}$$

and we find

$$\beta_h^{(t)} | \beta_h^{(t-1)}, \beta_h^{(t+1)}, \mathbf{\Phi}^{(t)}, \sigma^2_{\beta_i}, \boldsymbol{\nu}, \boldsymbol{Y^{(t)}}, V^{(t)}, \sigma^2 \sim \mathcal{N}\left(b^{-1}a, b^{-1}\right) \quad . \tag{3.5}$$

At $t = 5$ the distribution is almost equal to (3.2), the only difference is that here there is not dipendence on the future state, and hence a and b become:

$$a = \frac{1}{\sigma^2} \sum_{j=1}^{J[t]} \frac{V_{i,j}^{(t)}(y_j^{(t)} - \sum\limits_{k \neq h} V_k^{(t)} \beta_k^{(t)} - \phi_j^{(t)})}{\nu_{i[j]}} + \frac{\beta_h^{(t-1)}}{\sigma^2_{\beta_h}}$$

$$b = \frac{1}{\sigma^2} \sum_{j=1}^{J[t]} \frac{(V_{i,j}^{(t)})^2}{\nu_{i[j]}} + \frac{1}{\sigma^2_{\beta_h}}$$

$$\beta_h^{(t)} | \beta_h^{(t-1)}, \mathbf{\Phi}^{(t)}, \sigma^2_{\beta_h}, \boldsymbol{\nu}, \boldsymbol{Y^{(t)}}, V^{(t)}, \sigma^2 \sim \mathcal{N}\left(b^{-1}a, b^{-1}\right) \quad . \tag{3.6}$$

Update σ^2_β

The choice of a suitable prior distribution allows us to find a conjugate full conditional. In fact it is straightforward to prove that

$$\sigma^2_{\beta_h} | \beta_h^{(1)}, \ldots, \beta_h^{(5)} \sim IG\left(a + \frac{5}{2}, b + \frac{\sum\limits_{t=1}^{5} (\beta_h^{(t)} - \beta_h^{(t-1)})^2}{2}\right) \qquad h = 1, \ldots, 6 \quad . \tag{3.7}$$

with $a = b = 0.001$.

Update σ^2

Let $N^{(t)} \in \mathbb{R}^{J[t]\times J[t]}$ be a diagonal matrix such that $N^{(t)}_{j,j} = \nu_{i[j]}\ \forall t\ \forall t$. Then, even σ^2 has a conjugate prior. Set $a = b = 0.001$ and

$$f^{(t)} = (\boldsymbol{Y}^{(t)} - V^{(t)}\boldsymbol{\beta}^{(t)} - \boldsymbol{\Phi}^{(t)})'(N^{(t)})^{-1}(\boldsymbol{Y}^{(t)} - V^{(t)}\boldsymbol{\beta}^{(t)} - \boldsymbol{\Phi}^{(t)})$$

we obtain

$$\sigma^2|\boldsymbol{\beta}^{(1)},\ldots,\boldsymbol{\beta}^{(5)},\boldsymbol{\nu},\boldsymbol{\Phi}^{(1)},\ldots,\boldsymbol{\Phi}^{(5)},\boldsymbol{Y}^{(1)},\ldots,\boldsymbol{Y}^{(5)},V^{(1)},\ldots,V^{(5)} \sim$$
$$\sim IG\left(a + \frac{\sum_{t=1}^{5} J[t]}{2}, b + \frac{\sum_{t=1}^{5} f^{(t)}}{2}\right) \quad . \tag{3.8}$$

Update ν

The variable ν_i depends only on those census tracts that belong to $i - th$ county, we partition the dataset as $\boldsymbol{Y}_i^{(t)}, V_i^{(t)}, \boldsymbol{\Phi}_i^{(t)}$ such that they are data and parameters relative to county i . Set

$$k^{(t)} = \frac{(\boldsymbol{Y}_i^{(t)} - V_i^{(t)}\boldsymbol{\beta}^{(t)} - \boldsymbol{\Phi}_i^{(t)})'(\boldsymbol{Y}_i^{(t)} - V_i^{(t)}\boldsymbol{\beta}^{(t)} - \boldsymbol{\Phi}_i^{(t)})}{2\sigma^2}$$

and for $i = 1,\ldots,14$ we obtain

$$\nu_i|\boldsymbol{\beta}^{(1)},\ldots,\boldsymbol{\beta}^{(5)},\sigma^2,\boldsymbol{\Phi}_i^{(1)},\ldots,\boldsymbol{\Phi}_i^{(5)},\boldsymbol{Y}_i^{(1)},\ldots,\boldsymbol{Y}_i^{(5)},V_i^{(1)},\ldots,V_i^{(5)} \sim$$
$$\sim IG\left(\frac{r}{2} + \frac{\sum_{t=1}^{5} n_i^{(t)}}{2}, \frac{r}{2} + \sum_{i=1}^{5} k^{(t)}\right) \quad . \tag{3.9}$$

Update $\tau^{(t)}$

Set $a = 0.5$ and $b = 0.0005$, the full conditional distribution for $\tau^{(t)}$ is

$$\tau^{(t)}|\boldsymbol{\Phi}^{(t)},\rho^{(t)} \sim IG\left(a + \frac{J[t]}{2}, b + \frac{(\boldsymbol{\Phi}^{(t)})'(D_w^{(t)} - \rho^{(t)}W^{(t)})\boldsymbol{\Phi}^{(t)}}{2}\right) \tag{3.10}$$

where $D_w^{(t)}$ is a diagonal matrix with $(D_w^{(t)})_{j,j} = n_j^{(t)}$ and

$$W_{i,j}^{(t)} = \begin{cases} 1 & \text{if the census tract } i \text{ and } j \text{ are closed} \\ 0 & \text{otherwise} \end{cases}$$

Update $\rho^{(t)}$

The full conditional distribution for $\rho^{(t)}$ takes form

$$\begin{aligned} \rho^{(t)} | \mathbf{\Phi}^{(t)}, \tau^{(t)} \sim \left(\prod_{j=1}^{J[t]} (1 - \rho^{(t)} \lambda_j^{(t)}) \right)^{1/2} & \times \\ & \times \, exp \left\{ \frac{1}{2\tau^{(t)}} \rho^{(t)} (\mathbf{\Phi}^{(t)})' W^{(t)} \mathbf{\Phi}^{(t)} \right\} \mathbb{I}_{[0,1]}(\rho^{(t)}) \end{aligned} \tag{3.11}$$

where $\lambda_j^{(t)}$ are the eigenvalues of $A^{(t)}$. This distribution is unknown, therefore it is necessary to use a Metropolis step within the Gibbs sampler. Since ρ is defined on a (0,1) domain, a convenient strategy is to choose a uniform proposal $\mathcal{U}(0,1)$ and implement an independent random walk.

Algorithm 2 Metropolis-Hastings algorithm

```
At $k - th$ iteration , $\rho^{(t)} = r$
Set $f(x) = \left( \prod_{j=1}^{J[t]} (1 - x\lambda_j^{(t)}) \right)^{1/2} exp\left\{ \frac{1}{2\tau^{(t)}} x (\mathbf{\Phi}^{(t)})' D^{-1} A^{(t)} \mathbf{\Phi}^{(t)} \right\}$
Sample from the proposal $\delta \sim U(0,1)$
Compute the acceptance rate $\alpha = \frac{f(\delta)}{f(r)}$
Sample a probability $p \sim U(0,1)$
```

Update random effect parameters

Using the theorical results for the CAR model, we can update the random effects simply using the following full conditional distribution

$$\phi_j^{(t)} | \mathbf{\Phi}_{-j}^{(t)}, y_j^{(t)}, V_j^{(t)}, \nu_{i[j]}, \sigma^2, \boldsymbol{\beta}^{(t)}, \tau^{(t)}, \rho^{(t)} \sim \mathcal{N}(b^{-1}a, b^{-1}) \quad j = 1, \dots, J[t]$$

where

$$\begin{aligned} b &= \frac{1}{\sigma^2 \nu_{i[j]}} + \frac{n_j^{(t)}}{\tau^{(t)}} \\ a &= \frac{y_j^{(t)} - V_j^{(t)} \boldsymbol{\beta}^{(t)}}{\sigma^2 \nu_{i[j]}} + \frac{n_j^{(t)} \rho^{(t)} A_j^{(t)} \boldsymbol{\Phi}^{(t)}}{\tau^{(t)}} \quad . \end{aligned}$$

3.4.2 Implementation with Stan and Julia

We here justify the choice of the programming language used in this work. The first and the second model, i.e. the baseline model and the model with independent random effects, have been implemented in Stan. Stan is a package in R language that translate the code in C++. It is really useful because one has to specify only the prior distribution and the likelihood for the model, then the calculation is automatically done. This library are really intuitive to use, unfortunatly the calculation becomes very slow when there is a huge number of parameters and when there are a lot of matrix operations to do, especially for product. So models like our proper CAR are computationally unfeasible. In addition sampling from a multinormal distribution of such dimension has a high computational heaviness. R programming routines encourage operating on whole objects (i.e. vectorised code) because *while* and *for* loops are notoriously slow. Nevertheless, MCMC are not easily vectorised as every iteration depends on the previous one. Therefore we implement the code in the Julia language, that manages to combine computational efficiency with the easy scripting and interpretation typical of any other high-level programmming language. Julia language provides an extensive mathematical function library, in paticular random number generator libraries that are fundamental for a Bayesian analysis. In addition, in Julia objects are passed and assigned by reference, this allows the algorithm to reduce the memory usage and makes the algorithm faster, especially in MCMC computing where one deals with big matrices at every iterartion.

The baseline model and the model with independent random effects were implemented in Stan. Posterior estimations for the proper CAR model are computed via the Gibbs sampler algorithm described in Section 3.4, and the algorithm ran in Julia. In all cases, we compute 65000 iterations, while the first 15000 iterations were discarded, with a thinning of 10 to reduce autocorrelation of the Markov chain. The final sample size is then 5000. The codes are explained in detail in Appendix B.

Chapter 4

Application to Massachusettes census tracts data

In this chapter, we present the inference on the three models and we analyze the most important variables. Analysis and diagnostics of convergence of the MCMC have been done; we do not illustrate all the results of convergence because of the huge amount of parameters, among the others we show in Appendix D the diagnostic results for the distance's coefficient because it is the most relevant predictor.

4.1 Posterior inference on the regression coefficients

First of all, let us show which covariates result to be significative in leading the individual location choice. All the three models give results consistent with the observations of Section 3.2.
Basically, as we expected, the distance from Boston leads the population distribution. In a state where the most part of the economy develops in services, people detect fundamental to settle near to the biggest city, especially near to Boston, that is the centre of services and economy, i.e. the CBD. It is not a case that the counties that mainly increase in the periods under investigation are those near to Suffolk, e.g. Middlesex and Worcester, which allow for good services and fast connection with Boston.

Excluding the census of 1970, the population density decreases with the growth of

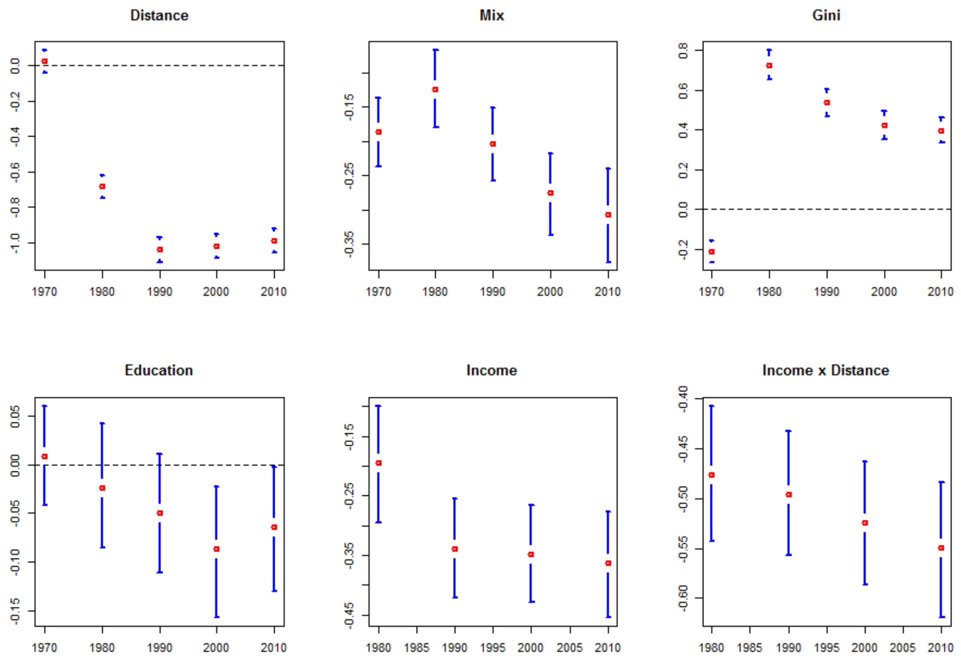

Figure 4.1: Credibility intervals for the regression coefficients at level 90%*, under the baseline model.*

distance. It is worth to notice that, for almost all the covariates, in 1970 and 1980 a different behavior appears in the estimations with respect to most recent years. As we explained in the descriprion of the dataset, that is because the second but especially the first census were collected with different criterions; one can notice that even for other variables in 1970 there is an evolution in opposition to the other times. Referring to Figures 4.1, 4.2 and 4.3 (the estimation and the quantiles for the regressors $\beta_h^{(t)}$ can be found in Appendix C), we can state that modelling the individual choice only considering the distance it would be restrictive. Also the ethnic composition plays a relevant role. People prefer to move where there is a more proportion of white inhabitants, that should guarantee the presence of wealthier people and a lower rate of criminality. Even the economic indicator results to be important. It turns out that richer people, with an high income, prefer to live far from the city, where there is a lower population density; in this way they can own big properties. Instead, if the income is equally distributed, the density of population increases, as people place where there is a diffused wellness. The education

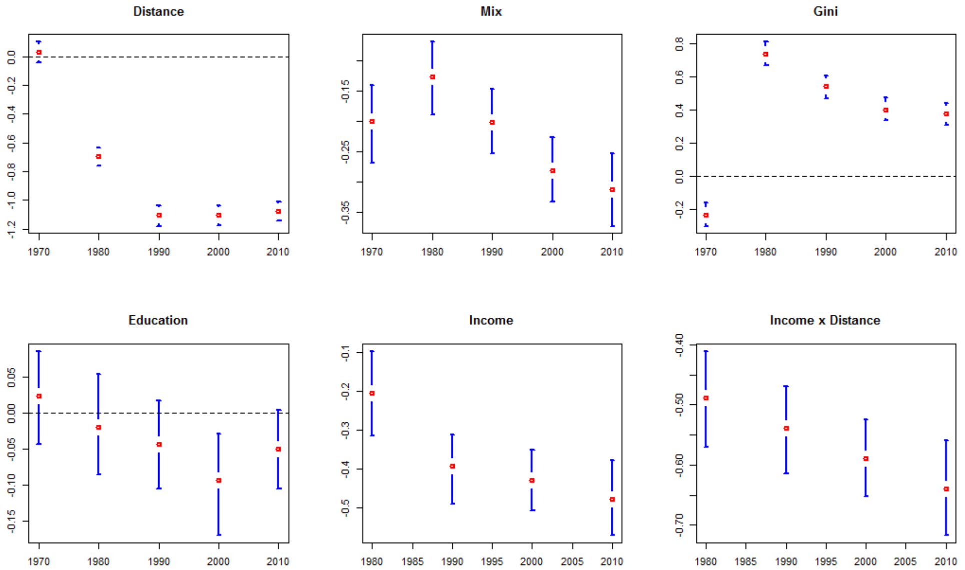

Figure 4.2: Credibility intervals for the regression coefficients at level 90%*, under the independent random effects model.*

