UNIVERSITY OF MILANO - BICOCCA

PHD IN STATISTICS - XXVIII

Latent Markov models for aggregate data: application to disease mapping and small area estimation

Candidate:
Gaia BERTARELLI

Supervisor:
Prof. M. Giovanna RANALLI
Internal Supervisor:
Prof. Fulvia MECATTI

A thesis submitted in fulfilment of the requirements of Doctor of Philosophy

in the

Department of Economics, Management and Statistics

"We'll make things right, we'll feel it all tonight
we'll find a way to offer up the night tonight
the indescribable moments of your life tonight
the impossible is possible tonight
believe in me as I believe in you, tonight."

Tonight, Tonight - Smashing pumpkins

A Federico.

UNIVERSITY OF MILANO - BICOCCA

Abstract

Department of Economics, Management and Statistics

Doctor of Philosophy

Latent Markov models for aggregate data: application to disease mapping and small area estimation

by Gaia BERTARELLI

Latent Markov Models (LMMs) are a particular class of statistical models in which a latent process is assumed. In studying LMMs, it is important to distinguish between two components: the measurement model, i.e. the conditional distribution of the response variables given the latent process, and the latent model, i.e. the distribution of the latent process. LMMs allow for the analysis of longitudinal data when the response variables measure common characteristics of interest, which are not directly observable. In LMMs the characteristics of interest, and their evolution in time, are represented by a latent process that follows a first order discrete Markov chain and units are allowed to change latent state over time. This thesis focuses on LMMs for aggregate data. It considers two fields of applications: disease mapping and small area estimation.

The goal of disease mapping is the study of the geographical pattern and variation of a disease measured through counts and incidence rates. From a methodological point of view, this work extends LMMs to include a spatial pattern in the latent model. This extension allows the probability of being in a latent state and the probability to move from a latent state to another over time to be influenced by the neighbouring areas. The model is fitted within a Bayesian framework using Gibbs and Random Metropolis-Hasting algorithms with augmented data that allows for a more efficient sampling of model parameters. Simulations studies are also conducted to investigate the performance of the proposed model on data generated under different settings. The model has also been applied to a data set of county specific lung cancer deaths counts in the state of Ohio, USA, during the years 1968-1988.

Small area estimation (SAE) methods are used in inference for finite populations to obtain estimates of parameters of interest when domain sample sizes are too small to provide adequate precision for direct domain estimators. The second work develops a new area-level SAE method using LMMs. In particular, since area-level SAE models consider a sampling and a linking model, a LMM is used as the linking model. In a hierarchical Bayesian framework the sampling model is introduced as the highest level of the hierarchy. In this context, data are considered aggregate because direct estimates are usually means and totals. Under the assumption of normality for the response variable, the model is estimated using a Gibbs sampling in a data augmentation context. The application field in this second work is particularly relevant: it uses yearly unemployment rates at Local Labour Market Areas level for the period 2004-2014 from the Labour Force Survey conducted by the Italian National Statistical Institute (ISTAT).

Acknowledgements

Trovare le parole per ringraziare coloro che mi hanno aiutata durante la realizzazione di questa tesi e negli ultimi tre anni rischia di essere complesso quanto la stesura di questo lavoro. Consapevole che non riuscirò a ringraziare ognuno nel modo in cui meriterebbe di essere ringraziato, devo comunque provare a riconoscere il merito a quelle persone senza le quali non sarei a questo punto in questo momento.

In primo luogo non posso che ringraziare la prof.ssa M. Giovanna Ranalli, esempio e modello costante: come ricercatrice, insegnante e persona.

Senza indugi devo i miei ringraziamenti alla prof.ssa Fulvia Mecatti, per le opportunità, le motivazioni ed il sostegno durante tutto questo percorso; al prof. Francesco Bartolucci per la sua disponibilità ed il suo sapere; al dott. Michele D'alò e al dott. Fabrizio Solari per il supporto e la disponibilità (tanta!!!!) nell'accontentare ogni mia richiesta e domanda riguardante i dati utilizzati in questa tesi e alla prof.ssa Paola Chiodini (capo!) per il suo essere una fine ascoltatrice e una saggia consigliera.

Agnese, che mi ha aiutata a camminare sempre a testa alta in questi anni e Igor, che con la sua determinazione e la sua forza mi ha spronata e motivata più di quanto lui possa immaginare, sono compagni di viaggio migliori di quelli che avrei sperato di incontrare prima di cominciare questo percorso. Grazie per avermi regalato la vostra amicizia prima che la vostra comprensione.

Un ringraziamento sincero, dovuto ma sopratutto voluto è per Daniele, Fabio, Gio', Giulio e Sara (rigorosamente in ordine alfabetico): la vera famiglia lontano da casa.

Grazie a Emanuela, Marcella, Nicola e Veronica, affidabile compagnia per le ore passate sui libri; a Edo, Marco, Silvia e Simone per aver trattato "quella di Milano" come se ci fosse sempre stata; a Francesca per aver sempre capito; ad Alberto, Arianna, Dede e i ragazzi di JSJ per le tante risate anche nei momenti più impensabili perché i sorrisi degli amici nascondono il vero supporto e a Raul, fratello silenzioso ma concreto.

Chiara, mamma e papà non posso ringraziarvi per il vostro incondizionato e costante amore, perché esso non sta nell' ambito delle cose, dei gesti e delle parole per le quali esista ringraziamento possibile. Voi, la certezza che non sarò mai sola, l'equilibrio di chi sa che sarà comunque sempre amata.

Contents

List of Figures

List of Tables

List of Abbreviations

AR(1)	first order **A**uto**R**egressive **S**tructure
ARE	**A**bsolute **R**elative **E**rror
ASE	**A**bsolute **S**quareD mean **E**rror
BYM	**B**esag, **Y**ork and **M**ollie model
EBLUP	**E**mpirical **B**est **L**inear **U**nbiased **P**rediction
FH	**F**ay Herriot
HB	**H**ierarchical **Bayesian**
HMM	**H**idden **M**arkov **Model**
HMRF	**H**idden **M**arkov **R**andom **F**ield
ISTAT	Italian National Statistical Institute
LAU	**L**ocal **A**dministrative **U**nit (LAU1: provincial level)
LCMs	**L**atent **C**lass **M**odels
LFS	**L**abour **F**orce **S**urvey
LMA	**L**abour **M**arket **A**reas
LMM	**L**atent **M**arkov **M**odel
MRF	**M**arkov **R**andom **F**ield
NSRA	**N**on **S**elf **R**epresenting **A**rea
NUTS	**N**omenclature des **U**nités **T**erritoriales **S**tatistiques (NUTS2: regional level)
PSU	**P**rimary **S**ampling **U**nit
SAE	**S**mall **A**rea **E**stimation
SAR	**S**imultaneous **A**utoregressive **M**odel
SRA	**S**elf **R**epresenting **A**rea
SSUs	**S**econdary **S**ampling **U**nit

Chapter 1

Introduction

Nowadays an always larger amount of data is often available thanks to modern tecnology and to planned longitudinal surveys. Even if this could be an advantage to gain pointwise information, occasionaly such data may not be easy to manage. Aggregate data can provide a feature rich data set with reduced dimensionality but without loosing too much information coming from longitudinal observation. Moreover such data could point out patterns and trends that would not normally be visible without long and sometimes not feasible computational procedures. In this thesis, Latent Markov models (LMMs) are applied to aggregate data. It considers two fields of applications where aggregated data have great relevance: disease mapping and small area estimation. In these fields longitudinal data are used to obtain information about areas, which are the subjects of interest.

Many models are proposed in the statistical literature for the analysis of longitudinal data. Among them, LMMs assume the existence of a latent process which affects the distribution of the response variables. The main assumption behind this approach is that the response variables are conditionally independent given this latent process, which follows a Markov chain with a finite number of states. The basic idea related to this assumption, which is referred to as *local independence*, is that the latent process fully explains the observable behavior of a subject together with possibly available covariates. Therefore, in studying LMMs, it is important to distinguish between two components: the *measurement model*, which concerns the conditional distribution of the response variables given the latent process, and the *latent model*, which concerns the distribution of the latent process.

LMMs can be applied in different kinds of analysis, e.g. they can asses the presence of measurement errors. In addition, they can account for unobserved heterogeneity between areas in the analysis including covariates in the measurement model which do not completely explain the heterogeneity in the response variable(s). The advantage of LMMs is that the effect of the unobservable variable has its own dynamics and it is not constrained to be time constant. Finally, through LMMs, a latent clustering of the population of interest can be pointed out. In fact, the latent states are identified as different subpopulations, with areas in the same subpopulation having a common distribution

for the response variable(s). In this context, a LMM may be seen as an extension of the latent class (LC) model (Lazarsfeld, Henry, and Anderson, 1968) in which areas are allowed to move between the latent classes during the period of observation. In this field, available covariates are included in the latent model and then may affect the initial and transition probabilities of the Markov chain.

In the second chapter of this thesis an application of LMMs to disease mapping in presented. The goal of disease mapping is to study the geographical pattern and variation of a disease measured through counts or incidence rates. It provides effective tools for identifying areas with potentially high risk, determining spatial trend, and formulating and validating hypotheses about the disease. Three aspects are relevant to disease mapping: computing accurate estimates of disease measures in small geographic areas, estimating the distribution of disease rates over the region and ranking the disease rates so that environmental investigation can be conducted. Our work wants to produce a classification useful to address all these aspects. From a methodological point of view, this work extends LMMs to include a spatial pattern in the latent model as an unobserved covariate of the latent states. This extension allows the probability of being in a latent state and the probability to move from a latent state to another over time to be influenced by the neighbouring areas. Moreover, in addition to create a classification of disease severity in the areas, the possibility of analysing the real influence and the significance of neighbour structure is admitted. The model is fitted within a Bayesian framework using Gibbs and Random Metropolis-Hasting algorithms with augmented data that allows for a more efficient sampling of model parameters. Simulation studies are conducted to investigate the performance of the proposed model on data generated under different settings. The model has then been applied to a data set of county specific lung cancer deaths counts in the state of Ohio, USA, during the years 1968-1988. Preliminary results of this work have been presented at the GEOMED conference in Florence, 9-11 September 2015.

The third chapter concerns the appliction of LMMs to small area estimation (SAE). SAE methods are used in inference for finite populations to obtain estimates of parameters of interest when domain sample sizes are too small to provide adequate precision for direct domain estimators. Unlike usual survey-sampling methods that treat each area's data independently, SAE models make assumptions that let areas "borrow strength" from each other and from longitudinal information. This usually leads to more precise and more stable estimates for the various areas. In this work, a new area-level SAE method using LMMs is introduced. In particular, since area-level SAE models consider a sampling and a linking model, a LMM is used as the linking model. In a hierarchical Bayesian framework the sampling model is introduced as the highest level of hierarchy. In this context, data are considered aggregate because direct estimates usually take the form of totlas or means (frequencies). Under the assumption of normality

for the response variable, the model is considered as a matched model and it is estimated using a Gibbs sampling in a data augmentation context. The proposed model is quite innovative because the definition of SAE methods which are able to take into account the non-observable nature of variables of interest is present in literature only in Fabrizi, Montanari, and Ranalli (2015), but the authors considere just the cross sectional nature of the problem without investigating its time extension. The application field in this second work is particularly relevant: it uses yearly direct unemployment rate estimates at Local Labour Market Areas level for the period 2004-2014 from the Labour Force Survey conducted by the Italian National Statistical Institute (ISTAT). Preliminary results of this second work have been presented at ITACOSM conference, Rome, 24-26 June 2015.