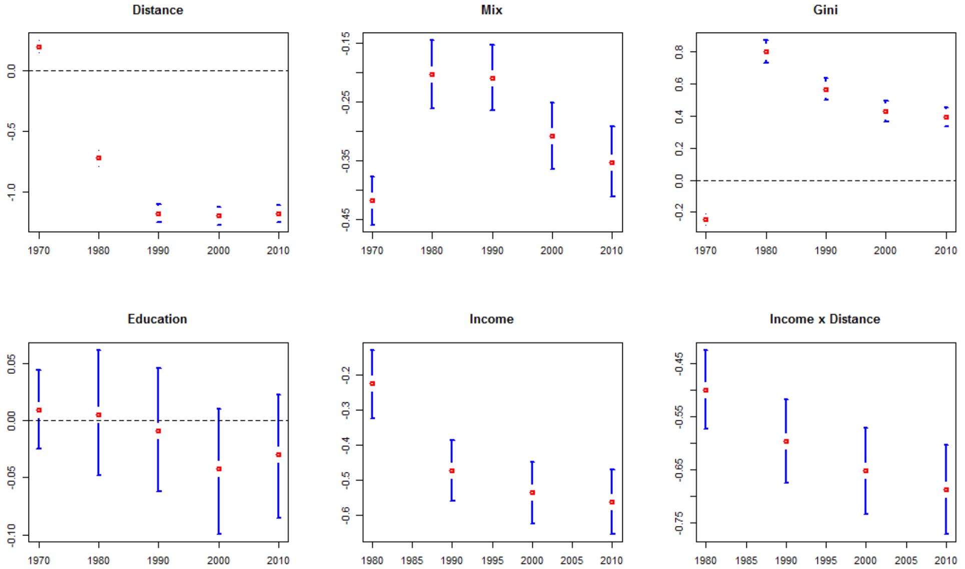

Figure 4.3: Credibility intervals for the regression coefficients at level 90%*, under the CAR model.*

level becomes important only in more recent time. It is interesting to notice that after 1990 the importance of the distance and the income influence remain almost constant, while the mix and the education factors become more and more relevant in a negative way. On the other hand the Gini index looses influence over time. These paths of the coefficients can be explained as an increase of the importance of the reputation effects: people are more and more interested in the feature of the cities they have selected.
Since these estimation are quite similar in all the three models, we can conclude that the introduction of spatial effects does not affect the dynamic trends of the regression coefficients.

4.2 Autocorrelation and heteroschedasticity

The study of spatial correlation is divided in two levels: the effect of belonging to a county and the effect of the neighbourhood.

The county effect translates into two aspects: the regressor given by the amenities and the variance's structure. The estimations for the amenities' coefficients is reported in Table 4.1. One can notice that, except for the first year, the amenities are pratically non influencial neither in the baseline model nor in the independent random effects. Instead, in the CAR model, $b_1^{(t)}$ is significative in a negative way. This is a logical behavior since, if the amenities are higher it means that the proportion of water area is extended and hence the population density would be lower.

	1970	1980	1990	2000	2010
Baseline model	**0.43**	**0.12**	0.02	**0.06**	**0.11**
	(0.38,0.49)	(0.07,0.18)	(-0.04,0.07)	(0.00,0.11)	(0.06,0.16)
Indipendent model	**0.43**	**0.12**	0,00	0.03	**0.09**
	(0.35,0.49)	(0.06,0.17)	(-0.05,0.05)	(-0.01,0.08)	(0.05,0.15)
CAR model	$\mathbf{-0.31}$	$\mathbf{-0.29}$	$\mathbf{-0.33}$	$\mathbf{-0.34}$	$\mathbf{-0.34}$
	(-0.39,-0.22)	(-0.37,-0.18)	(-0.41,-0.25)	(-0.43,-0.27)	(-0.42,-0.25)

Table 4.1: Estimation of the coefficient of the amenities β_1 in the three models. We show the mean (it is highligth if the coefficient results significative) and the 2.5% and 97.5% quantiles.

The variances $\sigma^2\nu$ of the log densities of population turn out to be different each others. This variance structure really allows to capture the heteroschedasticity of the data: the values obtained are different for each county, we can deduce that there is a strong county effect.

	Baseline model	Indipendent model	CAR model
Barnstable(1)	3.53	3.48	3.23
Hampshire(2)	10.07	8.26	6.41
Berkshire(3)	11.39	11.18	10.89
Middlesex(4)	3.82	1.94	1.17
Bristol(5)	5.25	3.39	2.56
Nantucket(6)	12.13	14.76	21.04
Dukes(7)	5.71	3.59	1.30
Norfolk(8)	2.63	1.35	0.81
Essex(9)	2.94	1.49	1.08
Plymouth(10)	4.87	3.36	1.86
Franklin(11)	9.51	8.30	5.57
Suffolk(12)	3.22	2.28	2.21
Hampden(13)	7.62	7.25	7.33
Worcester(14)	7.50	5.31	3.71

Table 4.2: Estimation of the variances $\sigma^2\nu$ of the population densities under the three alternative models.

The estimations obtained in the three different cases are perfectly consistent to each others. As the estimations in the CAR model were not precise, then we correct them by means of the other models estimations. We can divide the counties in two main groups. One one side, we have all these counties with a large number of census tracts, like Middlesex or Suffolk; in this case the variances are low, and the corresponding intervals of credibility is narrow (see Figures 4.4, 4.5 and 4.6). Actually in a very populous county there is a homogeneous population distribution, so there are no big differences. On the other hand, we expect that in a county with a little number of census tracts there would be more differences inside the county, due to the extended area of a census tract that includes different situations. We find validation to our observation in the obtained estimations, indeed, counties as Dukes or Franklin have higher variability than to the other counties. Even though the behaviors of the variances are quite similar, anyway, adding spatial effects make the estimation more precise (see Table 4.2).

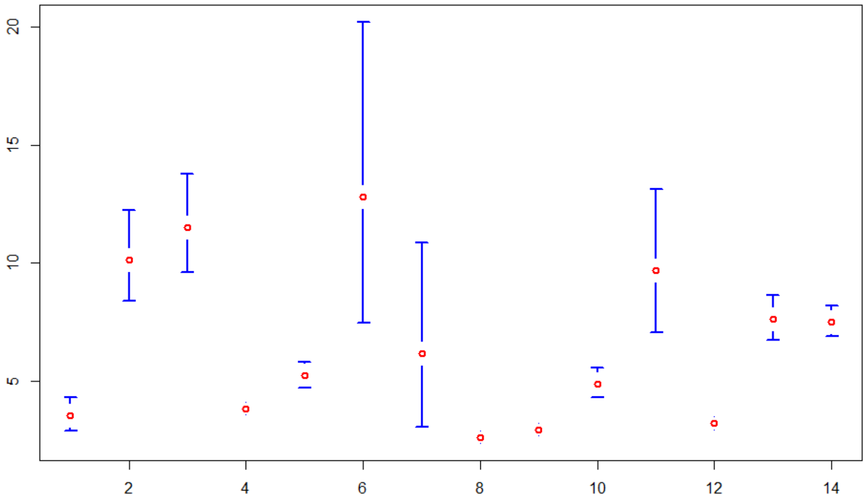

Figure 4.4: Credibility intervals of the variances of the density population Y on log scale at level 90%, under the baseline model.

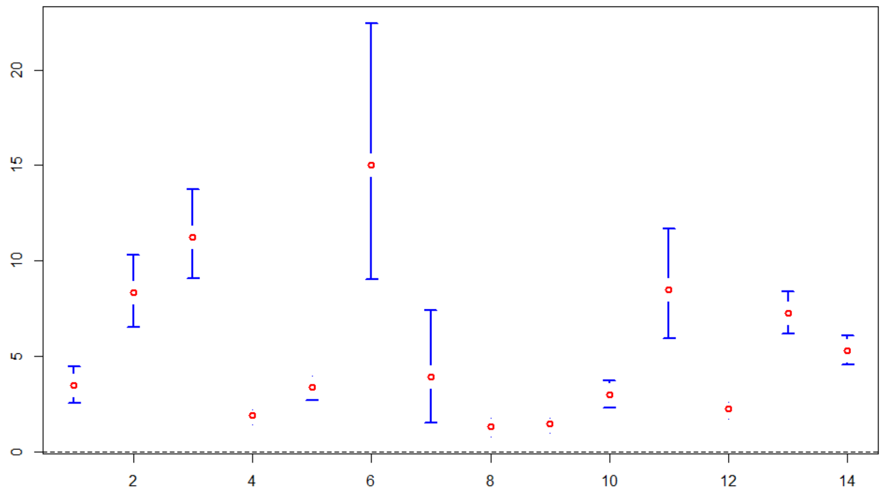

Figure 4.5: Credibility intervals of the variances of the density population Y on log scale at level 90%, under the random independent effects model.

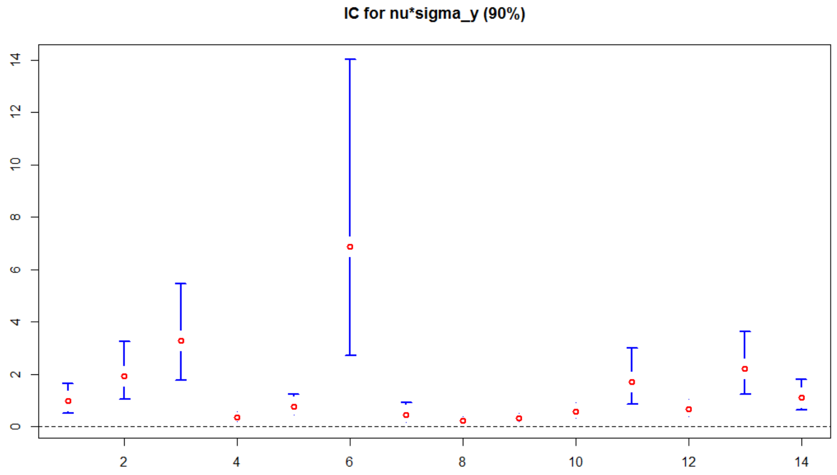

Figure 4.6: Credibility intervals of the variances of the density population Y on log scale at level 90%, under the CAR model.

Let us now investigate the local spatial correlation. The maps in Figures 4.7-4.11 show the dynamic evolution of the spatial independent random effects. One can notice that in the first two periods the random effects are significant, while, time after time, the estimations of the random effects approach zero for every census tracts. We expected this behavior because the number of census tracts is increasing over time, whereas as one can see in Figure 3.6 the density of a single census tract keeps almost constant, but the area becomes smaller over time. This lead to a homogeneus behavior of the population density in the single unit, the variability is all described by the covariates and there is no need of additional effects to interpretate the phenomenon. Concerning the census tracts in the country area, we notice that in the past the correspponding spatial independent random effects are significantly negative, whereas in more recent time they become null. This fact can be due to the improvement of transportation during the years. so that people can easily commute to Boston. On the other, the random effect on the census tracts in the big cities tend to be positive. The differnt sign of the random effects in the countryside with respect to the urban areas shows that actually there is a relevant part of the variability of the population density unexplained by the covariate.

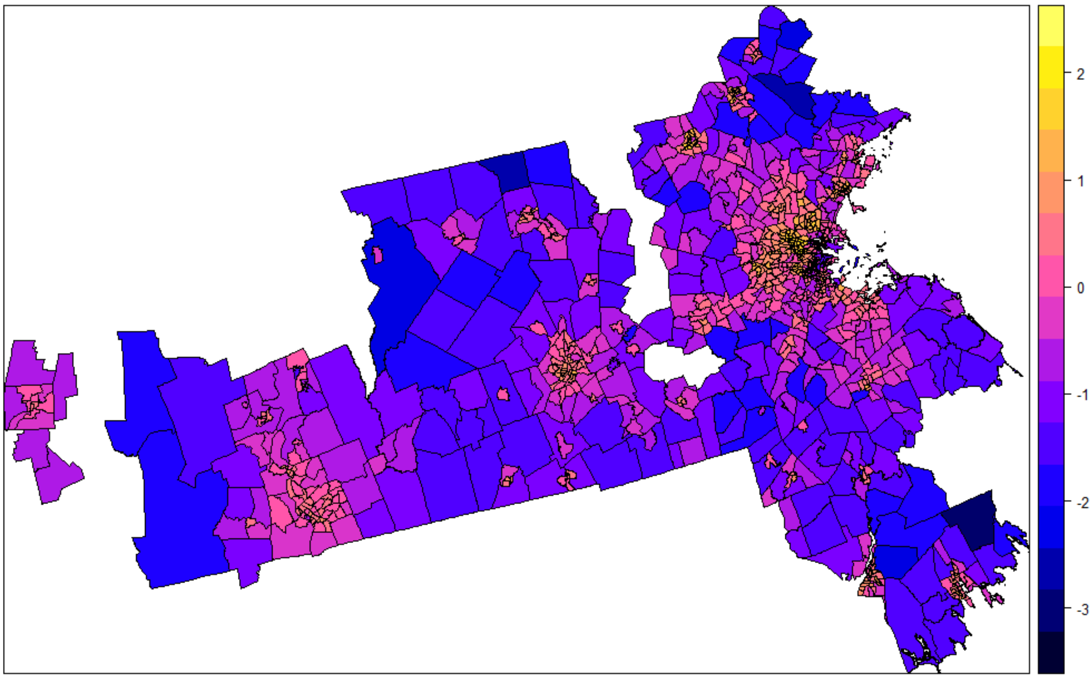

Figure 4.7: Estimation of the independent spatial random effects in 1970.

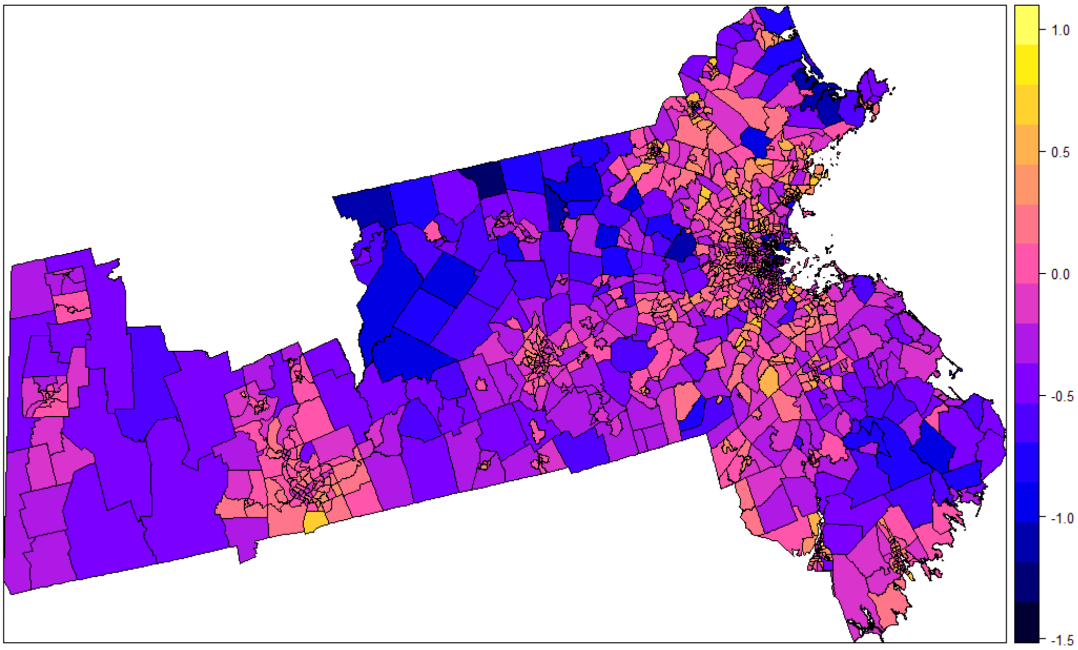

Figure 4.8: Estimation of the independent spatial random effects in 1980.