Chapter 2

Bayesian spatial latent Markov models for disease mapping

2.1 Introduction

The analysis of the geographical variation of a disease and its representation using maps are important tools to better understand its distribution in space. The observed cases of a particular disease are counted for each area in which the region under study is partitioned. These counts are then compared to the population size. Spatial dependency between counts has to be taken into account when analysing such data. Investigating the temporal pattern of a disease is also important to understand its evolution and trend. Most statistical methods for disease representation are based on risk mapping of aggregated data using in particular Poisson log-linear mixed models (Richardson et al., 1995; Mollié, 1999; Lawson et al., 2000) and the model proposed by Besag, York and Mollie (BYM) (Besag, York, and Mollié, 1991). The BYM model and its extensions (Clayton and Bernardinelli, 1992) are among the most popular approaches used in this context and use a Bayesian hierarchical modelling approach. BYM is based on an Hidden Markov Random Field where the latent risk field (the parameter of the Poisson distribution) is represented by a Markov random field with continuous state space modeled using Gaussian autoregressive spatial smoothing. Recent developments in this context include spatio-temporal mapping (Knorr-Held and Richardson, 2003; Robertson et al., 2010; Lawson and Song, 2010) and multivariate disease mapping (Knorr-Held, Raßer, and Becker, 2002).

The possibility of clustering the areas in different classes has the advantage of providing clearly delimited areas for different risk levels, which is helpful for decision makers to interpret the disease structure and enforce protection measures. These groups of areas can be viewed as spontaneous clusters (Knorr-Held, Raßer, and Becker, 2002), but we prefer to interpret them as incidence of disease or risk classes (Green and Richardson, 2002; Alfó, Nieddu, and Vicari, 2009). Latent Markov Models (LMMs) allow to obtain such classification and to consider different statistical distributions for

the response variable. In fact, the main difference between LMMs and clustering algorithms is that LMMs allow a "model-based clustering" approach that derives groups of observations using a probabilistic model that describes the distribution of the data. So, instead of finding clusters with some arbitrary chosen distance measure, LMMs use a model that assesses the probability that an area belongs to a particular class. So, we could say that it is a top-down approach (it starts with describing distribution of the data) while other clustering algorithms are rather bottom-up approaches (they find similarities between cases).

The basic LMM formulation is similar to that of hidden Markov models for time series data (MacDonald and Zucchini, 1997). In fact, a latent Markov chain, typically of first order, is used to represent the evolution of the latent characteristic over time. Moreover, the response variable observed at the different time points is assumed to be conditionally independent given this chain (assumption of local independence). The basic idea behind this assumption is that the latent process fully explains the observable behavior of an area. Furthermore, the latent state an area belongs to at a certain time only depends on the latent state at the previous time. A LMM may also be seen as an extension of the latent class model (LCM, Lazarsfeld, 1950; Lazarsfeld, Henry, and Anderson, 1968; Goodman, 1974), in which the assumption that each area belongs to the same latent class throughout the time period is suitably relaxed.

The basic LMM, relying on a homogenous Markov chain, has several extensions based on parameterizations that allow us to include hypotheses and constraints of interest. Generally speaking, these parameterizations may concern the conditional distribution of the response variable given the latent process (*measurement model*) and/or the distribution of the latent process (*latent model*). In this way, the introduction of covariates in the measurement or in the latent model is possible. When covariates are included in the measurement model, they affect the response variable and the latent process is seen as a way to account for unobserved heterogeneity between areas. The advantage with respect to a standard random effect model or a LCM with covariates is that the effect of unobservable covariates could be non constant over time and it could have its own dynamics. Differently, when covariates are included in the latent model, they influence initial and transition probabilities of the latent process. In this case we assume that the response variable measures and depends on the latent variable, which may evolve over time. In such case, the main research interest is in modelling the effect of the covariates on this latent variable distribution (Bartolucci, Lupparelli, and Montanari, 2009).

In our work, the spatial structure is modeled introducing defining a specific covariate into the latent model that influences the latent process and the resulting area classification. Due to the complexity of the resulting estimation we work in a Bayesian context and use Gibbs and Random Metropolis-Hasting algorithms with augmented data for efficently sampling model parameters. The paper is organized as follow: the

data are presented in Section 2.2. The model formulation and its estimation are presented in Sections 2.3 and 2.4, respectively. Section 2.5 presents an extensive simulation study, while the results of the application to the real data are provide in Section 2.6. Section 2.7 concludes with a summary discussion.

2.2 Data

We apply our method to a data set of county specific lung cancer deaths counts in the state of Ohio, USA. This dataset, originally studied in Devine (1992), contains information about the occurence of lung cancer deaths in 88 counties of Ohio, USA during the years 1968-1988. For each county the number of lung cancer deaths and people at risk are given for on each year. As we can see from the distribution of counts (fig. 2.1), incidence of death due to lung cancer in Ohio has dramatically increased during the period of observation.

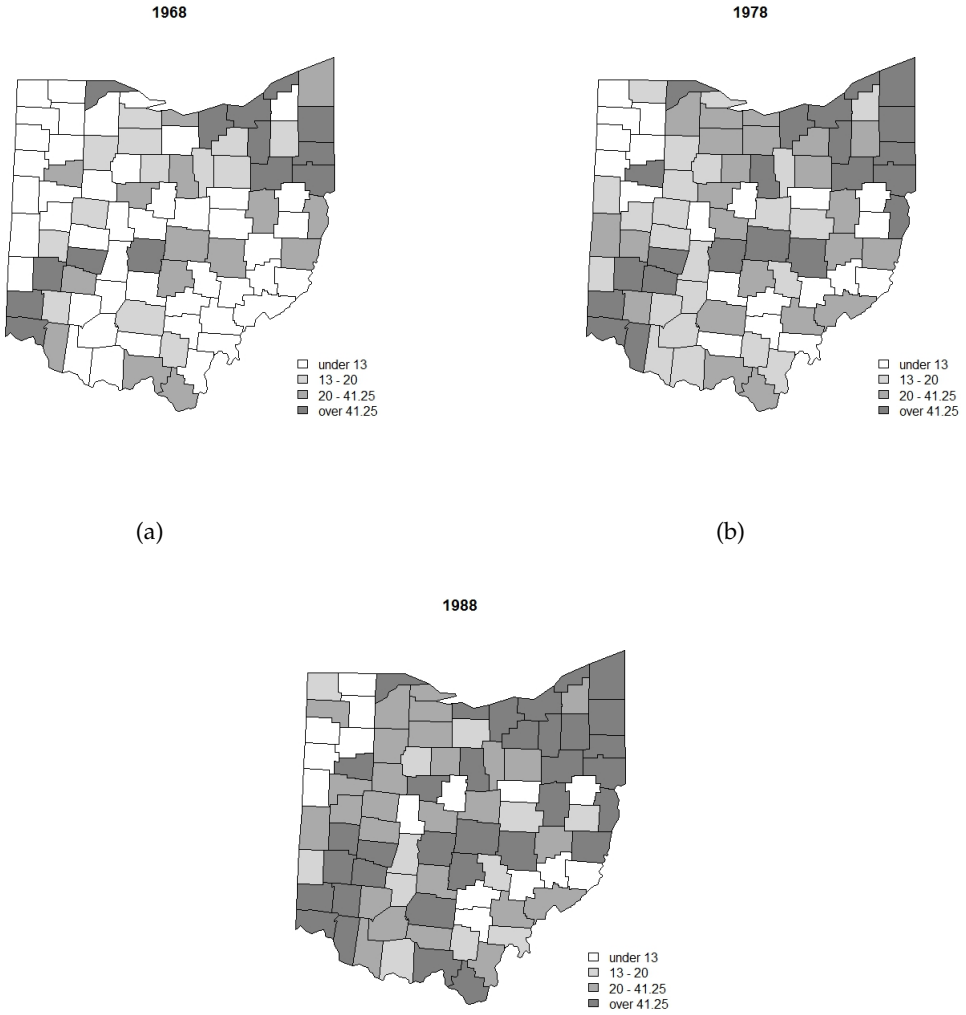
Data concerning the spatial structure of the counties is also available. In fact, for each county the index, the number of neighbouring counties and their corresponding indices are given. These informations are grouped in a neighbouring matrix $\boldsymbol{G}$ of dimension $n \times p$ where $n = 88$ is the number of counties and $p = 8$ is the the maximum number of neighbours considering all counties. As an example, we show the first six rows of $\boldsymbol{G}$, which represent counties labelled from 1 to 6:

$$\boldsymbol{G} = \begin{bmatrix} 8 & 36 & 66 & 73 & na & na & na & na \\ 6 & 32 & 33 & 69 & 81 & na & na & na \\ 38 & 39 & 42 & 47 & 52 & 70 & 85 & na \\ 28 & 43 & 78 & na & na & na & na & na \\ 37 & 53 & 58 & 64 & 82 & 84 & na & na \\ 19 & 2 & 33 & 46 & 54 & 75 & 81 & na \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix}. \tag{2.1}$$

The first row indicates that county labelled with number 1 has 4 neighbours and they are counties labelled as 8, 36, 66 and 73. Missing values na are reported in the matrix when there are not other neighbouring counties. The second row indicates that county labelled with number 2 has 5 neighbours and they are counties labelled as 6, 32, 33, 69, 81. Note that county 2, as a consequence, shows up in the sixth row.

2.3 Model

In a LMM the existence of two processes is assumed: an unobservable finite-state first-order Markov chain $U_i^{(t)}$, $i = 1, \dots, n$ and $t = 1, \dots, T$ with state space $\{1, \dots, m\}$ and



(a) (b)

(c)

FIGURE 2.1: Lung cancer deaths in Ohio 1968-1978-1988.

an observed process $Y_i^{(t)}$, $i = 1, \ldots, n$ and $t = 1, \ldots, T$, where $Y_i^{(t)}$ denotes the response variable for area i at time t and similary for $U_i^{(t)}$. We assume that the distribution of $Y_i^{(t)}$ depends only on $U_i^{(t)}$; specifically the $Y_i^{(t)}$, $t = 0, \ldots, T$ are conditionally independent given the $U_i^{(t)}$. We also denote by $\tilde{\boldsymbol{U}}_i^{(t)}$ the vector of latent states for the neighbours of area i at occasion t.

The state-dependent distribution, i.e. the distribution of $Y_i^{(t)}$ given $U_i^{(t)}$, can be either a continuous or a discrete distribution. It can be taken from the exponential family such as the Binomial, the Poisson or the Normal distribution. Thus, the unknown vector of parameters $\boldsymbol{\phi}$ in a LMM includes both the parameters of the Markov chain $\boldsymbol{\phi}_{lat}$ and the vector of parameters of the state-dependent distribution of $Y_i^{(t)}$ conditionally

on $U_i^{(t)}$, $\boldsymbol{\phi}_{obs}$. The *measurement model* involves $\boldsymbol{\phi}_{obs}$ and it can be written as

$$Y_i^{(t)}|U_i^{(t)} \sim f(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\phi}_{obs}).$$

The *latent model* includes two sets of parameters. The parameters $\boldsymbol{\phi}_{lat}$ of the Markov chain are the elements of the transition probability matrix

$$\boldsymbol{\Pi} = \{\pi_{u|\bar{u}}\} \text{ , with } u, \bar{u} = 1, \ldots, m;$$

where

$$\pi_{u|\bar{u}} = P(U_i^{(t)} = u|U_i^{(t-1)} = \bar{u})$$

is the probability that area i visits state u at time t given that at time $t-1$ it was in state $\bar{u}$, and the vector of initial probabilities

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_u, \ldots, \pi_m)'$$

where

$$\pi_u = P(U_i^{(1)} = u)$$

is the probability of being in state u at the initial time for $u = 1, \ldots, m$. For the sake of simplicity, in this work we consider homogeneous LMMs, i.e. LMMs where the transition probability matrix is constant as a function of time t, but the hidden markov chain could also be considered non homogeneous and the transition probabilities time-varying.

Spatial structure in this work is introduced as a covariate in the *latent model* based on the latent structure and it depends on the neighbouring matrix $\boldsymbol{G}$ from equation (2.1) which is fixed in time. Let $\tilde{\boldsymbol{u}}_i^{(t)}$ denote the particular configuration of latent states for the neighbouring counties of county i at time t. We consider a generic function $\boldsymbol{\eta}$ of $\tilde{\boldsymbol{u}}_i^{(t)}$ as a time-varying covariate affectecting $\boldsymbol{\phi}_{lat}$, i.e. the initial and transition probabilities, through the following parameterizations:

$$\log \frac{p(U_i^{(1)} = u|\tilde{\boldsymbol{U}}_i^{(1)} = \tilde{\boldsymbol{u}}_i^{(1)})}{p(U_i^{(1)} = 1|\tilde{\boldsymbol{U}}_i^{(1)} = \tilde{\boldsymbol{u}}_i^{1})} = \beta_{0u} + \boldsymbol{\eta}[\tilde{\boldsymbol{u}}_i^{(1)}]'\boldsymbol{\beta}_{1u} \text{ with } u \geq 2, \tag{2.2}$$

$$\log \frac{p(U^{(t)} = u|U^{t-1} = \bar{u}, \tilde{\boldsymbol{U}}_i^{(t)} = \tilde{\boldsymbol{u}}_i^{(t)})}{p(U^{(t)} = \bar{u}|U^{t-1} = \bar{u}, \tilde{\boldsymbol{U}}_i^{(t)} = \tilde{\boldsymbol{u}}_i^{(t)})} = \gamma_{0u\bar{u}} + \boldsymbol{\eta}(\tilde{\boldsymbol{u}}_i^{(t)})'\boldsymbol{\gamma}_{1u\bar{u}}, \quad \text{with } t \geq 2 \text{ and } u \neq \bar{u} \tag{2.3}$$

where $\boldsymbol{\beta_u} = (\beta_{0u}, \boldsymbol{\beta}_{1u}^{'})'$ and $\boldsymbol{\gamma}_{u\bar{u}} = (\gamma_{0u\bar{u}}, \boldsymbol{\gamma}_{1u\bar{u}}^{'})'$ are vectors of parameters to be estimated because they are part of $\boldsymbol{\phi}_{lat}$. There are different possible choices of $\boldsymbol{\eta}$ in this

context: e.g. it could be the mean of the elements in $\tilde{\boldsymbol{u}}_i^{(t)}$, its mode, the relative frequencies of the neighbouring latent states or the sum of the latent states realizations in the neighbours. The best choice of $\boldsymbol{\eta}$ depends on the application of interest. We will study the effect of alternativr choices in the simulation studies, and in the application on the Ohio data.

2.4 Model Estimation

We adopt the principle of data augmentation (Tanner and Wong, 1987) in which the latent states are introduced as missing data and augmented to the state of the sampler (Germain, 2010). In this way we can simplify the process of sampling from the posterior: we can use a Gibbs sampler for the parameters of the measurement model and we can estimate the initial and the transition probabilities by a Random Metropolis-Hasting step. We introduce a system of priors for the unknow model parameters $\boldsymbol{\phi}$. We adopt common non-informative prior distributions. In particular, for $\boldsymbol{\Pi}$ and $\boldsymbol{\pi}$ we use the uniform distribution $U(0,1)$, while for the vectors $\boldsymbol{\beta_u}$ and $\boldsymbol{\gamma}_{u\bar{u}}$ we assume that they are a priori independent with distribution $N(0,\sigma_\beta^2\boldsymbol{I})$ and $N(0,\sigma_\gamma^2\boldsymbol{I})$, respectively. The choice for σ_β^2 and σ_γ^2 depends on the context of the application. Tipically $5 \leq \sigma_\beta^2 = \sigma_\gamma^2 \leq 10$. The prior distribution for the parameters of the measurement model $\boldsymbol{\phi}_{obs}$ depends on the distribution assumed for the state-dependent distribution.

One of the goals of model inference is to estimate the set of the latent variables $\boldsymbol{u}$. In this context there is the possibility of choosing priors which are conjugate to the form of the complete data likelihood, therefore sampling from the conditional posterior of the model parameters given the latent states (the so called complete data posterior) is straightforward. Moreover, because the state space is discrete and finite, sampling from the conditional posterior of the latent states given the model parameters is also possible. So it is possible to generate samples from the joint posterior distribution of the model parameters and latent states as follows. Let $\boldsymbol{y}$ be the set of realizations of the response variable. Given

$$\pi(\boldsymbol{\theta},\boldsymbol{u}|\boldsymbol{y}) = \pi(\boldsymbol{\theta})p(\boldsymbol{u}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{u}) \tag{2.4}$$

$$\text{joint posterior} \;=\; \text{prior} \;\times\; \text{likelihood} \;\times\; \text{likelihood}\,,$$

samples can be generated alternating between sampling $\boldsymbol{u}$ from the conditional posterior distribution $\pi(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\phi})$ and drawing $\boldsymbol{\phi}$ from the conditional posterior distribution $\pi(\boldsymbol{\phi}|\boldsymbol{y},\boldsymbol{u})$. When a priori independence is assumed between $\boldsymbol{\phi}_{obs}$ and $\boldsymbol{\phi}_{lat}$, the complete data posterior can be written as

$$\pi(\boldsymbol{\theta}|\boldsymbol{u},\boldsymbol{y}) = \pi(\boldsymbol{\theta}_{obs}|\boldsymbol{u},\boldsymbol{y})\pi(\boldsymbol{\theta}_{lat}|\boldsymbol{u},\boldsymbol{y}) \tag{2.5}$$

and the MCMC sampling scheme leads to repeat for iterations $b = 1, \ldots, B$ the following steps:

1. Simulate $\boldsymbol{u}^b$ from $\pi(\boldsymbol{u}|\boldsymbol{\phi}^b, \boldsymbol{y})$.
2. Simulate $\boldsymbol{\phi}^b$ from $\pi(\boldsymbol{\phi}|\boldsymbol{u}^{b-1}, \boldsymbol{y})$ where:
   (a) $\boldsymbol{\phi}^b_{lat}$ is simulated from $\pi(\boldsymbol{\phi}_{lat}|\boldsymbol{u}^{b-1})$;
   (b) $\boldsymbol{\phi}^b_{obs}$ is simulated from $\pi(\boldsymbol{\phi}_{obs}|\boldsymbol{u}^{b-1}, \boldsymbol{y})$.

If each $\boldsymbol{y}_i$ is assumed independent, $\boldsymbol{u}_i$ can be sampled individually using a Gibbs sampler from the posterior and they can be drawn from

$$U_i^{(t)} \sim Mult(p(u_i^{(t)} = 1|\boldsymbol{y}^{(t-1)}, \boldsymbol{y}^{(t+1)}, \boldsymbol{\theta}), \ldots, p(u_i^{(t)} = m|\boldsymbol{y}^{t-1}, \boldsymbol{y}^{(t+1)}, \boldsymbol{\theta})). \tag{2.6}$$

The complete data posterior distribution is given by the Bayes Theroem as

$$\pi(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{u}) \propto \pi(\boldsymbol{\phi})p(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\phi}).$$

If a priori independence is assumed between $\boldsymbol{\phi}_{obs}$ and $\boldsymbol{\phi}_{lat}$, given $\boldsymbol{u}$ these parameters remain conditionally independent a posteriori and the complete data posterior can be decomposed in

$$\pi(\boldsymbol{\phi}_{obs}|\boldsymbol{y}, \boldsymbol{u}) \propto \pi(\boldsymbol{\phi}_{obs})p(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\phi}_{obs}) \tag{2.7}$$

and

$$\pi(\boldsymbol{\phi}_{lat}|\boldsymbol{y}, \boldsymbol{u}) \propto \pi(\boldsymbol{\phi}_{lat})p(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\phi}_{lat}). \tag{2.8}$$

The overall form of the complete data posterior distribution $\pi(\boldsymbol{\phi}_{obs}|\boldsymbol{y}, \boldsymbol{u})$ is specific to a particular latent Markov model. If the prior for a component of $\boldsymbol{\phi}_{obs}$ is conjugate to the form of the complete data likelihood, than the full conditional distribution belongs to the same family of distributions as the prior and can be sampled directly with a Gibbs sampler. Otherwise it is necessary to introduce Random Metropolis-Hasting steps. To estimate $\boldsymbol{\phi}_{lat} = \{\boldsymbol{\beta}_{\boldsymbol{u}}, \boldsymbol{\gamma}_{\boldsymbol{u}\bar{\boldsymbol{u}}}\}$ when we introduce the spatial structure, we need a Metropolis-Hasting step because we introduce a non-linear link function.

To clarify the estimation of the vector of parameters $\boldsymbol{\phi}$ we consider a binomial state-dependent distribution. This is the choice for the data set of lung cancer deaths in Ohio presented in Section 3.2. In this case

$$\begin{aligned} Y_i^{(t)}|U_i^{(t)} &\sim Bin(r_i^{(t)}, p_u) \\ &\text{for } u = 1, \ldots, m \end{aligned} \tag{2.9}$$

where $r_i^{(t)}$ is the number of people at risk for area i at time t and p_u is the succes probability given a specific latent state. We assume that each component of $\phi_{obs} = \{p_1 \ldots, p_m\}$ has a Uniform $U(0, 1)$ prior. In this way the vector $\boldsymbol{p} = (p_1, \ldots, p_m)'$ can be

simulated from a Dirichlet distribution such that

$$\boldsymbol{p} \sim Dir(\boldsymbol{\alpha}, \boldsymbol{\xi}) \tag{2.10}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ and $\alpha_u = 1 + \sum_t \sum_i y_i^{(t)} I(U_i^{(t)} = u)$ with $u = 1, \ldots, m$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$ and $\xi_u = 1 + \sum_t \sum_i s_i^{(t)} I(U_i^{(t)} = u)$ where $s_i^{(t)}$ is the number of subjects who are not dead for lung cancer in area i at time t. Latent parameters $\boldsymbol{\phi}_{lat} = \{\boldsymbol{\beta_u}, \boldsymbol{\gamma_{u\bar{u}}}\}$ have Normal priors with mean zero and fixed variance but their full conditional distribution can not be written directly because it depends on the logit parametrization and on the covariates so we have to use a Random Metropolis-Hasting sampler.

The choice of the number of latent states of the unobserved Markov chain underlying the observed data corresponds to the model selection procedure. This procedure is very important in the estimation process. We use a model selection method based on the choice of the maximum likelihood. Consider the marginal likelihood of the data $\boldsymbol{y}$. For any parameter configuration $\bar{\boldsymbol{u}}^{(1)}, \ldots, \bar{\boldsymbol{u}}^{(T)}, \overline{\phi}$, Bayes' rule implies that the marginal likelihood of the data $\boldsymbol{y}$ satisfies

$$p(\boldsymbol{y}) = \frac{p(\bar{\boldsymbol{u}}^{(1)}, \ldots, \bar{\boldsymbol{u}}^{(T)}, \overline{\phi}, \boldsymbol{y})}{p(\bar{\boldsymbol{u}}^{(1)}, \ldots, \bar{\boldsymbol{u}}^{(T)}, \overline{\phi}|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\bar{\boldsymbol{u}}^{(1)}, \ldots, \bar{\boldsymbol{u}}^{(T)}, \overline{\phi}) p(\bar{\boldsymbol{u}}^{(1)}, \ldots, \bar{\boldsymbol{u}}^{(T)}, \overline{\phi})}{p(\bar{\boldsymbol{u}}^{(1)}, \ldots, \bar{\boldsymbol{u}}^{(T)}, \overline{\phi}|\boldsymbol{y})}; \tag{2.11}$$

where $\bar{\boldsymbol{u}}^{(t)}$, with $t = 1, \ldots, T$ is a fixed $n \times 1$ vector of latent states and $\overline{\phi}$ a vector of fixed parameters. The numerator in (3.29) can be calculated but the posterior probability at denominator can be notoriously difficult to compute, and particularly so in high dimensional problems. To do that we use the Chib and Jeliazkov (2001) estimator. They employ the local reversibility of the Metropolis-Hasting algorithm to construct an estimator in models where full conditional densities are not available analytically but which does not require the normalising constant of $p(\bar{\boldsymbol{u}}^{(1)}, \ldots, \bar{\boldsymbol{u}}^{(T)}, \overline{\phi}|\boldsymbol{y})$. The estimator is free of distributional assumptions and is directly linked to the simulation algorithm.