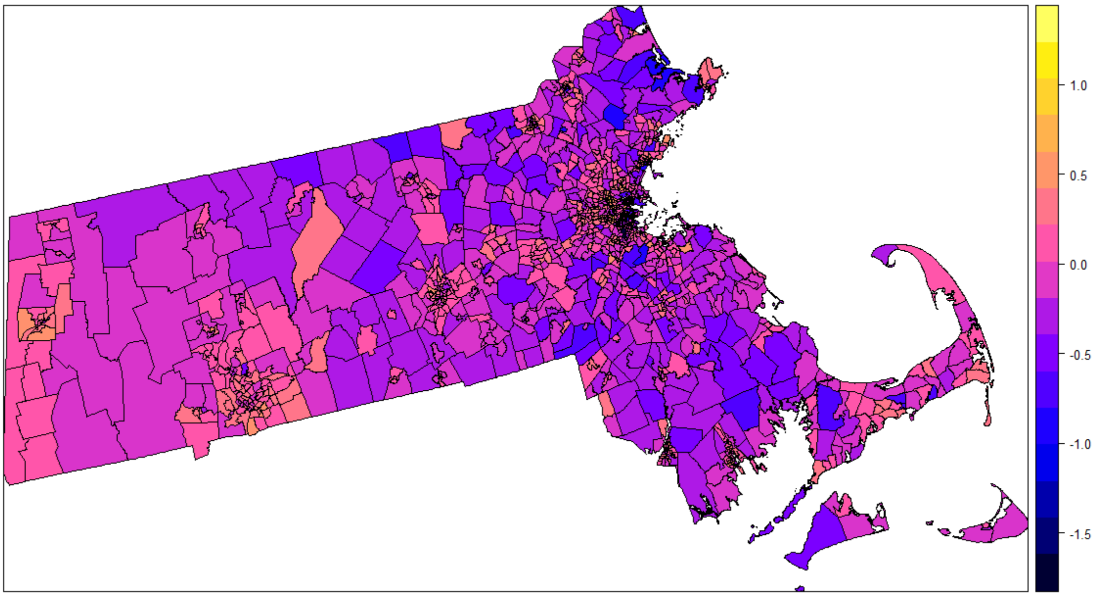

Figure 4.9: Estimation of the indipendt spatial random effects in 1990.

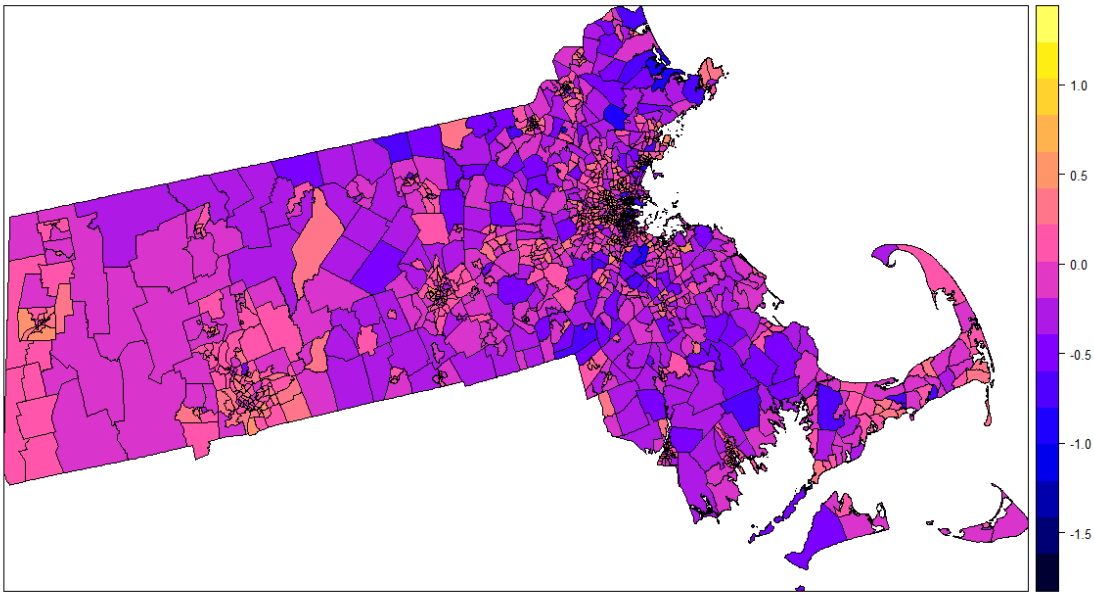

Figure 4.10: Estimation of the indipendt spatial random effects in 2000.

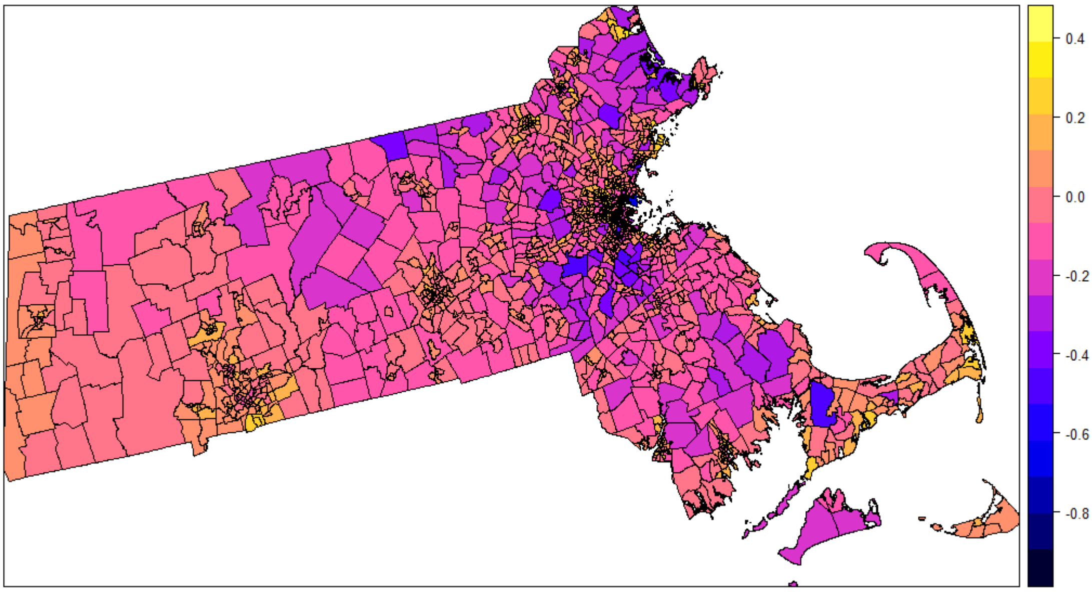

Figure 4.11: Estimation of the indipendt spatial random effects in 2010.

	1970	1980	1990	2000	2010
ρ	**0.99**	**0.72**	**0.57**	**0.50**	**0.53**
	(0.99,0.99)	(0.57,0.84)	(0.37,0.72)	(0.27,0.70)	(0.30,0.72)

Table 4.3: Estimation of ρ . We show the mean and the (2.5%,97.5%) quantiles.

From the analysis of the CAR model results, we know that there is spatial correlation among the census tracts. Referring to Table 4.3, the estimations of ρ represent the strenght of the autocorrelation, hence in 1970 and 1980 the spatial dependence is stronger than in the other periods. Anyway there is always positive spatial dependence among data. Maps in Figures 4.12-4.16 show the dynamic evolution of the random effects. Under the CAR specification on the random effects, the interpretation of the estimated random effects is not as clear as in the independent case. For example in 1990s $\mathbf{\Phi}^{(t)}$ are almost null, whereas 20 years later in 2010 they become almost all negative as one can see in Figures 4.13-4.14. Some furter investigations are required.

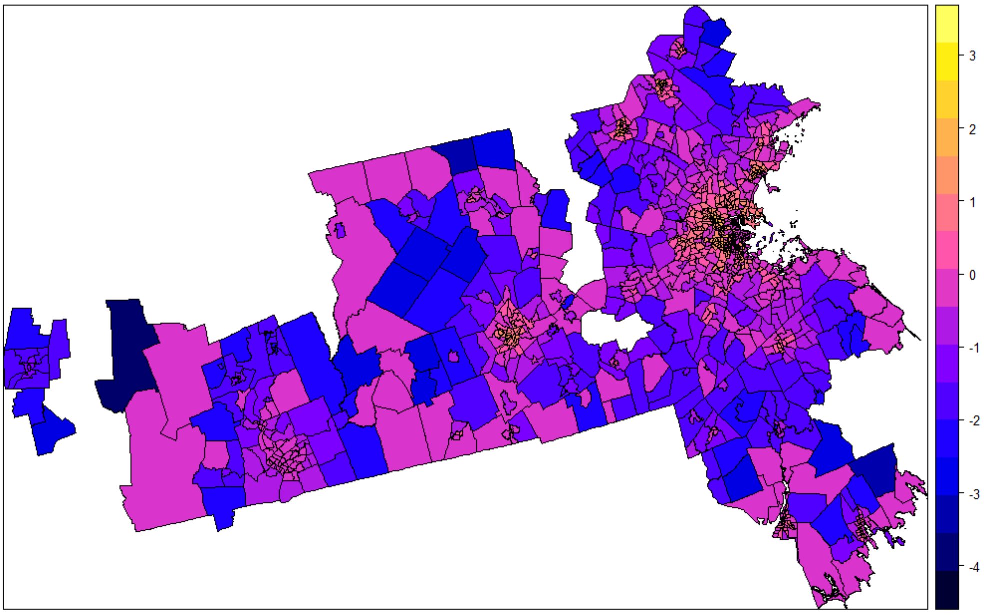

Figure 4.12: Estimation of the spatial random effects under the CAR model in 1970.

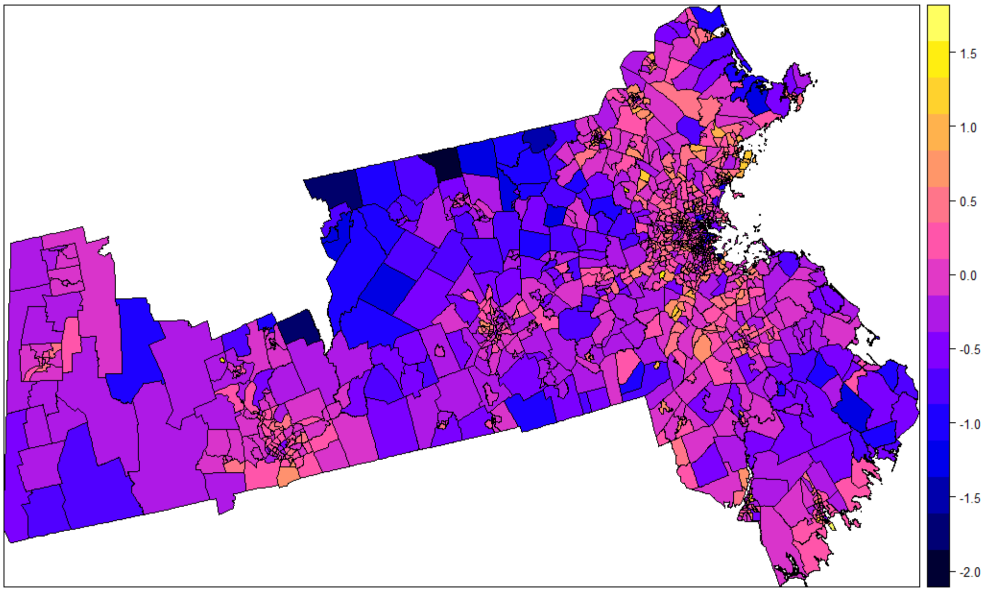

Figure 4.13: Estimation of the spatial random effects under the CAR model in 1980.

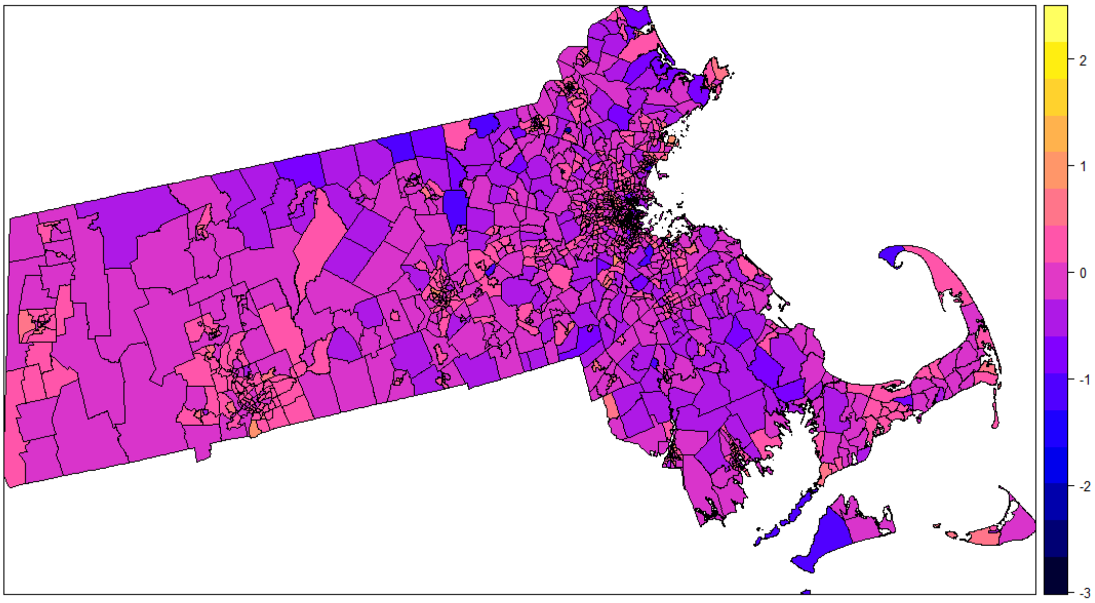

Figure 4.14: Estimation of spatial random effects under the CAR model in 1990.

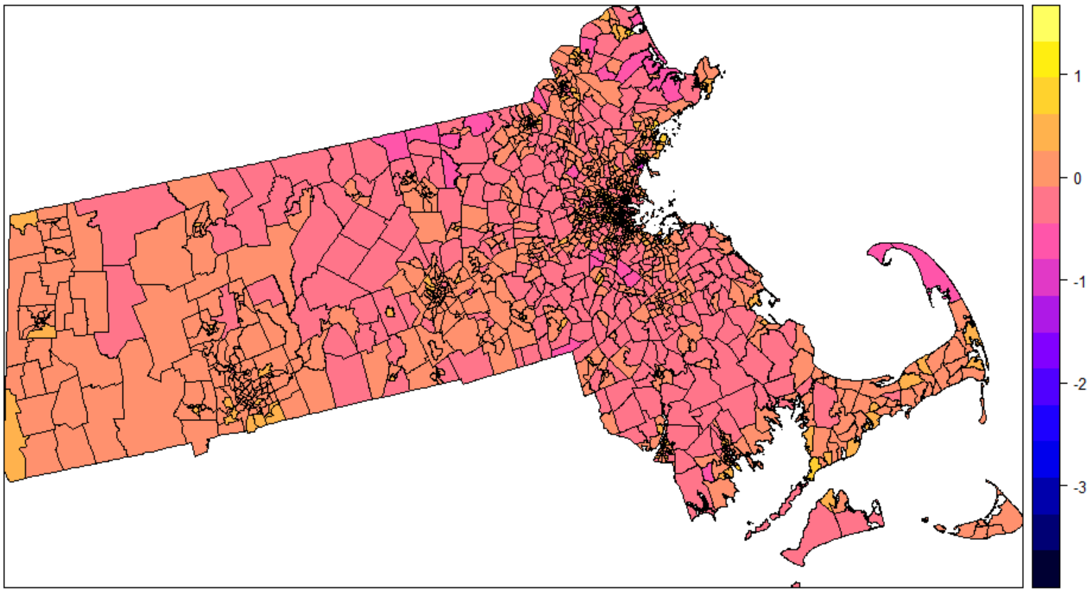

Figure 4.15: Estimation of spatial random effects under the CAR model in 2000.

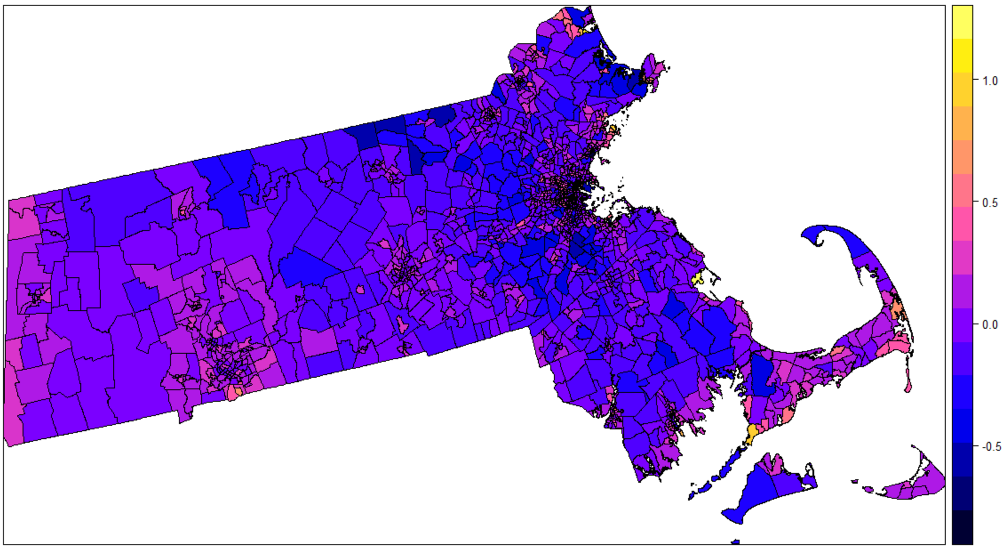

Figure 4.16: Estimation of the spatial random effects under the CAR model in 2010.

Finally, let us explore the estimations of the varinces of the random effects under both independent and CAR random effects models. Using the formula for the variance decomposition:

$$Var(log(Y_j^{(t)})) = Var(E[log(Y_j^{(t)})|\phi_j^{(t)}]) + E[Var(log(Y_j^{(t)})|\phi_j^{(t)})] \quad ,$$

one can show that

$$Var(Y_j^{(t)}|\beta_j^{(t)}, \rho^{(t)}, \tau^{(t)}, \sigma^2, \nu_{i[j]}) \propto c_j \tau^{(t)} + \sigma^2 \nu_{i[j]}$$

where

$$c_j \propto \begin{cases} 1 & \text{under independent random model for } \phi_j^{(t)} \\ \frac{1}{n_j^{(t)}} & \text{under CAR model for } \phi_j^{(t)} \end{cases} \quad .$$

Hence, the variance of the random effects influences the variance of the population density. Both $\lambda^{(t)}$ and $\tau^{(t)}$ show the same trend, the values estimated decreasing over time. All in all, the idea that emerges from the analysis is that over time there is a progressive homogeneity in population distribution. Census tracts effects are going to zero, and the population density is almost constantly distributed in the countryside (see Figure 3.6).

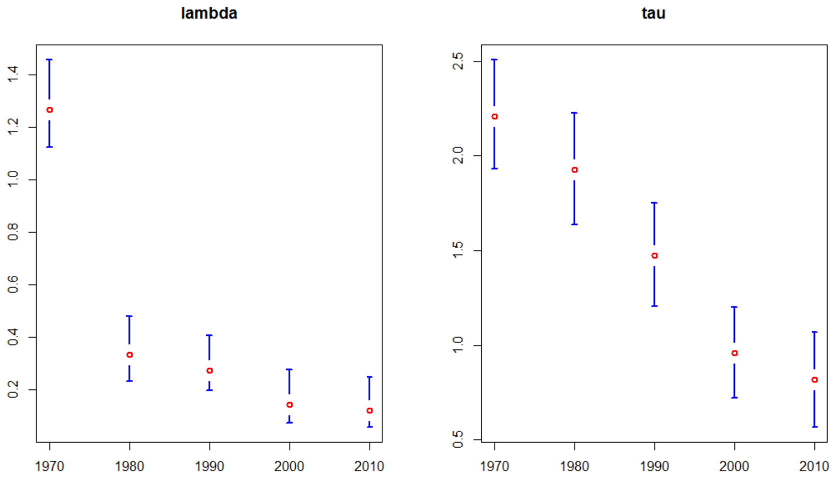

Figure 4.17: Credibility intervals of the common term $\lambda^{(t)}$ in the variances of the spatial random effects model at level 90%, under the random independent effects model. Credibility intervals of the common term $\tau^{(t)}$ in the variances of the spatial random effects at level 90%, under the random CAR model.

4.3 Models comparison

In this section we compare the results obtain in the three models. First of all an index for the goodness-of-fit of the model has to be introduced. In this work, we choosed the *Log Pseudo Marginal Likelihood* (LPML) in order to evaluate the performances of the models and compare them. The LPML is defined as the sum of the logarithms of the *Conditional Predictive Ordinates* (CPO) for each observation, i.e.

$$LPML = \sum_{i=1}^{N} log(CPO_i) \quad ,$$

where CPO_i is the value of the predictive distribution evaluated at y_i, conditional to the training sample y_{-i} not containing the i-th observation. This approach is very common in the cross validation techniques, when the data matrix is partitioned in two parts: one is used to estimate the parameters, and the other to measure the goodness of fit. Obviously, the larger the value of the CPO (and, subsequently, of the LPML) is, the better the model prevision is. The calculation of LPML consists in the evaluation of n predictive distributions, which can be computationally heavy. However, an alternative formula can be proved for $CPO_i = f_i(y_i|y_{-i})$. In fact

$$\begin{aligned} CPO_i = f_i(y_i|y_{-i}) &= \int_\Theta f_i(y_i|\boldsymbol{\theta})\mathcal{L}(d\boldsymbol{\theta}|y_{-i}) = \\ &= \int_\Theta f_i(y_i|\boldsymbol{\theta}) \frac{\prod\limits_{j\neq i} f_j(y_j|\boldsymbol{\theta})\mathcal{L}(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_\Theta \prod\limits_{j\neq i} f_j(y_j|\boldsymbol{\theta})\mathcal{L}(\boldsymbol{\theta})d\boldsymbol{\theta}} \quad , \end{aligned}$$

and using the Bayes' theorem we obtain:

$$\begin{aligned} CPO_i^{-1} &= \frac{\int_\Theta \prod\limits_{j\neq i} f_j(y_j|\boldsymbol{\theta})\mathcal{L}(d\boldsymbol{\theta})}{\int_\Theta \prod\limits_{i=1}^{n} f_i(y_i|\boldsymbol{\theta})\mathcal{L}(d\boldsymbol{\theta})} = \\ &= \int_\Theta \frac{1}{f_i(y_i|\boldsymbol{\theta})} \frac{\prod\limits_{i=1}^{n} f_i(y_i|\boldsymbol{\theta})\mathcal{L}(d\boldsymbol{\theta})}{\int\limits_\Theta \prod\limits_{i=1}^{n} f_i(y_i|\boldsymbol{\theta})\mathcal{L}(d\boldsymbol{\theta})} = \int\limits_\Theta \frac{1}{f_i(y_i|\boldsymbol{\theta})}\mathcal{L}(d\boldsymbol{\theta}|\boldsymbol{y}) \quad . \end{aligned}$$

In the light of the ergodic theorem, the las integral can be approximated by:

$$CPO_i^{-1} = \frac{1}{M}\sum_{m=1}^{M} \frac{1}{f_i(y_i|\boldsymbol{\theta}^{(m)})} \quad ,$$

where M is the number of iterations and $\theta^{(m)}$ is the value of the chain at iteration m. The estimated values of the LPML index for every model are listed in Table 4.4.

	1970	1980	1990	2000	2010	TOT
Baseline model	-2258	-2221	-2335	-2394	-2599	-11807
Indipendent model	-1908	-1951	$\mathbf{-2050}$	$\mathbf{-2114}$	$\mathbf{-2302}$	-10325
CAR model	$\mathbf{-1745}$	$\mathbf{-1816}$	-2206	-2150	-2311	$\mathbf{-10228}$

Table 4.4: LPML values for every years.

According to LPML criterion, both the models with spatial effects perform better than the baseline model, and they have similar performance since the values are almost the same.

As further measure of goodness of fit and comparison of the models, we calculate the outliers of the models. The outliers are determined by means of the bayesian predictive p-values. In a Bayesian context, the posterior (or predictive) p-value is the probability, given the data, that a future observation is more extreme than the data. Mathematically, a Bayesian p-value can be computed by averaging over the distribution of p-values (with distribution induced by uncertainty about unknown parameters); see, for example, Gelman, Meng and Stern (1996) and Bayarri and Berger (2000). Pratically, we compute the predictive value for each census tract, given by

$$\tilde{y}_i^{(m)} = f(y_i^{(t)}|\boldsymbol{\theta}^{(m)}), t = 1, \ldots, 5, m = 1, \ldots, M$$

where M is the number of iterations and $\theta^{(m)}$ is the value of the chain at iteration m. Once $\tilde{\boldsymbol{y}}_i^{(t)}$ has been obtained, we compute the the number M_1 of data such that $\tilde{y}_i^{(m)} < y_i^{(t)}$ and M_2 such that $\tilde{y}_i^{(m)} > y_i^{(t)}$. Hence the predictive p-value is computed as follows:

$$p.lower_i = \frac{M_1}{M}$$
$$p.upper_i = \frac{M_2}{M}$$
$$p\_value_i = min(p.upper_i, p.lower_i)$$

The predictive p-value ranges between 0.5 and 0, it il is close to zero it means that the actual datum is in the tails of the posterior distribution, hence it is not well explained by the model. In this work we consider as outliers those data with a p-value lower than 0.1.

The percentage of outliers in each model are listed in Table 4.5.

Even if the CAR model has more outlier than the others, anyway all the models fit well the observations. The grahs of the predictions in the independent random effects model and in the CAR model in 2010 are shown in Figures 4.16 and 4.17 The predictions of the models fit perfectly the data.

	1970	1980	1990	2000	2010
Baseline model	0.05	0.02	0.01	0.005	0.004
Indipendent model	0	0.007	0.004	0.004	0.005
CAR model	0.002	0.01	0.06	0.05	0.04

Table 4.5: Percentage of outliers for every years.

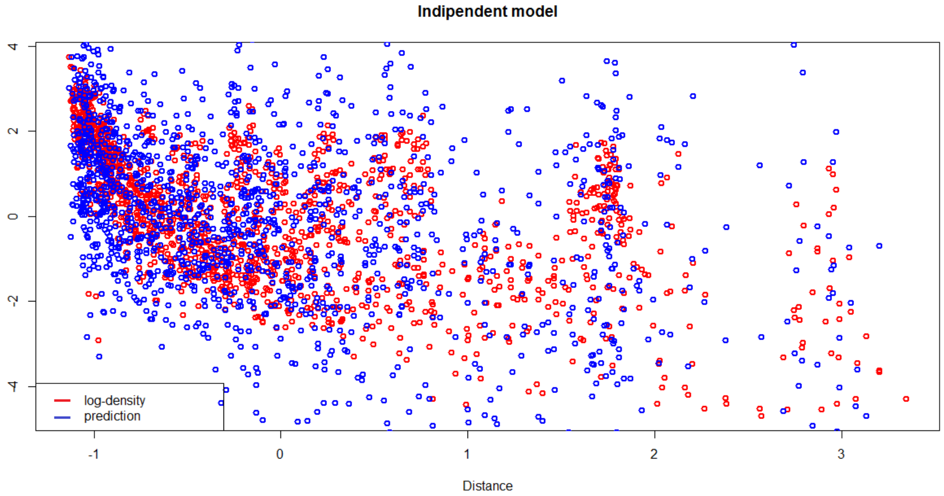

Figure 4.18: Predicted and actual log-density of population in 2010 under independet random effects model.

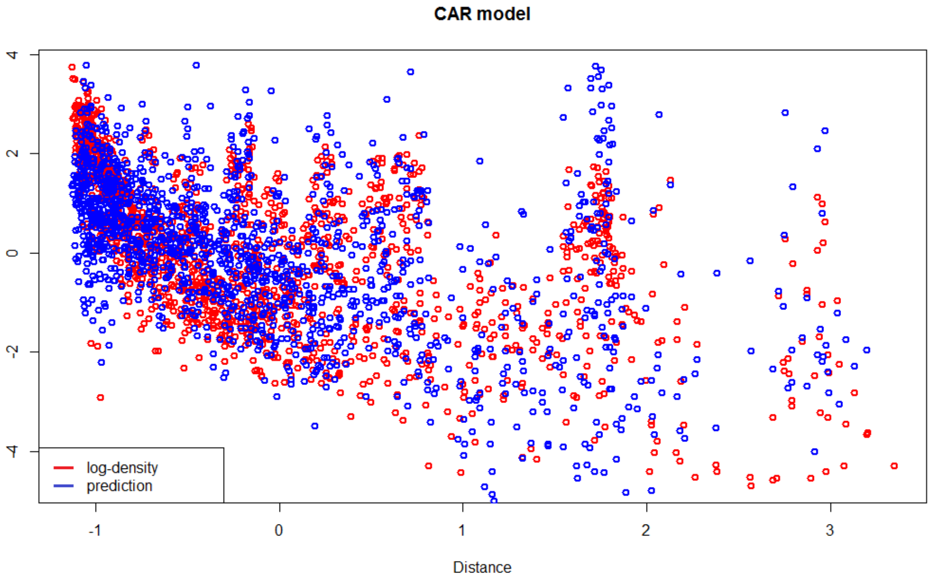

Figure 4.19: Predicted and actual log-density of population in 2010 under CAR model.

Chapter 5

Concluding Remarks

In this thesis we have applied three different dynamic Bayesian hierarchical regression lognormal models to population density at level of census tracts; we have compared the results to study spatial and time correlation among data of Massachusetts census tracts. Time dipendence has been introduced by an autoregressive structure on the regression coefficients. The spatial random effects take account for spatial correlation. As we expected, we found out that the reputation of a place plays an important role in influencing the individual decision. People prefer to settle in a wellness place, characterize by a good economic situation and similar ethnic composition. Anyway the distance from Boston remains the most important feature in leading population trend over time.
A remarkable result is that the introduction of spatial random effects improve the predictions of the actual population densities: it seems that it is sufficient to introduce independent random effects to better capture the population densities' variability. In addition, the Bayesian inference of the CAR model highlight an effective dependence by the neighbours features. It is worth to underline that the estimations for the regression coefficients are quite similar under baseline, independent and CAR random effects model.