A well-know problem occurring in Bayesian modeling is that of label switching in which random permutations of the hidden state labels occur over the course of the MCMC run. That can be seen as the non-identifiability of the components due to the invariance of the posterior distribution to the permutation in the parameters labeling. Several solutions have been proposed in the literature (Jasra, Holmes, and Stephens, 2005). We relabelled the MCMC output at every iteration following the ascending order of the parameters which affect the initial probabilities vector.

2.5 Simulation Studies

In the following we illustrate simulation results aimed at evaluating the behaviour of the proposed model. We want to study how the model can identify the underlying latent distribution and fit the parameters with different number of states, state-dependent distribution and function $\boldsymbol{\eta}(\cdot)$. In order to do that, we consider five simulation scenarios:

- Scenario 1a: Binomial distribution with $m = 2$ latent states and $\boldsymbol{\eta}(\cdot) =$ mean,
- Scenario 1b: Binomial distribution with $m = 3$ latent states and $\boldsymbol{\eta}(\cdot) =$ mean,
- Scenario 2a: Binomial distribution with $m = 2$ latent states and $\boldsymbol{\eta}(\cdot) =$ relative frequencies,
- Scenario 3a: Normal distribution with $m = 2$ latent states and $\boldsymbol{\eta}(\cdot) =$ mean,
- Scenario 3b: Normal distribution with $m = 3$ latent states and $\boldsymbol{\eta}(\cdot) =$ mean.

For each scenario we run three chains and then consider their mean.

2.5.1 Scenario 1a

Scenario 1a concerns the use of the binomial distribution as in (2.9). In this first scenario we consider $m = 2$ latent states and the spatial function covariate considered in the model is the mean of the neighbouring latent states for each area i at time t. We consider $n = 88$ units and $T = 21$ times as in the Ohio data. As neighbouring matrix we consider the matrix (2.1) which comes from the Ohio data. Following (2.2) and (2.3), with $k = 2$ latent states and $q = 1$ covariate on the initial and the transition probabilities, eight parameters have to be estimated. We fix these to be:

- $\boldsymbol{\beta}_1^{'} = (\beta_{01}, \beta_{11}) = (-1.5, 1)$, where $u = 1$ is the reference group,
- $\boldsymbol{\gamma}_{\boldsymbol{u\bar{u}}}^{'} = (\gamma_{012}, \gamma_{021}, \gamma_{21}, \gamma_{12}) = (-2, 1, 0.5, -1)$, where $\bar{u}$ is the latent state at time $(t-1)$,
- $\boldsymbol{p}^{'} = (p_1, p_2) = (0.0025, 0.005)$.

With these parameters we simulate a scenario where, considering the severity of the condition, an increase of the mean influences the area state to a more severe condition and where there is a balanced division of areas in the two states. Based on these parameters we simulate $U_i^{(t)}$ from its full conditional as

$$U_i^{(t)} \sim Mult(p(U_i^{(t)} = 1|\boldsymbol{y}^{(t-1)}, \boldsymbol{y}^{(t+1)}, \boldsymbol{\theta}), \ldots, p(U_i^{(t)} = m|\boldsymbol{y}^{t-1}, \boldsymbol{y}^{(t+1)}, \boldsymbol{\theta}))$$

and then

$$y_i^{(t)}|U_i^{(t)} = u_i^{(t)} \sim Bin(r_i^{(t)}, p_u)\ i =, 1, \ldots, 88; t = 1, \ldots, 21; u = 1, 2; \tag{2.12}$$

where $r_i^t \sim U(10000, 1800000)$. In this way the underlying latent distribution is divided as follow: 957 areas are classified in state $u = 1$, 891 areas are allocated in state $u = 2$.

To estimate $\boldsymbol{\beta}_u$ and $\boldsymbol{\gamma}_{u\bar{u}}$ a random Metropolis-Hasting is used, with $N(0, 1)$ as random walk and Normal proposal with mean zero and standard deviation equal to three. In the Metropolis-Hasting we obtain an acceptance rate of 21.64% for $\boldsymbol{\beta}_u$ and 20.64% for $\boldsymbol{\gamma}_{u\bar{u}}$. The parameter vector of the manifest distribution is sampled using a Gibbs sampler from its full conditional distribution $\boldsymbol{p} \sim Dir(\boldsymbol{\alpha}, \boldsymbol{\xi})$ where $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$ are defined after (2.10).

TABLE 2.1: Scenario 1a: Binomial $m = 2$ and $\boldsymbol{\eta}(\cdot)$ mean. State $u = 1$ distribution. Real value 957.

u=1	953	954	955	956	957	958	959	960
Freq	3	74	1734	12111	5284	743	51	1

TABLE 2.2: Scenario 1a: Binomial $m = 2$ and $\boldsymbol{\eta}(\cdot)$ mean. Parameter estimates and time series S.E. (Burn-in=20000).

$\widehat{\beta_{01}}$	$S.E.(\widehat{\beta_{01}})$	$\widehat{\beta_{11}}$	$S.E.(\widehat{\beta_{11}})$	$\widehat{\gamma_{012}}$	$S.E.(\widehat{\gamma_{012}})$	$\widehat{\gamma_{021}}$	$S.E.(\widehat{\gamma_{021}})$
-1.5101	0.0819	0.9782	0.0515	-2.0001	0.0215	0.3687	0.0227
$\widehat{\gamma_{21}}$	$S.E.(\widehat{\gamma_{21}})$	$\widehat{\gamma_{12}}$	$S.E.(\widehat{\gamma_{12}})$	$\widehat{p_1}$	$S.E.(\widehat{p_1})$	$\widehat{p_2}$	$S.E.(\widehat{p_2})$
0.7365	0.0144	-0.8133	0.0153	0.0025	3.100e-08	0.0049	5.153e-08

2.5.2 Scenario 1b

In this second scenario we consider $m = 3$ latent states and the spatial function covariate is considered in the model as the mean of the neighbouring latent states for each area at time t. We again consider $n = 88$ units and $T = 21$ times as in the Ohio data. As neighbouring matrix we consider the matrix in (2.1). Following (2.2) and (2.3), with $m = 3$ latent states and $q = 1$ covariate on the initial and the transition probabilities, 19 parameters have to be estimated. We fix these to be

- $\boldsymbol{\beta}_1^{'} = (\beta_{01}, \beta_{11}, \beta_{22}, \beta_{33}) = (1, -3.5, 3.5, 4)$, where $u = 1$ is the reference group,
- $\boldsymbol{\gamma}_{\boldsymbol{u\bar{u}}}^{'} = (\gamma_{021}, \gamma_{031}, \gamma_{012}, \gamma_{032}, \gamma_{013}, \gamma_{023}, \gamma_{21}, \gamma_{31}, \gamma_{12}, \gamma_{32}, \gamma_{13}, \gamma_{23}) =$
  $= (1, 1, -0.5, 1, -0.5, 0.5, 0.5, 0.5, -0.5, 0.5, -0.5 - 0.5)$, where $\bar{u}$ is the latent state at time $(t - 1)$
- $\boldsymbol{p}^{'} = (p_1, p_2, p_3) = (0.002, 0.003, 0.006)$

Then, the latent variable has the following distribution: over all the period of observation 425 area are in state $u = 1$, 641 are in $u = 2$ and 782 in $u = 3$.

TABLE 2.3A: Scenario 1b: Binomial distribution $m = 3$ and $\boldsymbol{\eta}(\cdot)$ mean. State $u = 1$ distribution. Real value 425.

u=1	418	419	420	421	422	423	424	425	426
Freq	1	7	87	629	2424	4731	1921	187	13

TABLE 2.3B: Scenario 1b: Binomial distribution $m = 3$ and $\boldsymbol{\eta}(\cdot)$ mean. State $u = 2$ distribution. Real value 641.

u=2	0	630	631	632	633	634	635	636	637	638
Freq	1	10	22	76	206	397	751	1164	1571	1718
u=2	639	640	641	642	643	644	645	646	647	648
Freq	1561	1194	744	369	130	63	14	7	2	1

TABLE 2.3C: Scenario 1b: Binomial distribution $m = 3$ and $\boldsymbol{\eta}(\cdot)$ mean. State $u = 3$ distribution. Real value 782.

u=3	0	777	779	780	781	782	783	784	785
Freq	1	1	1	10	30	92	242	624	1111
786	787	788	789	790	791	792	793	794	795
1588	1881	1731	1235	798	391	176	56	26	7

TABLE 2.4: Scenario 1b: Binomial distribution $m = 3$ and $\boldsymbol{\eta}(\cdot)$ mean. Latent parameters estimates and time series S.E.

$\widehat{\beta}_{01}$	$S.E.(\widehat{\beta}_{01})$	$\widehat{\beta}_{11}$	$S.E.(\widehat{\beta}_{11})$	$\widehat{\beta}_{22}$	$S.E.(\widehat{\beta}_{22})$	$\widehat{\beta}_{33}$	$S.E.(\widehat{\beta}_{33})$
0.822	0.008	-0.905	0.011	3.968	0.027	1.323	0.02
$\widehat{\gamma}_{021}$	$S.E.(\widehat{\gamma}_{021})$	$\widehat{\gamma}_{031}$	$S.E.(\widehat{\gamma}_{031})$	$\widehat{\gamma}_{012}$	$S.E.(\widehat{\gamma}_{012})$	$\widehat{\gamma}_{032}$	$S.E.(\widehat{\gamma}_{032}$
1.354	0.007	1.875	0.003	-0.639	0.003	1.853	0.007
$\widehat{\gamma}_{013}$	$S.E.(\widehat{\gamma}_{012})$	$\widehat{\gamma}_{023}$	$S.E.(\widehat{\gamma}_{022})$	$\widehat{\gamma}_{21}$	$S.E.(\widehat{\gamma}_{21})$	$\widehat{\gamma}_{31}$	$S.E.(\widehat{\gamma}_{31})$
-0.154	0.001	0.292	0.003	0.311	0.001	0.177	0.001
$\widehat{\gamma}_{12}$	$S.E.(\widehat{\gamma}_{12})$	$\widehat{\gamma}_{32}$	$S.E.(\widehat{\gamma}_{32})$	$\widehat{\gamma}_{13}$	$S.E.(\widehat{\gamma}_{13})$	$\widehat{\gamma}_{23}$	$S.E.(\widehat{\gamma}_{23})$
-0.022	0.002	0.917	0.003	-0.123	0.004	-0.933	0.001

TABLE 2.5: Scenario 1b: Binomial distribution $m = 3$ and $\boldsymbol{\eta}(\cdot)$ mean. Manifest parameters estimates and time series S.E.