Appendices

Appendix A

Full conditionals calculation

The calculus of the full conditional distributions are explained in details. Given the posterior distribution relative to the CAR model

$$\begin{aligned}
&\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(5)}, \sigma^2, \boldsymbol{\nu}, \boldsymbol{\Phi}^{(1)}, \dots, \boldsymbol{\Phi}^{(5)}, \Sigma_\beta, \tau^{(1)}, \dots, \tau(5), \dots \\
&\qquad\qquad\qquad\qquad \dots, \rho^{(1)}, \dots, \rho^{(5)} | \boldsymbol{Y}^{(1)}, \dots, \boldsymbol{Y}^{(5)}, V^{(1)}, \dots, V^{(5)} \\
&\propto \prod_{t=1}^{5} [\pi(\boldsymbol{Y}^{(t)} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\nu}, \sigma^2, \boldsymbol{\Phi}^{(t)}, V^{(t)})] \pi(\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(5)} | \boldsymbol{\beta}^{(0)}, \Sigma_\beta) \pi(\boldsymbol{\beta}^{(0)}) \pi(\Sigma_\beta) \pi(\sigma^2) \pi(\boldsymbol{\nu}) \times \\
&\qquad \times \prod_{t=1}^{5} [\pi(\boldsymbol{\Phi}^{(1)}, \dots, \boldsymbol{\Phi}^{(5)} | \rho^{(t)}, \tau^{(t)}) \pi(\rho^{(t)}) \pi(\tau^{(t)})] \\
&\propto \prod_{t=1}^{5} [\pi(\boldsymbol{Y}^{(t)} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\nu}, \sigma^2, \boldsymbol{\Phi}^{(t)}, V^{(t)}) \pi(\boldsymbol{\Phi}^{(1)}, \dots, \boldsymbol{\Phi}^{(5)} | \rho^{(t)}, \tau^{(t)}) \pi(\rho^{(t)}) \pi(\tau^{t}) \pi(\boldsymbol{\beta}^{(t)} | \boldsymbol{\beta}^{(t-1)}, \Sigma_\beta)] \times \\
&\qquad \times \pi(\boldsymbol{\nu}) \pi(\Sigma_\beta) \pi(\sigma^2) \pi(\boldsymbol{\beta}^{(0)}) \pi(\rho^{(t)}) \pi(\tau^{(t)})
\end{aligned}$$

we can compute the full conditionals for each parameters.

Update of regressiors coefficients

The update of the the regression coefficients depend on the time.

For $t = 0$

$$\begin{aligned}
\boldsymbol{\beta}^{(0)} | \boldsymbol{\beta}^{(1)}, \Sigma_\beta &\propto \pi(\boldsymbol{\beta}^{(1)} | \boldsymbol{\beta}^{(0)}, \Sigma_\beta) \pi(\boldsymbol{\beta}^{(0)}) \\
&\propto exp\left\{ -\frac{1}{2} (\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)})' \Sigma_\beta^{-1} (\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}) \right\} exp\left\{ -\frac{1}{200} (\boldsymbol{\beta}^{(0)})' \boldsymbol{\beta}^{(0)} \right\} \quad .
\end{aligned}$$

Since the covariance matrix is diagonal, we can factorize the distribution, for $h = 1, \dots, 8$

$$\beta_h^{(0)} | \beta_h^{(1)}, \sigma^2_{\beta_i} \propto exp\left\{ -\frac{1}{2\sigma^2_{\beta_h}} (\boldsymbol{\beta}_h^{(1)} - \boldsymbol{\beta}_h^{(0)})^2 \right\} exp\left\{ -\frac{1}{200} (\boldsymbol{\beta}_i^{(0)})^2 \right\}$$

$$\propto exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\beta}_h^{(0)})^2 \left(\frac{1}{\sigma^2_{\beta_h}} + \frac{1}{100} \right) - 2\boldsymbol{\beta}_h^{(0)} \frac{\boldsymbol{\beta}_h^{(1)}}{\sigma^2_{\beta_h}} \right] \right\}$$

therefore

$$\beta_h^{(0)} | \beta_h^{(1)}, \sigma_{\beta_h} \sim \mathcal{N} \left(\left(\frac{1}{\sigma^2_{\beta_h}} + \frac{1}{100} \right)^{-1} \frac{\beta_h^{(1)}}{\sigma^2_{\beta_h}}, \left(\frac{1}{\sigma^2_{\beta_h}} + \frac{1}{100} \right)^{-1} \right)$$

For $t = 1, 2, 3, 4$ the distribution depends on both the previous and the next values of β, we define the diagonal matrix N such that $N_{i,i} = \nu_{i[j]}$. As in the case above the distribution can be factorized, hence for $h = 1, \dots, 8$

$$\begin{aligned}
&\beta_h^{(t)} | \beta_h^{(t-1)}, \beta_h^{(t+1)}, \boldsymbol{\Phi}^{(t)}, \sigma^2_{\beta_i}, \boldsymbol{\nu}, \boldsymbol{Y}^{(\boldsymbol{t})}, V^{(t)}, \sigma^2 \\
&\propto \pi(\boldsymbol{Y}^{(t)} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\nu}, \sigma^2, \boldsymbol{\Phi}^{(t)}, V^{(t)}) \pi(\boldsymbol{\beta}^{(t)} | \boldsymbol{\beta}^{(t-1)}, \Sigma_\beta) \pi(\boldsymbol{\beta}^{(t+1)} | \boldsymbol{\beta}^{(t)}, \Sigma_\beta) \\
&\propto exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{Y}^{(t)} - \boldsymbol{V}_h^{(t)} \beta_h^{(t)} - \sum_{k \neq h} \boldsymbol{V}_k^{(t)} \beta_k^{(t)} - \boldsymbol{\Phi}^{(t)})' N^{-1} (\boldsymbol{Y}^{(t)} - \boldsymbol{V}_h^{(t)} \beta_h^{(t)} - \sum_{k \neq h} \boldsymbol{V}_k^{(t)} \beta_k^{(t)} - \boldsymbol{\Phi}^{(t)}) \right\} \times \\
&\qquad \times exp\left\{ -\frac{1}{2\sigma^2_{\beta_h}} (\beta_h^{(t)} - \beta_h^{(t-1)})^2 \right\} exp\left\{ -\frac{1}{2\sigma^2_{\beta_h}} (\beta_h^{(t+1)} - \beta_h^{(t)})^2 \right\} \\
&\propto \prod_{j=1}^{J[t]} exp\left\{ -\frac{1}{2\sigma^2 \nu_{i[j]}} (y_j^{(t)} - V_{h,j}^{(t)} \beta_h^{(t)} - \sum_{k \neq h} V_{k,j}^{(t)} \beta_k^{(t)} - \phi_j^{(t)})^2 \right\} \times \\
&\qquad \times exp\left\{ -\frac{1}{2\sigma^2_{\beta_h}} (\beta_h^{(t)} - \beta_h^{(t-1)})^2 \right\} exp\left\{ -\frac{1}{2\sigma^2_{\beta_h}} (\beta_h^{(t+1)} - \beta_h^{(t)})^2 \right\}
\end{aligned}$$

by defining the quantities

$$k^{(t)} = y_j^{(t)} - \sum_{k \neq h} V_{k,j}^{(t)} \beta_k^{(t)} - \phi_j^{(t)}$$

we rewrite

$$\begin{aligned}
&exp\left\{ -\frac{1}{2\sigma^2} \left[(\beta_h^{(t)})^2 \sum_{j=1}^{J[t]} \frac{(V_{h,j}^{(t)})^2}{\nu_{i[j]}} - 2\beta_h^{(t)} \sum_{j=1}^{J[t]} \frac{V_{h,j}^{(t)} k^{(t)}}{\nu_{i[j]}} \right] - \frac{1}{2\sigma^2_{\beta_h}} \left[2(\beta_h^{(t)})^2 - 2\beta_h^{(t)} (\beta_h^{(t+1)} + \beta_h^{(t-1)}) \right] \right\} \\
&\propto exp\left\{ -\frac{1}{2} \left[(\beta_h^{(t)})^2 \left(\frac{1}{\sigma^2} \sum_{j=1}^{J[t]} \frac{(V_{h,j}^{(t)})^2}{\nu_{i[j]}} + \frac{2}{\sigma^2_{\beta_h}} \right) - 2\beta_h^{(t)} \left(\frac{1}{\sigma^2} \sum_{j=1}^{J[t]} \frac{V_{h,j}^{(t)} k^{(t)}}{\nu_{i[j]}} + \frac{1}{\sigma^2_{\beta_h}} (\beta_h^{(t+1)} + \beta_h^{(t-1)}) \right) \right] \right\} \quad .
\end{aligned}$$

Set

$$\begin{aligned} a &= \frac{1}{\sigma^2}\sum_{j=1}^{J[t]}\frac{V_{h,j}^{(t)}k^{(t)}}{\nu_{i[j]}} + \frac{1}{\sigma^2_{\beta_h}}(\beta_h^{(t+1)} + \beta_h^{(t-1)}) \\ b &= \frac{1}{\sigma^2}\sum_{j=1}^{J[t]}\frac{(V_{h,j}^{(t)})^2}{\nu_{i[j]}} + \frac{2}{\sigma^2_{\beta_h}} \end{aligned}$$

a Gibbs sampling scheme can be adopted using

$$\beta_h^{(t)}|\beta_h^{(t-1)}, \beta_h^{(t+1)}, \mathbf{\Phi}^{(t)}, \sigma^2_{\beta_h}, \boldsymbol{\nu}, \boldsymbol{Y^{(t)}}, V^{(t)}, \sigma^2 \sim \mathcal{N}\left(b^{-1}a, b^{-1}\right) \quad .$$

The same strategy can be adopted for $t = 5$, with the only device that the distribution depens only on the previous step then the full distribution is

$$\begin{aligned} &\beta_h^{(t)}|\beta_h^{(t-1)}, \mathbf{\Phi}^{(t)}, \sigma^2_{\beta_h}, \boldsymbol{\nu}, \boldsymbol{Y^{(t)}}, V^{(t)}, \sigma^2 \\ &\propto \pi(\boldsymbol{Y}^{(t)}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\nu}, \sigma^2, \mathbf{\Phi}^{(t)}, V^{(t)})\pi(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^{(t-1)}, \Sigma_\beta) \end{aligned}$$

then the full distribution is

$$\beta_h^{(t)}|\beta_h^{(t-1)}, \mathbf{\Phi}^{(t)}, \sigma^2_{\beta_h}, \boldsymbol{\nu}, \boldsymbol{Y^{(t)}}, V^{(t)}, \sigma^2 \sim \mathcal{N}\left(b^{-1}a, b^{-1}\right)$$

with

$$\begin{aligned} a &= \frac{1}{\sigma^2}\sum_{j=1}^{J[t]}\frac{V_{h,j}^{(t)}k^{(t)}}{\nu_{i[j]}} + \frac{\beta_h^{(t-1)}}{\sigma^2_{\beta_h}} \\ b &= \frac{1}{\sigma^2}\sum_{j=1}^{J[t]}\frac{(V_{h,j}^{(t)})^2}{\nu_{i[j]}} + \frac{1}{\sigma^2_{\beta_h}} \quad . \end{aligned}$$

Update σ^2_β

The variances of the regression coefficients are conjugate to the model, we find a simple expression for the full distributions, let be $a = b = 0.001$

$$
\begin{aligned}
\sigma^2_{\beta_h}|\beta_h^{(1)},\ldots,\beta_h^{(5)} &\propto \prod_{t=1}^{5}\pi(\beta_h^{(t)}|\beta_h^{(t-1)},\sigma^2_{\beta_h})\pi(\sigma^2_{\beta_h})\\
&\propto \prod_{t=1}^{5}\left[\frac{1}{(\sigma^2_{\beta_h})^{1/2}}exp\left\{-\frac{(\beta_h^{(t)}-\beta_h^{(t-1)})^2}{2\sigma^2_{\beta_h}}\right\}\right]\frac{1}{(\sigma^2_{\beta_h})^{a+1}}exp\left\{-\frac{b}{\sigma^2_{\beta_h}}\right\}\mathbb{I}_{[0,\infty]}(\sigma^2_{\beta_h})\\
&\propto \frac{1}{(\sigma^2_{\beta_h})^{5/2+a+1}}exp\left\{-\frac{\sum\limits_{t=1}^{5}(\beta_h^{(t)}-\beta_h^{(t-1)})^2+2b}{2\sigma^2_{\beta_h}}\right\}\mathbb{I}_{[0,\infty]}(\sigma^2_{\beta_h})
\end{aligned}
$$

we find

$$
\sigma^2_{\beta_h}|\beta_h^{(1)},\ldots,\beta_h^{(5)} \sim IG\left(a+\frac{5}{2},b+\frac{\sum\limits_{t=1}^{5}(\beta_h^{(t)}-\beta_h^{(t-1)})^2}{2}\right) \quad \forall h=1,\ldots,8 \quad .
$$

Update σ^2

The variance σ^2 is simply updated with an inverse-gamma ditribution with hyperparameters $a=b=0.001$

$$
\begin{aligned}
&\sigma^2|\boldsymbol{\beta}_i^{(1)},\ldots,\boldsymbol{\beta}_i^{(5)},\boldsymbol{\nu},\boldsymbol{\Phi}^{(1)},\ldots,\boldsymbol{\Phi}^{(5)},\boldsymbol{Y}^{(\mathbf{1})},\ldots,\boldsymbol{Y}^{(\mathbf{5})},V^{(1)},\ldots,V^{(5)}\\
&\propto \prod_{t=1}^{5}\pi(\boldsymbol{Y}^{(t)}|\boldsymbol{\beta}^{(t)},\boldsymbol{\nu},\sigma^2,\phi_j^{(t)},V^{(t)})\pi(\sigma^2)\\
&\propto \prod_{t=1}^{5}\left[\frac{1}{(\sigma^2)^{J[t]/2}}exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Y}^{(t)}-V^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}^{(t)})'N^{-1}(\boldsymbol{Y}^{(t)}-V^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}^{(t)})\right\}\right]\times\\
&\qquad\times\frac{1}{(\sigma^2)^{a+1}}exp\left\{-\frac{b}{\sigma^2}\right\}\mathbb{I}_{[0,\infty]}(\sigma^2)\\
&\propto \frac{1}{(\sigma^2)^{\sum\limits_{t=1}^{5}J[t]/2+a+1}}exp\left\{-\frac{\sum\limits_{t=1}^{5}(\boldsymbol{Y}^{(t)}-V^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}^{(t)})'N^{-1}(\boldsymbol{Y}^{(t)}-V_i^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}^{(t)})+2b}{2\sigma^2}\right\}\mathbb{I}_{[0,\infty]}(\sigma^2)
\end{aligned}
$$

so we obtain

$$\sigma^2|\boldsymbol{\beta}^{(1)},\dots,\boldsymbol{\beta}^{(5)},\boldsymbol{\nu},\boldsymbol{\Phi}^{(1)},\dots,\boldsymbol{\Phi}^{(5)},\boldsymbol{Y}^{(1)},\dots,\boldsymbol{Y}^{(5)},V^{(1)},\dots,V^{(5)}\sim$$
$$\sim IG\left(a+\frac{\sum\limits_{t=1}^{5}J[t]}{2},b+\frac{\sum\limits_{t=1}^{5}(\boldsymbol{Y}^{(t)}-V^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}^{(t)})'N^{-1}(\boldsymbol{Y}^{(t)}-V^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}^{(t)})}{2}\right) \quad .$$

Update ν

Since $\nu_1,\dots,\nu_{14}$ are independent, the dataset and the parameters can be partitioned in this way: $\boldsymbol{Y}_i^{(t)}, V_i^{(t)}, \boldsymbol{\Phi}_i^{(t)}$ are the data and the parameters relative to $i-th$ county.

For $i=1,\dots,14$, let be$r=4$ and the diagonal matrix N such that $N_{i,i}=nu_{i[j]}$. Let us write the determinat of N as follow

$$|N|^{1/2}=\left(\prod_{j=1}^{n_1^{(t)}}\nu_1\cdots\prod_{j=1}^{n_i^{(t)}}\nu_i\cdots\prod_{j=1}^{n_{14}^{(t)}}\nu_{14}\right)^{1/2}\propto\nu_i^{n_i^{(t)}/2} \quad .$$

Therefore the conditional distribution is

$$\begin{aligned}
&\nu_i|\boldsymbol{\beta}^{(1)},\dots,\boldsymbol{\beta}^{(5)},\sigma^2,\boldsymbol{\Phi}_i^{(1)},\dots,\boldsymbol{\Phi}_i^{(5)},\boldsymbol{Y}_i^{(1)},\dots,\boldsymbol{Y}_i^{(5)},V_i^{(1)},\dots,V_i^{(5)}\\
&\propto\prod_{t=1}^{5}\pi(\boldsymbol{Y}_i^{(t)}|\boldsymbol{\beta}^{(t)},\nu_i,\sigma^2,\boldsymbol{\Phi}_i^{(t)},V_i^{(t)})\pi(\nu_i)\\
&\propto\prod_{t=1}^{5}\left[\frac{1}{\nu_i^{n_i^{(t)}/2}}exp\left\{-\frac{1}{\sigma^2\nu_i}(\boldsymbol{Y}_i^{(t)}-V_i^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}_i^{(t)})'(\boldsymbol{Y}_i^{(t)}-V_i^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}_i^{(t)})\right\}\right]\times\\
&\qquad\times\frac{1}{(\nu_i)^{r/2+1}}exp\left\{-\frac{r}{2\nu_i}\right\}\mathbb{I}_{[0,\infty]}(\nu_i)\\
&\propto\frac{1}{(\nu_i)^{\sum_{t=1}^{5}n_i^{(t)}/2+r/2+1}}\times\\
&\qquad\times exp\left\{-\frac{1}{\nu_i}\left(\sum_{i=1}^{5}\frac{(\boldsymbol{Y}_i^{(t)}-V_i^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}_i^{(t)})'(\boldsymbol{Y}_i^{(t)}-V_i^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}_i^{(t)})}{2\sigma^2}+\frac{r}{2}\right)\right\}\mathbb{I}_{[0,\infty]}(\nu_i)
\end{aligned}$$

by defining the quantities

$$k^{(t)}=\frac{(\boldsymbol{Y}_i^{(t)}-V_i^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}_i^{(t)})'(\boldsymbol{Y}_i^{(t)}-V_i^{(t)}\boldsymbol{\beta}^{(t)}-\boldsymbol{\Phi}_i^{(t)})}{2\sigma^2}$$

we obtain for all $i = 1, \dots, 14$

$$\nu_i | \boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(5)}, \sigma^2, \boldsymbol{\Phi}_i^{(1)}, \dots, \boldsymbol{\Phi}_i^{(5)}, \boldsymbol{Y}_i^{(1)}, \dots, \boldsymbol{Y}_i^{(5)}, V_i^{(1)}, \dots, V_i^{(5)} \sim$$
$$\sim IG\left(\frac{r}{2} + \frac{\sum\limits_{t=1}^{5} n_i^{(t)}}{2}, \frac{r}{2} + \sum_{i=1}^{5} k^{(t)}\right).$$

Update $\tau^{(t)}$

The hyperparameters fot $\tau^{(t)}$ are $a = 0.5$ and b=0.005. As the other variances prior, even in this case it is a conjugate prior, hence the full conditional distribution is simply

$$\tau^{(t)} | \boldsymbol{\Phi}^{(t)}, \rho^{(t)} \propto \pi(\boldsymbol{\Phi}^{(t)} | \rho^{(t)}, \tau^{(t)}) \pi(\tau^{(t)})$$
$$\propto \frac{1}{(\tau^{(t)})^{J[t]/2}} exp\left\{-\frac{1}{2\tau^{(t)}}(\boldsymbol{\Phi}^{(t)})'(D_w^{(t)} - \rho^{(t)} W^{(t)})\boldsymbol{\Phi}^{(t)}\right\} \frac{1}{(\tau^{(t)})^{a+1}} exp\left\{-\frac{b}{\tau^{(t)}}\right\} \mathbb{I}_{[0,\infty]}(\tau^{(t)})$$
$$\propto \frac{1}{(\tau^{(t)})^{a+J[t]/2+1}} exp\left\{-\frac{1}{\tau^{(t)}}\left[b + \frac{(\boldsymbol{\Phi}^{(t)})'(D_w^{(t)} - \rho^{(t)} W^{(t)})\boldsymbol{\Phi}^{(t)}}{2}\right]\right\} \mathbb{I}_{[0,\infty]}(\tau^{(t)})$$

$$\tau^{(t)} | \boldsymbol{Y}^{(t)}, V^{(t)}, \boldsymbol{\Phi}^{(t)}, \rho^{(t)} \sim IG\left(a + \frac{J[t]}{2}, b + \frac{(\boldsymbol{\Phi}^{(t)})'(D_w^{(t)} - \rho^{(t)} W^{(t)})\boldsymbol{\Phi}^{(t)}}{2}\right)$$

Update $\rho^{(t)}$

In order to sample from the full conditional distribution we need the relation between the eigenvalues of matrix $A^{(t)}$ and $I - \rho^{(t)} A^{(t)}$. Let $\mu_j^{(t)}$ be the eigenvalues of $I - \rho^{(t)} A^{(t)}$ and $\lambda_i^{(t)}$ be the eigenvalues of $A^{(t)}$, then the following holds

$$\mu_j^{(t)} = 1 - \rho^{(t)} \lambda_j^{(t)}$$

for the demostration we recall Section 3.3 . Hence the determinant of $I - \rho^{(t)} A^{(t)}$ can be written as

$$|(I - \rho^{(t)} A^{(t)})^{-1} D|^{1/2} = |I - \rho^{(t)} A^{(t)}|^{-1/2} |D|^{1/2} \propto \left(\prod_{j=1}^{J[t]} \mu_j^{(t)}\right)^{-1/2} \propto \left(\prod_{j=1}^{J[t]} (1 - \rho^{(t)} \lambda_j^{(t)})\right)^{-1/2} .$$

The full conditional distribution for $\rho^{(t)}$ is

$$\begin{aligned}
&\rho^{(t)}|\mathbf{\Phi}^{(t)},\tau^{(t)} \propto \pi(\mathbf{\Phi}^{(t)}|\rho^{(t)},\tau^{(t)})\pi(\tau^{(t)})\\
&\propto \frac{1}{|(I-\rho^{(t)}A^{(t)})^{-1}D|^{1/2}} exp\left\{-\frac{1}{2\tau^{(t)}}(\mathbf{\Phi}^{(t)})'D^{-1}(I-\rho^{(t)}A^{(t)})\mathbf{\Phi}^{(t)}\right\}\mathbb{I}_{[0,1]}(\rho^{(t)})\\
&\propto \left(\prod_{j=1}^{J[t]}(1-\rho^{(t)}\lambda_j^{(t)})\right)^{1/2} exp\left\{\frac{1}{2\tau^{(t)}}\rho^{(t)}(\mathbf{\Phi}^{(t)})'D^{-1}A^{(t)}\mathbf{\Phi}^{(t)}\right\}\mathbb{I}_{[0,1]}(\rho^{(t)}).
\end{aligned}$$

This distribution is unknown, so we sample with a step of Metropolis Hastings algorithm.

Update random effect parameters

We want to determine the full conditional distribution for $\Phi^{(t)}$

$$\begin{aligned}
&\mathbf{\Phi}^{(t)}|\boldsymbol{Y}^{(t)},V^{(t)},\boldsymbol{\nu},\sigma^2,\boldsymbol{\beta}^{(t)},\tau^{(t)},\rho^{(t)} \propto \pi(\boldsymbol{Y}^{(t)}|\boldsymbol{\beta}^{(t)},\boldsymbol{\nu},\sigma^2,\mathbf{\Phi}^{(t)},V^{(t)})\pi(\mathbf{\Phi}^{(t)}|\tau^{(t)},\rho^{(t)})\\
&\propto exp\left\{\frac{(\boldsymbol{Y}^{(t)}-V^{(t)}\boldsymbol{\beta}^{(t)}-\mathbf{\Phi}^{(t)})'N^{-1}(\boldsymbol{Y}^{(t)}-V^{(t)}\boldsymbol{\beta}^{(t)}-\mathbf{\Phi}^{(t)})}{2\sigma^2}\right\} exp\left\{-\frac{1}{2}(\mathbf{\Phi}^{(t)})'\frac{D_w^{(t)}-\rho^{(t)}W^{(t)}}{\tau^{(t)}}\mathbf{\Phi}^{(t)}\right\}\\
&\propto \left\{-\frac{1}{2}\left[(\mathbf{\Phi}^{(t)})'\frac{N^{-1}}{\sigma^2}\mathbf{\Phi}^{(t)}-2(\mathbf{\Phi}^{(t)})'\frac{N^{-1}}{\sigma^2}(\boldsymbol{Y}^{(t)}-V^{(t)}\boldsymbol{\beta}^{(t)})+(\mathbf{\Phi}^{(t)})'\frac{D_w^{(t)}-\rho^{(t)}W^{(t)}}{\tau^{(t)}}\mathbf{\Phi}^{(t)}\right]\right\}\\
&\propto \left\{-\frac{1}{2}\left[(\mathbf{\Phi}^{(t)})'\left(\frac{N^{-1}}{\sigma^2}+\frac{D_w^{(t)}-\rho^{(t)}W^{(t)}}{\tau^{(t)}}\right)\mathbf{\Phi}^{(t)}-2(\mathbf{\Phi}^{(t)})'\frac{N^{-1}}{\sigma^2}(\boldsymbol{Y}^{(t)}-V^{(t)}\boldsymbol{\beta}^{(t)})\right]\right\}
\end{aligned}$$

hence this is the kernel of a multinormal distribution with parameters updated as follow

$$\begin{aligned}
&\mathbf{\Phi}^{(t)}|\boldsymbol{Y}^{(t)},V^{(t)},\boldsymbol{\nu},\sigma^2,\boldsymbol{\beta}^{(t)},\tau^{(t)},\rho^{(t)} \sim\\
&\quad\sim \mathcal{MN}\left(\left(\frac{N^{-1}}{\sigma^2}+\frac{D_w^{(t)}-\rho^{(t)}W^{(t)}}{\tau^{(t)}}\right)^{-1}\frac{N^{-1}}{\sigma^2},\left(\frac{N^{-1}}{\sigma^2}+\frac{D_w^{(t)}-\rho^{(t)}W^{(t)}}{\tau^{(t)}}\right)^{-1}\right) \quad .
\end{aligned}$$

Since sampling from a multinormal distribution of such dimension is computationally slow and some numeric error could occur, we propose a second full conditional distribution obtained startinf from Equation 1.9. We condition the previos distribution with respect to $\boldsymbol{\Phi}_j^{(t)}$ and then the Bayes' theorem is applied

$$\begin{aligned}
&\pi(\phi_j^{(t)}|\boldsymbol{\Phi}_{-j}^{(t)}, \boldsymbol{Y}^{(t)}, V^{(t)}, \boldsymbol{\nu}, \sigma^2, \boldsymbol{\beta}^{(t)}, \tau^{(t)}, \rho^{(t)}) \\
&= \frac{\pi(\boldsymbol{Y}^{(t)}, \boldsymbol{\nu}, \sigma^2, \boldsymbol{\beta}^{(t)}, \tau^{(t)}, \rho^{(t)}|\phi_j^{(t)}, \boldsymbol{\Phi}_{-j}^{(t)})\pi(\phi_j^{(t)}|\boldsymbol{\Phi}_{-j}^{(t)})}{\pi(\boldsymbol{Y}^{(t)}, \boldsymbol{\nu}, \sigma^2, \boldsymbol{\beta}^{(t)}, \tau^{(t)}, \rho^{(t)}|\boldsymbol{\Phi}_{-j}^{(t)})} \\
&\propto \pi(\boldsymbol{Y}^{(t)}|\boldsymbol{\Phi}^{(t)})\pi(\phi_j^{(t)}|\boldsymbol{\Phi}_{-j}^{(t)}) \quad .
\end{aligned}$$

Therefore we use the simpler full conditionals

$$\phi_j^{(t)}|\boldsymbol{\Phi}_{-j}^{(t)}, y_j^{(t)}, V_j^{(t)}, \nu_{i[j]}, \sigma^2, \boldsymbol{\beta}^{(t)}, \tau^{(t)}, \rho^{(t)} \sim \mathcal{N}(b^{-1}a, b^{-1})$$

where

$$\begin{aligned}
b &= \frac{1}{\sigma^2 \nu_{i[j]}} + \frac{n_j^{(t)}}{\tau^{(t)}} \\
a &= \frac{y_j^{(t)} - V_j^{(t)}\boldsymbol{\beta}^{(t)}}{\sigma^2 \nu_{i[j]}} + \frac{n_j^{(t)}\rho^{(t)}A_j^{(t)}\boldsymbol{\Phi}^{(t)}}{\tau^{(t)}} \quad .
\end{aligned}$$