$\widehat{p_1}$	$S.E.(\widehat{p_2})$	$\widehat{p_2}$	$S.E.(\widehat{p_2})$	$\widehat{p_3}$	$S.E.(\widehat{p_3})$
0.003	5.603e-08	0.005	7.203e-08	0.006	5.692e-08

2.5.3 Scenario 2a

In Scenario 2a we consider a Binomial state-dependent distribution with $m = 2$ latent states and the spatial function covariate $\boldsymbol{\eta}(\cdot)$ is considered in the model as the relative frequencies of the neighbouring latent states for each area at time t. In this way m covariates are introduced in the model. The first covariate is given by the relative frequencies of neighbouring areas with estimates $u = 1$, the last covariate is the relative frequencies of neighbouring area with estimated latent state $u = m$. With the same aim of Scenario 1, for $m = 2$ we fix the 11 parameters as follow:

- $\boldsymbol{\beta}_1^{'} = (\beta_{01}, \beta_{11_1}, \beta_{11_2}) = (-3.5, -3.5, 0.5)$;
- $\boldsymbol{\gamma}_{\boldsymbol{u\bar{u}}}^{'} = (\gamma_{012}, \gamma_{021}, \gamma_{21_1}, \gamma_{21_2}, \gamma_{12_1}, \gamma_{12_2}) = (-1.5, 0.5, -1, 1, 1.5, -0.5)$;
- $\boldsymbol{p}^{'} = (p_1, p_2) = (0.003, 0.007)$.

Then, we simulate with a Gibbs sampler the distribution of the latent states for every $i = 1, \ldots, 88$ at every $t = 1, \ldots, 21$ and then we generate the state space distribution as in eq. (2.12). To estimate $\boldsymbol{\beta}_u$ and $\boldsymbol{\gamma}_{u\bar{u}}$ a random Metropolis-Hasting is used, with $N(0, 1)$ as random walk and Normal proposal of mean zero and standard deviation equal to six. In the Metropolis-Hasting we obtain an acceptance rate of 22.6% for $\boldsymbol{\beta}_u$ and 46.1% for $\boldsymbol{\gamma}_{u\bar{u}}$. Latent states are divided as follows: 931 areas are assigned to $u = 1$ and 916 to $u = 2$.

TABLE 2.6: Scenario 2a: Binomial distribution $m = 2$ and $\boldsymbol{\eta}(\cdot)$ relative frequencies. State $u = 1$ distribution. Real value 931.

u=1	917	918	919	929	921	922	923	924
Freq	1	19	117	363	958	1788	2211	
u=1	925	926	927	928	929	939	931	1793
Freq	1411	689	256	57	11	3	1	1

TABLE 2.7: Scenario 2a: Binomial distribution $m = 2$ and $\boldsymbol{\eta}(\cdot)$ relative frequencies. Latent parameters estimates and time series S.E.

$\widehat{\beta}_{01}$	$S.E.(\widehat{\beta}_{01})$	$\widehat{\beta}_{11_1}$	$S.E.(\widehat{\beta}_{11_1})$	$\widehat{\beta}_{11_2}$	$S.E.(\widehat{\beta}_{11_2})$
-3.5310	0.0201	-3.5310	0.0228	0.3486	0.1122
$\widehat{\gamma}_{012}$	$S.E.(\widehat{\gamma}_{012})$	$\widehat{\gamma}_{021}$	$S.E.(\widehat{\gamma}_{021})$	$\widehat{\gamma}_{21_1}$	$S.E.(\widehat{\gamma}_{21_1})$
-1.5494	0.0037	0.5534	0.0038	-0.8984	0.0117
$\widehat{\gamma}_{21_2}$	$S.E.(\widehat{\gamma}_{21_2})$	$\widehat{\gamma}_{12_1}$	$S.E.(\widehat{\gamma}_{12_1})$	$\widehat{\gamma}_{12_2}$	$S.E.(\widehat{\gamma}_{12_2})$
1.0946	0.0129	1.4505	0.0037	-0.4466	0.0033

TABLE 2.8: Scenario 2a: Binomial distribution $m = 2$ and $\boldsymbol{\eta}(\cdot)$ relative frequencies. Manifest parameters estimates and time series S.E. (Burn-in=20000).

$\widehat{p}_1$	$S.E.(\widehat{p}_1)$	$\widehat{p}_2$	$S.E.(\widehat{p}_2)$
0.003	5.381e-08	0.0069	1.609e-07

2.5.4 Scenario 3a

Following the aim of testing the adaptability of the proposed method to different state-dependent distributions, we simulate a sample from a Normal distribution. In this first scenario we consider $m = 2$ latent states and the spatial function $\boldsymbol{\eta}(\cdot)$ is in the model as the mean of the neighbouring latent states for each area at time t. We consider $n = 88$ units and $T = 21$ times as in the Ohio data. As neighbouring matrix we consider the matrix (2.1) which comes from the Ohio data. In this case, the state-dependent distribution is

$$\begin{aligned} Y_i^{(t)}|U_i^{(t)} &\sim N(\mu_u, \sigma_u) \\ &\text{for } u = 1, 2. \end{aligned} \tag{2.13}$$

We simulate a scenario where, considering the severity of the condition, an increase of the mean influences the area state to more severe condition and where there is a balanced division of areas in the two states. For $m = 2$ we fix the parameters to:

- $\boldsymbol{\beta}_1^{'} = (\beta_{01}, \beta_{11}) = (-1.5, 1)$, where $u = 1$ is the reference group;
- $\boldsymbol{\gamma}_{\boldsymbol{u\bar{u}}}^{'} = (\gamma_{012}, \gamma_{021}, \gamma_{21}, \gamma_{12}) = (-2, 1, 0.5, -1)$, where $\bar{u}$ is the latent state at time $(t-1)$ and we have differen refereces group;
- $\boldsymbol{\mu}^{'} = (\mu_1, \mu_2) = \{10, 20\}$;
- $\boldsymbol{\sigma}^{2^{'}} = (\sigma_1, \sigma_2) = (1, 1)$.

In this way, 957 areas are in state $u = 1$ and 891 in $u = 2$. We estimate the parameters of the latent model with a Metropolis-Hasting algorithm , with $N(0, 1)$ as random walk and Normal proposal of mean zero and standard deviation equal to three. The Gibbs full conditional distributions are

- $\mu_u \sim \mathrm{N}(\overline{Y}, \sigma_u^2/n_u)$;
- $\tau \sim \mathrm{InvGamma}\Big(\frac{n_u}{2}, \frac{1}{2}\sum_{i=1}^{n_u}\sum_{t=1}^{T}(Y_i^t - \mu_u)^2\Big)$;

where n_u is the number of areas classified in state u over all time points. The acceptance rate for $\boldsymbol{\beta}$ is 26.52% and for $\boldsymbol{\gamma}$ is 21.84%.

TABLE 2.9: Scenario 3a: Normal distribution $m = 2$ and $\boldsymbol{\eta}(\cdot)$ mean. State $u = 1$ distribution. Real value 957.

u=1	0	955	957	958
Freq	1	51	9546	2

TABLE 2.10: Scenario 3a: Normal distributionl $m = 2$ and $\boldsymbol{\eta}(\cdot)$ mean. Latent parameters estimates and time series S.E.

$\widehat{\beta}_{01}$	$S.E.(\widehat{\beta}_{01})$	$\widehat{\beta}_{11}$	$S.E.(\widehat{\beta}_{11})$	$\widehat{\gamma}_{012}$	$S.E.(\widehat{\gamma}_{012})$
-0.986	0.034	0.728	0.022	-1.769	0.063
$\widehat{\gamma}_{021}$	$S.E.(\widehat{\gamma}_{021})$	$\widehat{\gamma}_{21}$	$S.E.(\widehat{\gamma}_{21})$	$\widehat{\gamma}_{12}$	$S.E.(\widehat{\gamma}_{12})$
0.46	0.04	0.311	0.04	-0.89	0.033

TABLE 2.11: Scenario 3a: Normal distributionl $m = 2$ and $\boldsymbol{\eta}(\cdot)$ mean. Manifest parameters estimates and time series S.E. (Burn-in=10000).

$\widehat{\mu}_1$	$S.E.(\widehat{\mu}_1)$	$\widehat{\mu}_2$	$S.E.(\widehat{\mu}_2)$	$\widehat{\sigma}_1$	$S.E.(\widehat{\sigma}_1)$	$\widehat{\sigma}_2$	$S.E(\widehat{\sigma}_2)$
9.972	3.071e-05	19.981	3.502e-05	1.015	0.002	1.029	0.001

2.5.5 Scenario 3b

In this scenario we consider $m = 3$ latent states and the spatial function $\boldsymbol{\eta}(\cdots)$ is in the model as the mean of the neighbouring latent states for each area at time t. In this case, the state-dependent distribution is

$$\begin{aligned} Y_i^{(t)} | U_i^{(t)} &\sim N(\mu_u, \sigma_u) \\ &\text{for } u = 1, 2, 3. \end{aligned} \tag{2.14}$$

We simulate a scenario where, considering the severity of the condition, an increase of the mean influences the area state to more severe condition. For $m = 3$ we fix the parameters to:

- $\boldsymbol{\beta}_1^{'} = (\beta_{01}, \beta_{11}, \beta_{22}, \beta_{33}) = (1, -3.5, 3.5, 4)$, where $u = 1$ is the reference group;
- $\boldsymbol{\gamma}_{\boldsymbol{u\bar{u}}}^{'} = (\gamma_{021}, \gamma_{031}, \gamma_{012}, \gamma_{032}, \gamma_{013}, \gamma_{023}, \gamma_{21}, \gamma_{31}, \gamma_{12}, \gamma_{32}, \gamma_{13}, \gamma_{23}) =$
  $= (1, 1, -0.5, 1, -0.5, 0.5, 0.5, 0.5, -0.5, 0.5, -0.5 - 0.5)$,
  where $\bar{u}$ is the latent state at time $(t - 1)$;
- $\boldsymbol{\mu}^{'} = (\mu_1, \mu_2) = \{5, 10, 15\}$;
- $\boldsymbol{\sigma}^{2^{'}} = (\sigma_1, \sigma_2) = (1, 1, 1)$.

In this way, 384 areas are in state $u = 1$,602 in $u = 2$ and 862 in $u = 3$. We estimate the parameters of the latent model with a Metropolis-Hasting algorithm , with $N(0, 0.5)$ as

random walk and Normal proposal of mean zero and standard deviation equal to five. The acceptance rate for $\boldsymbol{\beta}_1$ is 44.82% and for $\boldsymbol{\gamma}_{u\bar{u}}$ is 36.86%.

TABLE 2.12: Scenario 3b: Normal distribution $m = 3$ and $\boldsymbol{\eta}(\cdot)$ mean. State $u = 1$ distribution. Real value 384.

u=1	363	383	384	399
Freq	4	33	9551	2

TABLE 2.13: Scenario 3b: Normal distribution $m = 3$ and $\boldsymbol{\eta}(\cdot)$ mean. State $u = 2$ distribution. Real value 602.

u=2	592	622	602	599
Freq	4	33	9551	2

TABLE 2.14: Scenario 3b: Normal distribution $m = 3$ and $\boldsymbol{\eta}(\cdot)$ mean. State $u = 3$ distribution. Real value 862.

u=1	843	850	862	893
Freq	4	33	9551	2

TABLE 2.15: Scenario 3b: Normal distributionl $m = 3$ and $\boldsymbol{\eta}(\cdot)$ mean. Latent parameters estimates and time series S.E.