Appendix B

Implementation codes

The first code is the Stan code for the model with spatial independent random effects.

```
# block of data
data
{
int<lower = 1> TT;                        # times
int<lower = 1> RIGHE;                     # max number of row
int<lower =1 > I;                         # number of counties
int<lower =0> p;                          # number of regressors
int r ;
int<lower = 0> J[TT];                     # number of census tracts
vector[I] zeta;
matrix[RIGHE, TT] y;
int <lower=0,upper=14> cc[RIGHE,TT];
matrix[RIGHE,TT] dist;
matrix[RIGHE,TT] mix;
matrix[RIGHE,TT] gini;
matrix[RIGHE,TT] educ ;
matrix[RIGHE,TT] income;
}

# block of parameters to sample
parameters
{
vector<lower = 0>[I] nu;
vector<lower = 0>[p+2] sigma_B ;
real<lower=0> sigma_comune;
matrix [p,TT] B;
matrix [2,TT-1] BI;                #there is no income data in 1970
vector [p+2] valori_iniziali;

vector[J[1]] phi1;
vector[J[2]] phi2;
vector[J[3]] phi3;
vector[J[4]] phi4;
vector[J[5]] phi5;
vector[TT] sigma_phi;
}
#block for the operations
transformed parameters{
vector[I] sigma ;
matrix [RIGHE,TT] media_beta;
matrix[RIGHE,TT] media_comune;

for (i in 1:I){ sigma[i] <- pow(nu[i],0.5)*pow(sigma_comune,0.5)*pow(r,-0.5); }

for(j in 1:J[1]){
media_beta[j,1]<-dist[j,1]*B[3,1]+ mix[j,1]*B[4,1] + gini[j,1]*B[5,1] + educ[j,1]*B[6,1];
}

for(t in 2:TT){
for(j in 1:J[t]){
media_beta[j,t]<-dist[j,t]*B[3,t]+ mix[j,t]*B[4,t] + gini[j,t]*B[5,t]+ educ[j,t]*B[6,t]+                        income[j,t]*BI[1,t-1]+income[j,t]*dist[j,t]*BI[2,t-1];
}
}
for(t in 1:TT){
for(j in 1:J[t]){ media_comune[j,t]<-B[1,t]+ B[2,t]* zeta [cc[j,t]];  } }
}

model
{
# Prior:
# 1) variance prior
sigma_comune ~ inv_gamma(0.001, 0.001);

for (i in 1:(p+2)){
sigma_B[i] ~ uniform(0.001,10); }

for (i in 1:I) {
nu[i] ~ chi_square (r); }
```

```

# 2) AR(1)
for (i in 1:(p+2)){valori_iniziali[i]~ normal (0,pow(10,0.5));}

for (i in 1:p){
B[i,1]~ normal(valori_iniziali[i],sigma_B[i]);
for (t in 2:TT){
B[i,t]~ normal(B[i,t-1],sigma_B[i]);  }
}
for (i in 1:2){
BI[i,1]~ normal(valori_iniziali[i+6],sigma_B[i+6]);
for (t in 2:4){
BI[i,t]~ normal(BI[i,t-1],sigma_B[i+6]);  }
}

#3) independent random effects
for (t in 1:TT){ sigma_phi[t] ~ inv_gamma(0.001, 0.001); }

for(i in 1:J[1]){ phi1[i] ~ normal(0,pow(sigma_phi[1],0.5)); }
for(i in 1:J[2]){ phi2[i] ~ normal(0,pow(sigma_phi[2],0.5)); }
for(i in 1:J[3]){ phi3[i] ~ normal(0,pow(sigma_phi[3],0.5)); }
for(i in 1:J[4]){ phi4[i] ~ normal(0,pow(sigma_phi[4],0.5)); }
for(i in 1:RIGHE){ phi5[i] ~ normal(0,pow(sigma_phi[5],0.5)); }

# Likelihood:
for (j in 1:J[1]){
y[j,1] ~ lognormal(media_beta[j,1]+ phi1[j]+media_comune[j,1] , sigma[cc[j,1]] );
     }
for (j in 1:J[2]){
y[j,2] ~ lognormal(media_beta[j,2]+ phi2[j]+media_comune[j,2] , sigma[cc[j,2]] );
       }
for (j in 1:J[3]){
y[j,3] ~ lognormal(media_beta[j,3]+ phi3[j]+media_comune[j,3] , sigma[cc[j,3]] );
       }
for (j in 1:J[4]){
y[j,4] ~ lognormal(media_beta[j,4]+ phi4[j]+media_comune[j,4] , sigma[cc[j,4]] );
       }
for (j in 1:J[5]){
y[j,5] ~ lognormal(media_beta[j,5]+ phi5[j]+media_comune[j,5] , sigma[cc[j,5]] );
     }

}
```

The second code is written in Julia, it is the implemenatation of the CAR model. For this second one we do not shown all the code, only the most important parts.

The functions for sampling from the full distribution of the regression coefficient, the variance of the coefficients and the variable ρ are reported:

```

#update of regression coefficients

function upd_b0(beta, sigma_beta, j, diago, sigma_y, y , p ,shift)
d =length(beta)
out=zeros(Float64,d)
var=1./(1./sigma_beta +1/ 100)
med=var*beta[2]./sigma_beta
out[1]=rand(Normal(med,var),1)[]

for t in 2:(d-1)
temp= diago[1:J[t-shift],t-shift]
b_inv = 1/(sum(temp)/sigma_y + 2/sigma_beta)
a=dot(temp,y[1:J[t-shift],t-shift]-p[1:J[t-shift],t-shift])./sigma_y + (out[t-1]+beta[t+1])./sigma_beta
out[t]=rand(Normal(b_inv*a,b_inv),1)[]
end

t=d
temp= diago[1:J[t-shift],t-shift]
b_inv = 1/(sum(temp)/sigma_y + 1/sigma_beta)
a=dot(temp,y[1:J[t-shift],t-shift]-p[1:J[t-shift],t-shift])./sigma_y + out[t-1]./sigma_beta
out[t]=rand(Normal(b_inv*a,b_inv),1)[]

return(out)
end;

function upd_beta(beta, data, sigma_beta, j, diago, sigma_y,y,p,shift)
d = length(beta)
out=zeros(Float64,d)
var=1/(1./sigma_beta +1/ 100)
med=var*beta[2]/sigma_beta
out[1]=rand(Normal(med,var),1)[]
for t in 2:(d-1)
temp= data[1:J[t-shift],t-shift].*diago[1:J[t-shift],t-shift]
b_inv = 1/(dot(temp,data[1:J[t-shift],t-shift])./sigma_y + 2/sigma_beta)
```

```
a=dot(temp,y[1:J[t-shift],t-shift]-p[1:J[t-shift],t-shift])./sigma_y + (out[t-1]+beta[t+1])./sigma_beta
out[t]=rand(Normal(b_inv*a,b_inv),1)[]
end

t=d
temp= data[1:J[t-shift],t-shift].*diago[1:J[t-shift],t-shift]
b_inv = 1/(dot(temp,data[1:J[t-shift],t-shift])./sigma_y + 1/sigma_beta)
a= dot(temp,y[1:J[t-shift],t-shift]-p[1:J[t-shift],t-shift])./sigma_y + out[t-1]./sigma_beta
out[t]=rand(Normal(b_inv*a,b_inv),1)[]

return(out)
end;
```

```
function logfun(x,p,M,aut,t)
temp=sum(log(1 - x*aut));
out = 0.5*temp - 0.5*(x/t)*transpose(p)*M*p
return(out)
end;
function upd_rho ( r, phi,M,aut,t)
out = r
delta =  rand(Uniform(0,1), 1) [] # campiono da proposal
acp = logfun(delta,phi,M,aut,t)-logfun(r,phi,M,aut,t)
acp = minimum([0.0, acp[]])
ta = log(rand(Uniform(0,1),1)[])
if   ta < acp
out = delta
end
return (out)
end;

function sigma_beta(a_sigma,b_sigma, b)
temp=0
temp=sum((b[2:end]-b[1:(end-1)]).^2)
out = rand(InverseGamma(a_sigma, b_sigma + 0.5*temp),1 )[]
end;
```

The Gibbs sampling algorithm has been implemented as follow

```
#Gibbs

for k in 1:n_iter
t=1
TR[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
```

```
beta_educ[t+1]*educ[1:J[t],t]-phi[1:J[t],t]

for t in 2:TT
TR[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-beta_income[t]*income[1:J[t],t]-beta_income_dist[t]*L[1:J[t],t]-phi[1:J[t],t]
end
#update nu --------------------------------------------------------------------
for i in 1:14
b=0
for t in 1:TT
b=b+sum(TR[indici_contea[i,1:n_i[i,t],t],t].^2)
end
nu[i]=rand(InverseGamma(a_nu[i],(b_nu+0.5*b/sigma_y)[]),1)[]
end

#update sigma_y --------------------------
for t in 1:TT
diago[1:J[t],t]=1./nu[cc[1:J[t],t]]
end

K = 0
for t in 1:TT
#K=K+ transpose(TR[1:J[t],t])*(diago[1:J[t],t].*TR[1:J[t],t])
K=K+dot(TR[1:J[t],t],diago[1:J[t],t].*TR[1:J[t],t])
end
sigma_y = rand(InverseGamma(a_sigma_y,b_sigma_y + 0.5*K),1)[]
## update beta e sigma_beta--------------------------------------
sigma_beta_dist = sigma_beta(a_sigma_beta,b_sigma_beta, beta_dist);
sigma_b0 = sigma_beta(a_sigma_beta,b_sigma_beta, beta_dist);
sigma_b1 = sigma_beta(a_sigma_beta,b_sigma_beta, beta_dist);
sigma_beta_mix = sigma_beta(a_sigma_beta,b_sigma_beta, beta_mix);
sigma_beta_gini = sigma_beta(a_sigma_beta,b_sigma_beta, beta_gini);
sigma_beta_educ =  sigma_beta(a_sigma_beta,b_sigma_beta, beta_educ);
sigma_beta_income = sigma_beta(a_sigma_beta,b_sigma_beta, beta_income);
sigma_beta_income_dist = sigma_beta(a_sigma_beta,b_sigma_beta, beta_income_dist);

t=1
delta[1:J[t],t]=y[1:J[t],t]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-phi[1:J[t],t]

for t in 2:TT
```

```
delta[1:J[t],t]=y[1:J[t],t]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-beta_income[t]*income[1:J[t],t]-beta_income_dist[t]*L[1:J[t],t]-phi[1:J[t],t]
end
b0=upd_b0(b0,sigma_b0,J,diago,sigma_y,y,phi,1)
t=1
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-beta_dist[t+1]*dist[1:J[t],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-phi[1:J[t],t]

for t in 2:TT
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-beta_dist[t+1]*dist[1:J[t],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-beta_income[t]*income[1:J[t],t]-beta_income_dist[t]*L[1:J[t],t]-phi[1:J[t],t]
end
b1=upd_beta(b1,z, sigma_b1,J,diago,sigma_y,y,phi,1)
t=1
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-phi[1:J[t],t]

for t in 2:TT
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-beta_income[t]*income[1:J[t],t]-beta_income_dist[t]*L[1:J[t],t]-phi[1:J[t],t]
end
beta_dist=upd_beta(beta_dist, dist, sigma_beta_dist,J,diago,sigma_y,y,phi,1)
t=1
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-phi[1:J[t],t]

for t in 2:TT
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-beta_income[t]*income[1:J[t],t]-beta_income_dist[t]*L[1:J[t],t]-phi[1:J[t],t]
end
beta_mix=upd_beta(beta_mix, mix, sigma_beta_mix,J,diago,sigma_y,y,phi,1)
t=1
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t
```

```
    ],t]-beta_mix[t+1]*mix[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-phi[1:J[t],t]

for t in 2:TT
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t
    ],t]-beta_mix[t+1]*mix[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-beta_income[t]*income[1:J[t],t]-beta_income_dist[t]*
    L[1:J[t],t]-phi[1:J[t],t]
end
beta_gini=upd_beta(beta_gini, gini, sigma_beta_gini,J,diago,sigma_y,y,phi,1)
t=1
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t
    ],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
phi[1:J[t],t]

for t in 2:TT
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t
    ],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_income[t]*income[1:J[t],t]-beta_income_dist[t]*L[1:J[t],t]-phi[1:J[t],t]
end
beta_educ=upd_beta(beta_educ,educ,sigma_beta_educ,J,diago,sigma_y,y,phi,1)
t=1
delta[1:J[t],t]=zeros(Float64,J[t])
for t in 2:TT
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t
    ],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-beta_income_dist[t]*L[1:J[t],t]-phi[1:J[t],t]
end
beta_income=upd_beta(beta_income, income, sigma_beta_income,J,diago,sigma_y,y,phi
    ,0)
t=1
delta[1:J[t],t]=zeros(Float64,J[t])

for t in 2:TT
delta[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t
    ],t]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-beta_income[t]*income[1:J[t],t]-phi[1:J[t],t]
end
beta_income_dist=upd_beta(beta_income_dist, L, sigma_beta_income_dist,J,diago,
    sigma_y,y,phi,0)

#update CAR ----------------------
t=1
TR[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t],t
```

```
    ]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]

for t in 2:TT
TR[1:J[t],t]=y[1:J[t],t]-b0[t+1]-b1[t+1]*z[1:J[t],t]-beta_dist[t+1]*dist[1:J[t],t
    ]-beta_mix[t+1]*mix[1:J[t],t]-beta_gini[t+1]*gini[1:J[t],t]-
beta_educ[t+1]*educ[1:J[t],t]-beta_income[t]*income[1:J[t],t]-beta_income_dist[t]*
    L[1:J[t],t]
end

for  t in 1:TT
rho[t]=upd_rho(rho[t],phi[1:J[t],t],W[1:J[t],1:J[t],t],autovalori[1:J[t],t],tau[t
    ])

b_tau=0.5*dot(phi[1:J[t],t],(D[1:J[t],1:J[t],t] - rho[t] * W[1:J[t],1:J[t],t] )*
    phi[1:J[t],t] )
tau[t] = rand(InverseGamma(0.5 + 0.5*J[t], 0.005 + b_tau),1)[]



b=1./(diago[1:J[t],t]./sigma_y +n[1:J[t],t]./tau[t])
a= TR[1:J[t],t].*diago[1:J[t],t]./sigma_y
for j in 1:J[t]
   med=b[j].*(a[j]+rho[t]*(n[j,t]./tau[t])*A[j,1:J[t],t]*phi[1:J[t],t])
   phi[j,t]=rand(Normal(med,b[j]),1)[]
end #for j
end # for t

end
```

Appendix C

Tables of the posterior quantiles of the regression coefficients

	1970	1980	1990	2000	2010
Overall Intercept(b_0)	**0.22**	0.05	$\mathbf{-0.13}$	$\mathbf{-0.17}$	$\mathbf{-0.17}$
	(0.15,0.29)	(-0.02,0.12)	(-0.13,-0.07)	(-0.23,-0.10)	(-0.24,-0.11)
Distance(β_1)	0.03	$\mathbf{-0.68}$	$\mathbf{-1.04}$	$\mathbf{-1.02}$	$\mathbf{-0.99}$
	(-0.05,0.10)	(-0.76,-0.61)	(-1.12,-0.96)	(-1.10,-0.94)	(-1.07,-0.91)
Mix(β_2)	$\mathbf{-0.19}$	$\mathbf{-0.12}$	$\mathbf{-0.20}$	$\mathbf{-0.28}$	$\mathbf{-0.31}$
	(-0.25,-0.13)	(-0.19,-0.60)	(-0.27,-0.14)	(-0.35,-0.21)	(-0.39,-0.23)
Gini(β_3)	$\mathbf{-0.21}$	**0.73**	**0.53**	**0.42**	**0.40**
	(-0.27,-0.15)	(0.64,0.81)	(0.04,0.62)	(0.04,0.51)	(0.05,0.53)
Education(β_4)	0.01	-0.02	-0.05	$\mathbf{-0.08}$	-0.06
	(-0.05,0.07)	(-0.10,0.06)	(-0.12,0.02)	(-0.18,-0.01)	(-0.14,0.01)
Income(β_5)	NA	$\mathbf{-0.20}$	$\mathbf{-0.34}$	$\mathbf{-0.35}$	$\mathbf{-0.36}$
		(-0.31,-0.08)	(-0.44,-0.24)	(-0.45,-0.24)	(-0.47,-0.24)
Income*Distance(β_6)	NA	$\mathbf{-0.48}$	$\mathbf{-0.50}$	$\mathbf{-0.52}$	$\mathbf{-0.55}$
		(-0.55,-0.39)	(-0.57,-0.42)	(-0.60,-0.45)	(-0.64,-0.47)

Table C.1: Estimation of β coefficients in the baseline model. For each regressor we show the mean (it is highligth if the coefficient results significative) and 2.5%,97.5% quantiles.