$\widehat{\beta}_{01}$	$S.E.(\widehat{\beta}_{01})$	$\widehat{\beta}_{11}$	$S.E.(\widehat{\beta}_{11})$	$\widehat{\beta}_{22}$	$S.E.(\widehat{\beta}_{22})$	$\widehat{\beta}_{33}$	$S.E.(\widehat{\beta}_{33})$
1.139	0.026	-2.538	0.016	3.787	0.021	1.47	0.06
$\widehat{\gamma}_{021}$	$S.E.(\widehat{\gamma}_{021})$	$\widehat{\gamma}_{031}$	$S.E.(\widehat{\gamma}_{031})$	$\widehat{\gamma}_{012}$	$S.E.(\widehat{\gamma}_{012})$	$\widehat{\gamma}_{032}$	$S.E.(\widehat{\gamma}_{032}$
0.745	0.013	0.577	0.015	-0.324	0.043	1.256	0.103
$\widehat{\gamma}_{013}$	$S.E.(\widehat{\gamma}_{012})$	$\widehat{\gamma}_{023}$	$S.E.(\widehat{\gamma}_{022})$	$\widehat{\gamma}_{21}$	$S.E.(\widehat{\gamma}_{21})$	$\widehat{\gamma}_{31}$	$S.E.(\widehat{\gamma}_{31})$
-0.310	0.015	0.3258	0.01	0.793	0.024	0.440	0.01
$\widehat{\gamma}_{12}$	$S.E.(\widehat{\gamma}_{12}$	$\widehat{\gamma}_{32}$	$S.E.(\widehat{\gamma}_{32})$	$\widehat{\gamma}_{13}$	$S.E.(\widehat{\gamma}_{13})$	$\widehat{\gamma}_{23}$	$S.E.(\widehat{\gamma}_{23})$
-0.435	0.08	0.327	0.02	-0.675	0.04	-0.0795	0.019

TABLE 2.16: Scenario 3b: Normal distributionl $m = 3$ and $\boldsymbol{\eta}(\cdot)$ mean. Manifest parameters estimates and time series S.E.

$\widehat{\mu}_1$	$S.E.(\widehat{\mu}_1)$	$\widehat{\mu}_2$	$S.E.(\widehat{\mu}_2)$	$\widehat{\mu}_3$	$S.E.(\widehat{\mu}_3)$
4.683	5.071e-05	11.901	2.502e-04	13.978	4.065e-04
$\widehat{\sigma}_1$	$S.E.(\widehat{\sigma}_1)$	$\widehat{\sigma}_2$	$S.E(\widehat{\sigma}_2)$	$\widehat{\sigma}_3$	$S.E(\widehat{\sigma}_3)$
1.002	0.002	1.029	0.001	0.981	0.001

2.5.6 Conclusions from the simulation studies

We can notice from these simulation studies that:

- the choice of the function $\boldsymbol{\eta}(\cdot)$ does not affect the precision of the parameters estimates.
- Manifest parameters are always well estimated, but the estimation procedure works better if they are quite distinct.
- When m increases, latent parameters estimates are less accurate. They still keep the sign of the real value, but sometimes there is bias and the true value is not in the confidence interval. This may be due to the fact that we fix the real value of the parameters, then we generate the latent variable structure with a Gibbs sampling and considering its mode to obtain latent variable realizations for area i at time t. After that we generate the undelying latent structure, we sample the manifest variable for area i at time t and then finally we estimate the fixed parameters. In this way the final error is not just MCMC error but depends on all the steps.
- The latent distribution is always quite accurate in all the scenarios.

In appendix A chains diagnostics for scenarios 1a and 2a are reported.

2.6 Application

We use our method on the real dataset Ohio described in Section 2.2. We consider $m = 2$ and $m = 3$ and we selected $m = 2$ latent states based on the proposed model selection method. The marginal likelihood of the data $\boldsymbol{y}$ under the model with $m = 2$ latent states is 19932. In the other case, it is 17006. The selection of $m = 2$ latent states is also quite clear from the MCMC output. In fact, even under the model with $m = 3$ latent states, latent state $u = 3$ appears just in some iterations of the chain. With just two states we decide to consider the function $\boldsymbol{\eta}(\cdot)$ to be the mean of the neighbouring latent states. We use the priors in Section 2.4, with $\sigma^2_\beta = \sigma^2_\gamma = 10$.

We assume that latent states are ordered following the severity of lung cancer death scenario, so that in state 1 the incidence of lung cancer deaths is smaller than in state 2. We use this constraint to avoid label switching.

TABLE 2.17: Ohio dataset estimated parameters.

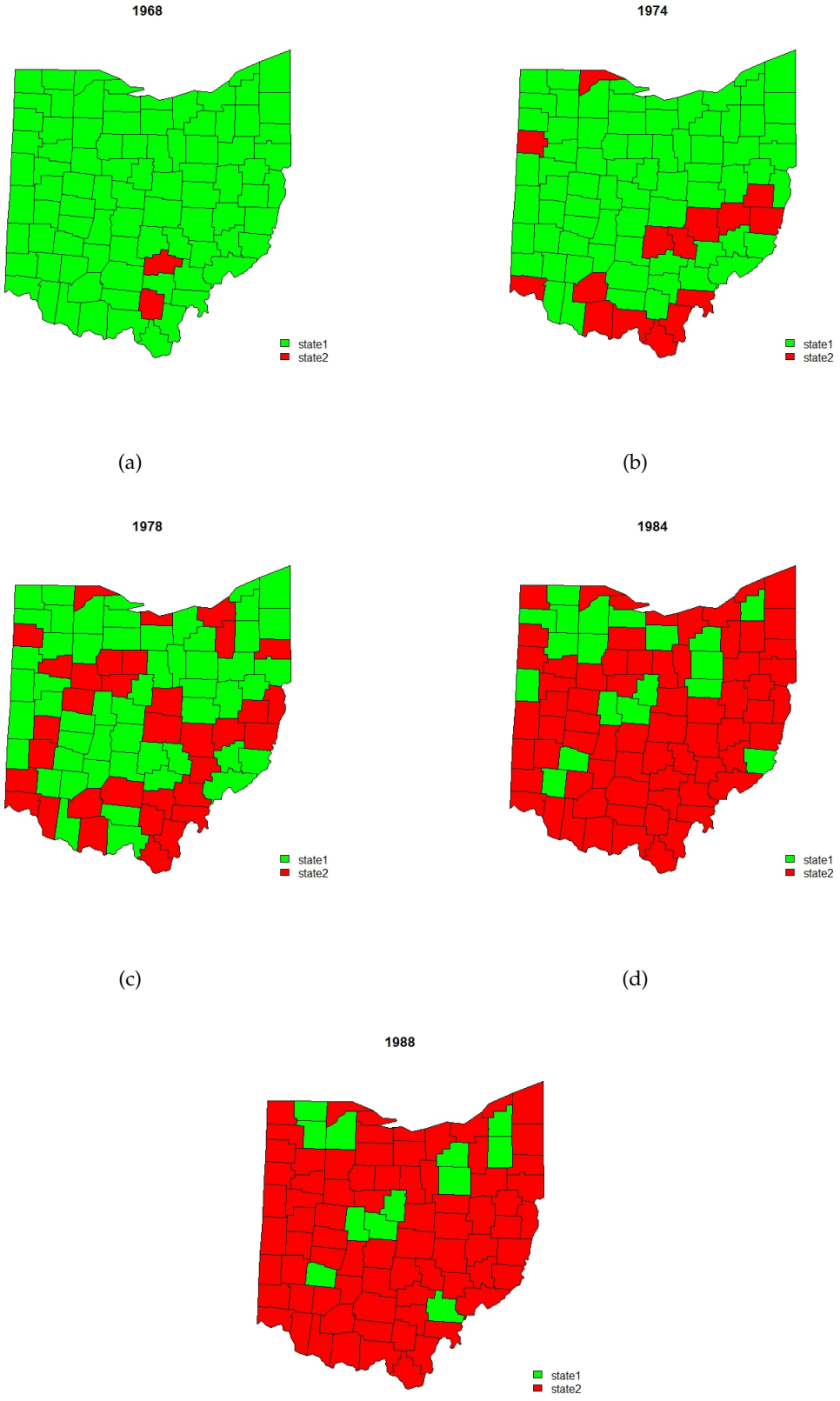
$\widehat{\beta}_{01}$	$S.E.(\widehat{\beta}_{01})$	$\widehat{\beta}_{11}$	$S.E.(\widehat{\beta}_{11})$	$\widehat{\gamma}_{012}$	$S.E.(\widehat{\gamma}_{012})$	$\widehat{\gamma}_{021}$	$S.E.(\widehat{\gamma}_{021})$
-0.0793	0.0030	-0.1155	0.0037	-5.4036	0.1331	0.1048	0.03
$\widehat{\gamma}_{21}$	$S.E.(\widehat{\gamma}_{21})$	$\widehat{\gamma}_{12}$	$S.E.(\widehat{\gamma}_{12})$	$\widehat{p}_1$	$S.E.(\widehat{p}_1)$	$\widehat{p}_2$	$S.E.(\widehat{p}_2)$
2.2528	0.0848	-2.3517	0.2998	0.0003	2.573e-07	0.0005	4.162e-07

We can see from Tab. 2.17 that when the mean of the neighbouring latent states increases, the probability of unit i to remain in state 2 is higher than the probability of moving to state 1 because $\widehat{\gamma}_{21}$ is positive. In contrast, the increase of the spatial covariate has a negative effect on the probability of moving to the lowest state. Moreover, a one-unit increase in the neighbouring latent states mean is associated with a 0.1155 decrease in the relative log odd of being in latent state 2 at the first time of observation. This means that the initial probability of area i to be in state 2 decreases when its neighbouring counties occupy state 2. This is not what we expected. In fact, Fig. 2.1 shows how areas with higher lung cancer death incidence are neighbours. Morover, Fig. 2.1 provides evidence of a strong temporal pattern. Probably, this temporal pattern influences also the initial probabilities and the Markov chain can not estimate them properly because the distribution of the latent state frequencies is not balanced. In fact a strong temporal trend can capture a large amount of information and it does not permit the correct estimates of the initial probabilities, because they are estimated using only data at time $t = 1$. This temporal pattern is also evident in the latent states classification (as in Fig. 2.2).

To avoid this problem and try to capure the temporal trend in the data a covariate that considers data time evolution (Hubert, 1973) may be usefully included in the model. If this covariate is introduced in the transition probabilities parameters of the latent model, the interpretation of the latent states does not change. This solution can also give a measure of the strength of the changes in time of the latent states. Another way to take the temporal trend into account in the data is to introduce the trend covariate in the measurement model. In this way, it affects the response variable estimates and it probably will fix the initial probabilities estimates. However, with this procedure the interpretaton of the latent states changes. They do not provide no more a classification of areas following the severity of incidence of lung cancer deaths, but we can only be interpreted as time space residuals. After some trials, we decide to introduce it in a linear way. We just test for the linear trend introducing this variable as a covariate in the transition probabilities model, according to the metodology developed in this work. With this solution the interpretability of the latent states classification does not change. In fact, it considers just the influence of time in the latent state model. As expected we can see that the increase in time has a positive influence on the probability to occupy state 2 instead of remaining in state 1. Also, the initial probabilities seem to fit better. Probably the inclusion of the covariates in the latent state generation helps the estimation.

TABLE 2.18: Ohio dataset: estimated $\boldsymbol{\beta}_1$ parameters with time variable.

$\widehat{\beta}_{01}$	$S.E.(\widehat{\beta}_{01})$	$\widehat{\beta}_{11}$	$S.E.(\widehat{\beta}_{11})$
-0.0613	0.026	1.363	0.011



(a) (b)

(c) (d)

(e)

FIGURE 2.2: Lantent states classification Ohio 1968-1974-1978-1984-1988.

TABLE 2.19: Ohio dataset: estimated $\gamma_{u\bar{u}}$ parameters with time variable.

$\widehat{\gamma}_{012}$	$S.E.(\widehat{\gamma}_{012})$	$\widehat{\gamma}_{021}$	$S.E.(\widehat{\gamma}_{021})$	$\widehat{\gamma}_{21_1}$	$S.E.(\widehat{\gamma}_{21_1})$
-0.337	0.018	0.046	0.010	1.375	0.011
$\widehat{\gamma}_{21_2}$	$S.E.(\widehat{\gamma}_{21_2})$	$\widehat{\gamma}_{12_1}$	$S.E.(\widehat{\gamma}_{12_1})$	$\widehat{\gamma}_{12_2}$	$S.E.(\widehat{\gamma}_{12_2})$
1.159	0.014	-0.692	0.017	-4.590	0.601

TABLE 2.20: Ohio dataset estimated manifest parameters with time variable.

$\widehat{p}_1$	$S.E.(\widehat{p}_1)$	$\widehat{p}_2$	$S.E.(\widehat{p}_2)$
0.003	7.812e-08	0.005	1.182e-07

2.7 Conclusion

In this work we develop a method to include a spatial structure in LMMs. This extension allows the probability of being in a latent state and the probability to move from a latent state to another over time to be influenced by the neighbouring areas. The model is fitted within a Bayesian framework using Gibbs and Random Metropolis-Hasting algorithm with augmented data that allows for a more efficient sampling of model parameters. Spatial structure is introduced as a function of the latent states in the neighbouring areas. It is important to notice that in this way the spatial structure depends on the latent process, so it's not fixed during the observation period.

We have run simulation studies in order to test for the robustness of the model procedure to the following factors:

- the number of latent states,
- the choice of the spatial function,
- different manifest distributions of the response variables.

Simulations show that the latent state structure and the parameters of the manifest distribution are always well fitted, but the estimation of latent parameters is less precise when the number of latent states increases. The choice of manifest distribution and the choice of the function of latent states which affects the spatial structure do not influence model efficiency. Other simulations can be conducted to investigate prior sensibility and classification of latent state when they are not well divided.

We have applied the proposed model to the Ohio dataset about mortality due to lung cancer in Ohio from 1968 to 1988. We notice that there is a strong temporal pattern in the data, and so we adjuste for it including a time trend variable in the latent model for the estimation of the transition probabilities. Both the spatial latent covariate than the time trand covariate are significant in our data. We find a good classification of the

areas in two groups. The transition probability from state one to state two is higher than the probability that an area i in state one remains in its state. Moreover, the probability of moving towards the better latent states is even lower.

Future research includes a multivariate extension of the model in order to consider more than one response variable and the inclusion of other different covariates in the latent or measurement model. Moreover different neighbouring matrix could be considered with a weighted spatial function.

Chapter 3

Time series SAE for unemployment rates using latent Markov models

3.1 Introduction

In Italy, the Labour Force Survey (LFS) is conducted quarterly by ISTAT, the National Statistical Institute, to produce estimates of the labour force status of the population at a national, regional (NUTS2) and provincial (LAU1) level. Since 1996 ISTAT produces LFS estimates of employed and unemployed counts at labour market areas (LMAs) level. Until 2011 LMAs were 686 and they were sub-regional geographical areas where the bulk of the labour force lives and works, and where establishments can find the largest amount of the labour force necessary to occupy the offered jobs. They were developed through an allocation process based on the analysis of commuting patterns. Since 2011 LMAs are based on commuting data stemming from the 15^{th} Population Census. Now they are redefined in 611 distinct areas (Istat, 2014).

Traditional direct estimation requires sufficiently large samples. Unlike NUTS2 and LAU1 areas, LMAs are unplanned domains and direct estimators have overly large sampling errors particulary for areas with small sample sizes. This makes it necessary to "borrow strength" from data on auxiliary variables from other neighbouring areas through appropriate models, leading to indirect or model based estimates. Small Area Estimation (SAE) methods are used in inference for finite populations to obtain estimates of parameters of interest when domain sample sizes are too small to provide adequate precision for direct domain estimators. Statistical models for SAE can be formulated at the individual or area (i.e. aggregate) levels. When information about the geographic indicators for target areas are available for all individuals in the sample, the usual approach is to estimate regression coefficients and variance components based on a unit-level linear mixed model. Since 2004, after the redesign of LFS sampling strategy, ISTAT uses an empirical best linear unbiased prediction (EBLUP) estimator based on a unit level linear mixed model with spatially autocorrelated random area effects and where individual covariates, such as sex by age classes, are inserted in the fixed part of

the model (Istat, 2006). As mentioned earlier, in 2011 LMAs have been redefined and this leads to re-thinking the SAE strategy. In particular, it is also possible to aggregate the data to area level and estimate SAE parameters based on a linear model for the areas. Area level data are computationally easier to manage because they are widely smaller in number, in particular with the application at hand of LFS where 11 years of data are available.

The Fay-Herriot model (Fay and Herriot, 1979, FH) is considered the basic area level SAE model. It combines cross-sectional information at each time for computing the estimate, but does not borrow strength over the past time periods. When longitudinal data is available, the idea is to borrow strength over time, too. In the last two decades, several approaches that allow to borrow strength simultaneously in space and in time have been developed. Estimators based on the approach developed by Rao and Yu (1994) successfully use space and time informations to produce improved estimated with desiderable properties for small areas. Ghosh, Nangia, and Kim (1996) apply a fully Bayesian analysis using a time series model to the estimation of median income of four-person families. Datta et al. (1999) apply this model to a longer time series accross small areas from the U.S. Current Population Survey using a random walk model. You, Rao, and Gambino (2003) apply the same model to unemployment rate estimation for the Canadian Labour Force Survey using short time series data accross small areas, so that they do not consider seasonal parameters. Finally, Marhuenda, Molina, and Morales (2013) develop a spatio-temporal FH with simultaneous autoregressive model in space (SAR) plus first order autoregressive (AR(1)) covariance structure in time.

Hierarchical Bayes (HB) models have been largely used in SAE (see Rao, 2003, Chapter 10). A HB structure allows to rewrite complex models for the data as simple models building blocks and it also allows to take into account the different sources of variation. Ghosh, Nangia, and Kim (1996) consider HB generalized linear models for an unified analysis of both discrete and continuous data. Fabrizi et al. (2011) develop a model-based SAE method for calculating estimates of poverty rates based on different thresholds for subsets of the Italian population in a HB framework. Finally, Boonstra (2014) uses a time-series HB multilevel model to estimate municipal unemployment based on the Dutch Labour Force Survey at a quarterly frequency including random municipality effects and random municipality by quarter effects.

This work wants to develop a new area level SAE method based on Latent Markov Models (LMMs, see Bartolucci, Farcomeni, and Pennoni, 2014, for an introduction). In particular, we wish to use this model to estimate unemployment rates in LMAs from 2004 to 2014 within a HB framework. Area-level SAE models consist of two parts, a sampling model formalizing the assumptions on direct estimators and their relationship with underlying area parameters and a linking model that relates these parameters to area specific auxiliary information. In this work a LMM is used as the linking model and the sampling model is introduced as the highest level of hierarchy. The definition

of SAE methods which are able to take into account the non-observable nature of variables of interest is presented in literature only in Fabrizi, Montanari, and Ranalli (2015), but the authors consider just the cross sectional nature of the problem without investigating its time extension. They develop a latent class unit-level model for predicting disability small area counts from survey data.

LMMs, introduced by Wiggins (1973), allow for the analysis of longitudinal data when the response variables measure common characteristics of interest which are not directly observable. The basic LMMs formulation is similar to that of Hidden Markov models for time series data (MacDonald and Zucchini, 1997). In these models the characteristics of interest, and their evolution in time, are represented by a latent process that follows a Markov chain, tipically of first order. Latent models represent the evolution of the latent characteristic over time and areas are allowed to move between the latent states during the period. LMMs can be seen as an extension of latent class models (Lazarsfeld, Henry, and Anderson, 1968) to longitudinal data. Moreover, LMMs may be seen as an extension of Markov chain models to control for measurement errors. The model presented in this work is fitted within a Bayesian framework using Gibbs sampler with augmented data that allows for a more efficient sampling of model parameters.

This chapter is organised as follows. Section 3.2 provides a more detailed description of the available LFS data. In Section 3.3, the model is described in detail, while in Section 3.4 the estimation procedure is presented. Section 3.5 is devoted to application results. Conclusions and possible future developments are outlined in the final Section 3.6.

3.2 Data

In Italy, the LFS is conducted quarterly by ISTAT to produce estimates of the labour force status of the population at national, NUTS2 and LAU1 level (D'Alo et al., 2012). Survey results are produced and disseminated on a quarterly basis and once a year as annual averages. Since 1996 ISTAT produces estimates also for LMAs. LMAs are unplanned domains for the LFS. In fact, the sampling design is as follows. Whitin a given LAU1, municipalities are classified as Self-Representing Areas (SRAs; larger municipalities) and Non Self-Representing Areas (NSRAs; smaller municipalities). In SRAs a stratified cluster sampling design is applied. Each municipality is a single stratum and households are selected by means of systematic sampling. In NSRAs, the sample is based on a stratified two stage sampling design. Municipalities are primary sampling units (PSUs), while households are Secondary Sampling Units (SSUs). PSUs are divided into strata of the same dimension in terms of population size. One PSU is drawn from each stratum without replacement and with probability proportional to the PSU

population size. SSUs are selected by means of systematic sampling in each PSU. All members of each sample household, both in SRAs and in NSRAs are interviewed. In each quarter, about 70,000 households and 1.350 municipalities are included in the sample. Note that some LMAs (generally the smallest ones) may have a very small sample size. Furthermore, usually about a third of the LMAs is not included in the sample at all (i.e. they have a zero sample size).

Households are rotated according to a 2-(2)-2 rotation scheme. Households are interviewed during two consecutive quarters. After a two-quarter break, they are again interviewed twice in the corresponding two quarters of the following year. As a result, each household is included in four waves of the survey (Eurostat, 2015). This work uses yearly unemployment incidences for 611 LMAs for the period 2004-2014 from the LFS. For a sake of semplicity, in this paper we call unemployment incidences unemployment rates. LFS yearly direct estimats of unemployment at LMAs level are obtained as the arithmetic mean of the quarterly direct estimates. The aggregation of quarterly variance estimates to produce the annual ones has to take into account the correlation between quarters due to the partial overlap of the sample during the four quarters in a year. Therefore it is obtained by multiplying the estimate of each of the four quarterly variances by a rotation coefficient, which depends on the correlation between estimates that are based on overlapping sample units.

Over all times and areas, 1895 direct estimates cannot be computed because the sample dimension is zero. In addition, missing values are more frequent in CVs compared to direct estimates (see Tab. 3.1 and Tab. 3.2) for the reason that when direct estimates are exactly equal to zero, CVs can not be calculated. Moreover it is necessary to underline that estimates in LFS are produced quarterly and then they are aggregate with an arithmetic mean to obtain yearly direct estimates. As a consequence, a zero value in a quarter does not influence the annual estimate, but its respective missing CV leads to a missing annual CV.

Tab. 3.3 shows the classification of the goodness of estimates based on their CV (Statistics Canada, 2005). Estimates with a CV greater than 33.3% are considered too unreliable to be published. Estimates with CV from 16.6% to 33.3% must be used with caution because their sampling variability is quite high while estimates with CV smaller than 16.6% are considered reliable. In our data, the vast majority of direct estimates have large CV and can not be considered reliable estimates. In particular, in 2004 the 56.7% of direct estimates cannot be considered reliable and more that 50% in all our sample.