	1970	1980	1990	2000	2010
Overall Intercept(b_0)	**0.18** (0.01,0.27)	0.05 (-0.02,0.11)	**−0.17** (-0.22,-0.08)	**−0.20** (-0.26,-0.14)	**−0.21** (-0.27,-0.15)
Distance(β_1)	0.03 (-0.06,0.12)	**−0.69** (-0.77,-0.62)	**−1.10** (-1.20,-1-02)	**−1.11** (-1.19,-1.03)	**−1.08** (-1.15,-0.99)
Mix(β_2)	**−0.20** (-0.28,-0.13)	**−0.13** (-0.19,-0.06)	**−0.20** (-0.26,-0.14)	**−0.28** (-0.35,-0.22)	**−0.31** (-0.39,-0.24)
Gini(β_3)	**−0.24** (-0.32,-0.15)	**0.74** (0.65,0.82)	**0.54** (0.46,0.62)	**0.41** (0.33,0.48)	**0.38** (0.31,0.46)
Education(β_4)	0.02 (-0.06,11)	-0.02 (-0.11,0.04)	-0.05 (-0.11,0.03)	-0.08 (-0.16,-0.01)	-0.05 (-0.12,0.02)
Income(β_5)	NA	**−0.20** (-0.31,-0.08)	**−0.34** (-0.44,-0.24)	**−0.35** (-0.45,-0.24)	**−0.36** (-0.47,-0.24)
Income*Distance(β_6)	NA	**−0.48** (-0.55,-0.39)	**−0.50** (-0.57,-0.42)	**−0.52** (-0.60,-0.45)	**−0.55** (-0.64,-0.47)

Table C.2: Estimation of β coefficients in the model with independent random effects. For each regressor we show the mean (it is highligth if the coefficient results significative) and 2.5%,97.5% quantiles.

	1970	1980	1990	2000	2010
Overall Intercept(b_0)	**0.54** (0.48,0.59)	0.15 (0.06,0.22)	**−0.09** (-0.15,-0.02)	**−0.13** (-0.18,-0.06)	**−0.13** (-0.19,-0.06)
Distance(β_1)	0.19 (0.14,0.25)	**−0.72** (-0.78,-0.65)	**−1.17** (-1.24,-1.10)	**−1.19** (-1.26,-1.12)	**−1.17** (-1.25,-1.10)
Mix(β_2)	**−0.41** (-0.45,-0.37)	**−0.20** (-0.26,-0.14)	**−0.21** (-0.26,-0.15)	**−0.31** (-0.36,-0.25)	**−0.35** (-0.41,-0.29)
Gini(β_3)	**−0.24** (-0.28,-0.20)	**0.80** (0.72,0.88)	**0.57** (0.48,0.64)	**0.43** (0.35,0.50)	**0.39** (0.32,0.46)
Education(β_4)	0.01 (-0.03,0.05)	-0.00 (-0.06,0.08)	-0.01 (-0.07,0.05)	-0.04 (-0.11,0.02)	-0.03 (-0.09,0.03)
Income(β_5)	NA	**−0.22** (-0.32,-0.13)	**−0.47** (-0.55,-0.91)	**−0.53** (-0.62,-0.44)	**−0.56** (-0.65,-0.47)
Income*Distance(β_6)	NA	**−0.50** (-0.57,-0.43)	**−0.59** (-0.67,-0.52)	**−0.66** (-0.73,-0.58)	**−0.69** (-0.77,-0.61)

Table C.3: Estimation of β coefficients in the proper CAR model. For each regressor we show the mean (it is highligth if the coefficient results significative) and 2.5%,97.5% quantiles.

Appendix D

Convergence diagnostic of MCMC chains

In this section we present some diagnostic analysis for the convergence of the model. Because of the huge number of parameters, we present the results only for some parameters under the CAR model. The traceplot and the autocorrelation for some fundamental parameters are represented in Figures D.1, D.2 and D.3. The graps show that the chains have reached the convergence, since the traceplots are "fat" and the autocorrelations decrease to zero very quickly. The convergence of the chains is checked via Geweke's

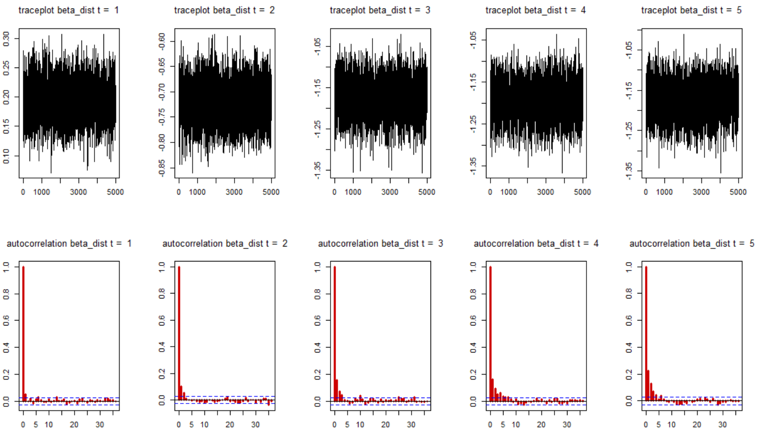

Figure D.1: Traceplot and autocorrelation for the distance regressor under the CAR model.

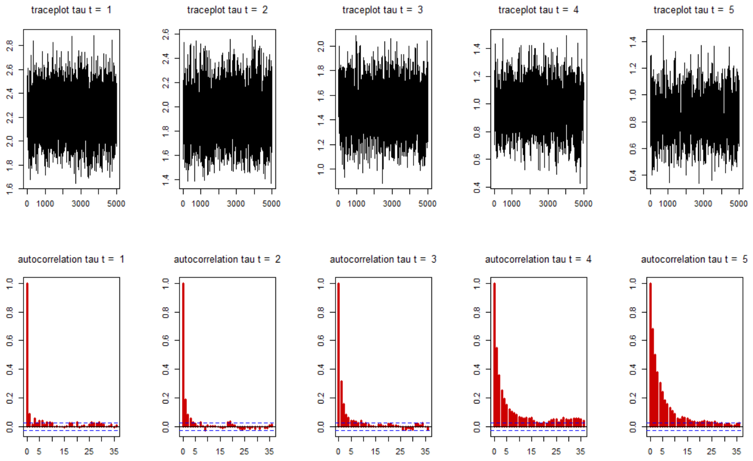

Figure D.2: Traceplot and autocorrelation for τ under the CAR model.

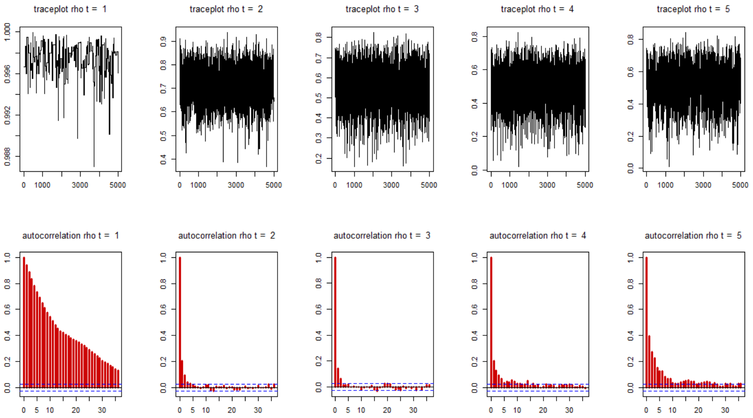

Figure D.3: Traceplot and autocorrelation for ρ under the CAR model.

statistics. The idea behind this test is simple: it is analogous to test the equality of the means of the first and the last part of a Markov chain (by default the first 10% and the last 50%). If the samples are drawn from the stationary distribution of the chain, the two means are equal and Geweke's statistic has an asymptotically standars normal distribution. In Figures D.4, D.5 and D.6 the test statistics are displayed. Since the values are in the interval $[-2, 2]$ in the majority of cases, we can conclude thata the MCMC chains are stationary.

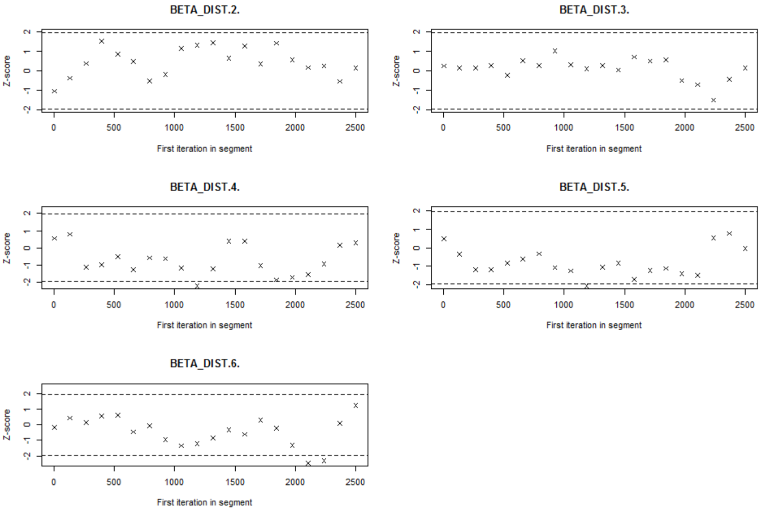

Figure D.4: Geweke test for the distance regressor under the CAR model in 1970, 1980, 1990, 2000 and 2010.

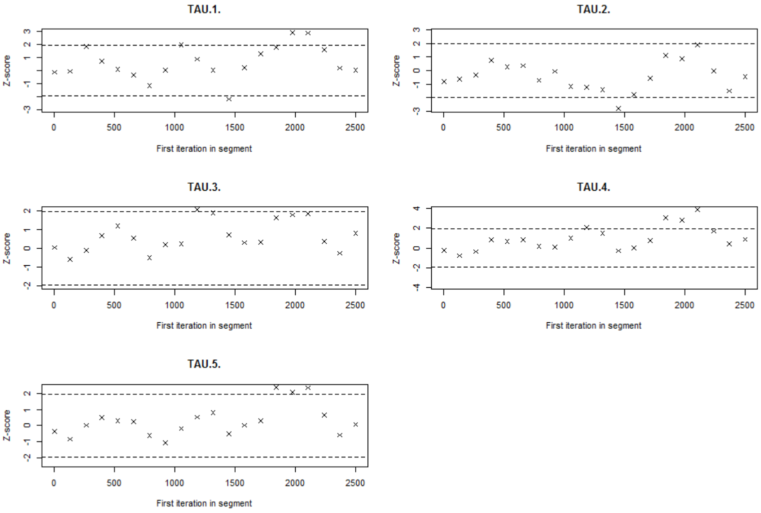

Figure D.5: Geweke test for τ under the CAR model in 1970, 1980, 1990, 2000 and 2010.

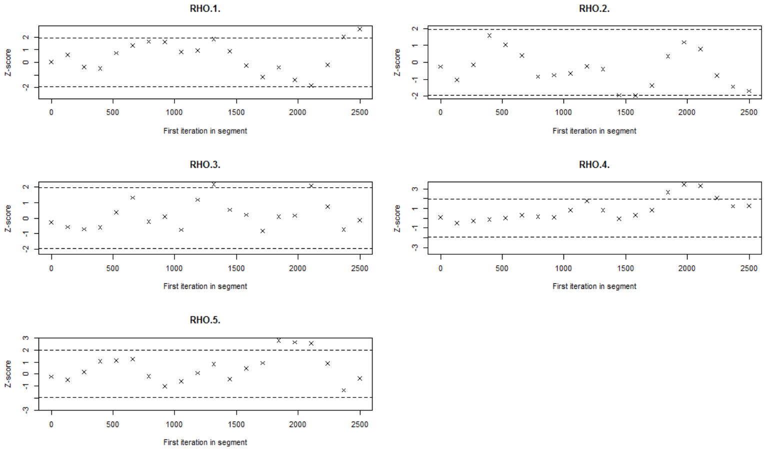

Figure D.6: Geweke test for ρ under the CAR model in 1970, 1980, 1990, 2000 and 2010.

In Table D.1, values of the effective numbers of iteration for convergence are listed. The effective number of iteration that garantee convergence is really lower than the iteration number, hence the convergence is reached.

	1970	1980	1990	2000	2010
β_{dist}	4127	3715	3063	2708	2334
τ	3122	2886	2085	1061	628
ρ	122	2951	3431	1856	1137

Table D.1: Effective size for the chains of β_{dist}, τ, ρ.

Bibliography

[1] Sudipto BANERJEE, Bradley P.CARLIN, Atan E. GELFAND (2015), *Hierarchical Modeling and Analysis for Spatial Data.* CRC Press.

[2] Julian BESAG, *Spatial Interaction and the Statistical Analysis of Lattice System.* Royal Statistical Society, 1974.

[3] Bradley P.CARLIN, Sudipto BANERJEE (2003a), *Hierarchical Multivariate CAR Model for Spatio-Temporally Correlated Survival Data.* Bayesian Statistics, **7**, 45-63.

[4] Bradley P.CARLIN, Sudipto BANERJEE (2003b), *Semiparametric spatio-temporal frailty modeling.* Environmetrics **14**, 523-535.

[5] Ilenia EPIFANI, Rosella NICOLINI (2013), *On the density distribution across space: a probabilistic approach.* Journal of Regional Science, **53**, 481-510.

[6] Ilenia EPIFANI, Rosella NICOLINI (2015a), *Modelling Population Density Over Time: How Spatial Distance Matters.* Regional Studies .

[7] Ilenia EPIFANI, Rosella NICOLINI (2015b), *The importance of historical linkages in shaping population density across space.* 63rd NARSC meeting, November 9-12.

[8] Andrew GELMAN (2005), *Fuzzy and Bayesian p-Values and u-Values.* Statistical Science, **20**, 380-381.

[9] Arthur GETIS (2008), *A History ot the Concept of Spatial Autocorrelation: A Geographer's Perspective.* Geographical Analysis, **40**, 297-309.

[10] Arthur Getis (1991), *Spatial Interaction and Spatial Autocorrelation: A Cross-Product Approach.*. Spatial Interaction and Spatial Autocorrelation: Across-Product Approach. Environment and Planning A 23:1269-1277.

[11] J.Geweke (1993), *Bayesian Treatment of the independent student-t Linear Model.* Journal of applied econometrics, **8**, 19-40.

[12] Simon Jackman (2009), *Bayesian Analysis for the Social Science.* Wiley.

[13] Duncan Lee (2013), *CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors.* Journal of Statistical Software, **55**, 1-24.

[14] James LeSage, R. Kelley Pace (2009), *Introduction to Spatial Econometrics.* CRC Press .

[15] Raquel Prado, Mike West (2014), *Time series: Modeling, Computation and Inference.* CRC Press.

[16] C.P. Robert, G. Casella (2010), *Introducing Monte Carlo Methods with R.* Springer.

[17] Hal S.Stern, Noel Cressie (1999), *Inference for Extremes in Disease Mapping.* Disease Mapping and Risk Assessment for Public Health, Wiley, 61-82

[18] Hal S. Stern, Noel Cressie (2000), *Posterior predictive model checks for disease mapping models.*. Statistics in medicine, **19**, 2377–2397.

[19] G. TOPA, Y. ZENOU (2015), *Neighbourhood Effects versus Network Effects.* Handbook of Regional and Urban Economics, **5**, 561-624.

[20] M. Sawada (2009), *Global Spatial Autocorrelation Indices - Moran's I, Geary's C and the General Cross-Product Statistic.* http://www.lpc.uottawa.ca/publications/moransi/moran.htm, University of Ottawa.

[21] *Julia Documentation.* www.julialang.org .

[22] R CORE TEAM (2014), *R: A Language and Environment for Statistical Computing.* URL http://www.R-project.org/, R foundation for Statistical Computing, Vienna, Austria

[23] STAN DEVELOPMENT TEAM (2015), *Stan Modeling Language, User's Guide and Reference Manual.* Stan version 2.8.0.

Ringraziamenti

"Se ho visto più lontano è perché sono salito sulle spalle dei giganti " (cit. Newton), ringrazio pertanto tutti i professori, in particola modo le professoresse Ilenia Epifani ed Alessandra Guglielmi che mi hanno seguito nel lavoro di tesi.

Ringrazio la mia famiglia e lo zio Gian.

Non sarebbe stato lo stesso senza i miei amici.