The basic idea of SAE is to introduce a statistical model to exploit the relationship between the variable of interest and some covariates for which population information is available or which characterize each area. Auxiliary variables available for these data are the following:

TABLE 3.1: Summary of Unemployment Rates direct estimates (%) from 2004 to 2014.

year	min	1st Qu	Median	Mean	3rd Qu.	Max	NA
2004	0.00	1.78	2.75	3.32	4.55	11.09	160
2005	0.00	1.81	2.82	3.23	4.41	10.25	174
2006	0.23	1.62	2.41	2.76	3.63	10.01	178
2007	0.00	1.49	2.26	2.60	3.49	9.96	176
2008	0.00	1.65	2.44	2.88	3.91	13.04	169
2009	0.00	2.10	2.87	3.21	4.02	15.06	167
2010	0.00	2.13	3.15	3.41	4.32	14.44	164
2011	0.21	2.16	3.02	3.43	4.57	10.52	169
2012	0.00	3.06	4.06	4.56	5.83	14.18	166
2013	0.00	3.42	4.67	5.03	6.52	12.37	185
2014	0.38	3.58	4.88	5.33	6.71	17.58	187
2004-2014	0.00	2.05	3.23	3.61	4.71	17.58	1895

- Time varying variables:
  - population rates in $sex \times 14\, age$ classes (0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34,35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65+).
- Fixed in time variables:
  - *Cultural Vocation* : a qualitative nominal variable with five categories:
    - Great beauty (70 LMAs that are both artistic and naturalistic centers and which have a manufacturing based on cultural connotation);
    - Potential heritage (138 LMAs that are artistic and naturalistic centers but where the entrepreneurial dimension is less developed);
    - Cultural activity (138 LMAs that have a manufacturing base to cultural connotation even if they are not artistic or naturalistic centers);
    - Tourism (194 LMAs that have a poorly developed cultural dimension or industry but they are a tourist destination);
    - Cultural remotness (71 LMAs that are under the EU standard in any cultural or naturalistic classes).
  - *Prevalent Specialization*: a qualitative nominal variable with four levels:
    - not specialized,
    - not manufacturing,
    - made in Italy,
    - industrial.

TABLE 3.2: Summary of unemployment rates direct estimates CV (%) from 2004 to 2014.

year	min	1st Qu	Median	Mean	3rd Qu.	Max	NA
2004	4.91	25.81	36.73	39.34	48.34	148.8	216
2005	5.10	25.25	36.12	38.61	48.16	102.7	231
2006	5.89	25.86	38.01	42.42	52.28	154.2	264
2007	6.88	27.80	39.18	43.58	54.58	144.2	257
2008	6.82	26.95	36.49	41.32	49.75	150.3	238
2009	6.89	25.00	35.06	39.06	47.31	154.3	234
2010	6.56	23.83	34.48	37.77	46.11	122.0	234
2011	6.16	24.38	33.35	37.24	44.86	125.7	229
2012	4.96	22.11	30.56	33.08	39.99	116.6	211
2013	4.14	21.37	28.24	31.72	37.96	142.3	219
2014	4.14	20.08	28.30	31.40	38.26	122.3	214
2004-2014	4.14	23.85	33.72	37.64	46.30	154.3	2547

These qualitative variables are defined in the 2015 annual report by ISTAT (2015).

3.2.1 Smoothing MSE

The estimated mean square errors and CVs show wide variation (See Tab. 3.2). These wide variations can be seen as a function of the sample size. Errors in estimates of sampling error can affect small area modelling in different ways because the estimates of final unemployment rates depend on it (Rao and Yu, 1994). For this reason, smoothing estimated mean square errors is necessary. In this work, we propose to smooth them using a factor regression model with a logarithmic trasformation of the coefficient of variation and of the mean square error. The model is divided into two steps which are applied to quarterly estimates. The first step aims at computing the smoothed MSE when an estimate of the unemployment rate is available. The second step imputes smoothed MSEs for the LMAs with a zero sample size.

Let $\hat{\theta}_{it_j}$ be the direct survey estimate for small area $i = 1, \ldots, m$, with $m = 611$, at year $t = 1, \ldots, T$, with $T = 11$, and where $j = 1, \ldots, 4$ are the quarters of observation. Let $CV^2_{it_j}$ be the correspondig squared coefficient of variation. In order to obtain smooth estimated MSEs, the following auxiliary information is used:

- Dummy variables indicating the geographic macro-area LMA i belongs. In particular, δ_{hi}, for $h = 1, \ldots, 4$, are north-west, north-east, center, south respectively.
- M_{it_j} the population size at time t, quarter j, of the macro-area LMA i belongs to;
- N_{it_j} the population size of LMA i at time t and quarter j.

TABLE 3.3: Number of small areas with values of CV less than 16.6%, between 16.6% and 33.3% and over 33.3% for Direct estimator (from 2004 to 2014).

year	< 16.6	16.6-33.3	> 33.3
2004	41	130	224
2005	35	125	220
2006	24	115	208
2007	18	113	223
2008	24	132	217
2009	22	154	201
2010	25	155	197
2011	25	166	191
2012	42	188	170
2013	55	197	140
2014	60	198	139
2004-2014	371	1673	2130

The first step is based on predictions from the following model:

$$\log(CV_{it_j}^2) = \sum_{h=1}^{4} \beta_h \delta_{hi} + \sum_{h=1}^{4} \beta_{4+h} \log(\hat{\theta}_{it_j}) + \sum_{h=1}^{4} \beta_{8+h} \frac{N_{it_j}}{M_{it_j}}. \tag{3.1}$$

Smoothed $\widehat{CV}_{it_j}$ are obtained as the square root of the exponential of such predictions. Then, smoothed MSEs are obtained as

$$\widehat{MSE}_{it_j} = \widehat{CV}_{it_j} \times \hat{\theta}_{it_j}. \tag{3.2}$$

When direct estimates are missing, the model presented in this first step can not be use because $CV_{it_j}^2$ and $\hat{\theta}_{it_j}$ are not available. In this second second step smoothed MSEs are computed directly for LMA i at point t_j using the prediction from the following model:

$$\log(MSE_{it_j}) = \sum_{h=1}^{4} \beta_h \delta_{hi} + \sum_{h=1}^{4} \beta_{4+h} \frac{N_{it_j}}{M_{it_j}}. \tag{3.3}$$

Annual smoothed MSEs are obtained as the mean of quarterly MSEs adjusted in order to take into account the partial overlap of the quarterly samples.

3.3 Model

3.3.1 General SAE framework

Rao and Yu (1994) propose an area level model involving autocorrelated random effects and sampling errors using both time series and cross sectional data. It consists of a

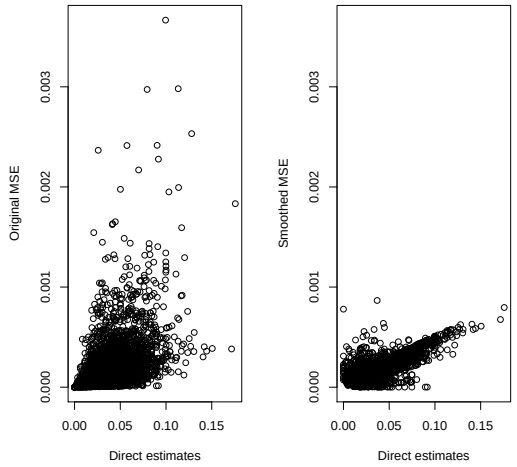


FIGURE 3.1: Direct estimates vs original MSEs and Direct estimates vc smoothed MSEs.

sampling model

$$\begin{aligned} \hat{\theta}_{it} &= \theta_{it} + e_{it}, \\ i = 1, \ldots, m,\ t &= 1, \ldots, T, \end{aligned} \tag{3.4}$$

and an area-linking model

$$\begin{aligned} \theta_{it} &= \boldsymbol{x}_{it}^T \boldsymbol{\beta} + v_i + u_{it}, \\ i = 1, \ldots, m,\ t &= 1, \ldots, T, \end{aligned} \tag{3.5}$$

where $\hat{\theta}_{it}$ is the direct survey estimator for small area i at time t, $\theta_{it} = g(\bar{Y}_{it})$ is a function of the small area mean, $e_{it}|\theta_{it}$ are normal sampling errors with zero mean and with known covariance matrix $\boldsymbol{\Psi} = \text{blockdiag}\{\boldsymbol{\Psi}_i\}$, $\boldsymbol{x}_{it}$ are area specific covariates (possibly time-varying), $v_i \sim N(0, \sigma_v^2)$ is the area effect and $u_{it} = \rho u_{i,t-1} + \epsilon_{it}$ with $|\rho| < 1$ and $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ is the area-by-time effect. In this model, e_{it}, v_i and ϵ_{it} are assumed independent of each other.

The linking model is basically a linear model with mixed coefficients. You, Rao, and Gambino (2003) translate this model into a HB framework as follows:

$$\begin{aligned} \widehat{\theta}_{it}|\theta_i &\sim N(\theta_{it}, \boldsymbol{\Psi}_i) \\ \theta_{it}|\boldsymbol{\beta}, u_{it}, \sigma_v^2 &\sim N(\boldsymbol{x}_{it}^T \boldsymbol{\beta} + \rho u_{it}, \sigma_v^2) \\ u_{it}|u_{i,t-1}, \sigma_u^2 &\sim N(\rho u_{i,t-1}, \sigma_u^2) \end{aligned} \tag{3.6}$$

where $\boldsymbol{\beta}$, σ_v^2 and σ_u^2 are mutually independent with priors given as $\boldsymbol{\beta} \propto 1$, $\sigma_v^2 \sim IG(a_1, b_1)$ and $\sigma_u^2 \sim IG(a_2, b_2)$. Even if a proper informative prior distribution on the hyperparameters would be appropriate for a fully Bayesian analysis, non-informative

priors are often used because information about these priors is seldom available in real HB applications. The choice of a diffuse prior is not unique and some diffuse improper priors could lead to improper posteriors. In the HB approach, MCMC methods make inference direct and computationally feasible. For this reason, in a HB approach it is easier to consider more complicated and realistic methods to analyse SAE problems. Usually the Rao-Blackwellization approach is used to obtain estimators for the posterior mean and the posterior variance of interest. In the next subsections, the new model based on LMMs is illustrated.

3.3.2 Introducing elements on LMMs

As previously said, LMMs allow for the analysis of longitudinal data using latent variables. A fundamental assumption of LMMs is that of local independence, according to which the response variables are conditionally independent given the latent variables. The motivation of this assumption is that the latent variables represent the only explanatory factor of the response variables. In fact, the response variables provide a measure of the latent ones and the latent process fully explains the observable behaviour of an area. Furthermore, the latent state to which an area belongs to at a certain time point only depends on the latent state at the previous occasion.

In LMMs the existence of two processes is assumed: an unobservable finite-state first order Markov chain U_{it}, $i = 1, \ldots, m$ and $t = 1, \ldots, T$ with state space $U = \{1, \ldots, k\}$ and an observed process θ_{it}, $i = 1, \ldots, m$ and $t = 1, \ldots, T$, where θ_{it} denotes the response variable for area i at time t and similary for U_{it}. It is assumed that the distribution of θ_{it} depends only on U_{it}, specifically the θ_{it} are conditionally independent given the U_{it}. The state-dependent distribution, i.e. the distribution of θ_{it} given U_{it}, can be a continuous or a discrete distribution. It can be taken for example from the exponential family. Thus, the unknown vector of parameters $\boldsymbol{\phi}$ in a LMM includes both the parameters of the Markov chain $\boldsymbol{\phi}_{lat}$ and the vector of parameters of the state-dependent distribution of the random variables θ_{it} conditionally on U_{it} , $\boldsymbol{\phi}_{obs}$.

Due to the existence of these two processes, it is possible to differentiate between two components of the model which are formulated through specific assumptions: the measurement model and the latent model. The measurement model concerns the conditional distribution of the response variables given the latent variables. The latent model concerns the distribution of the latent variables, instead. By jointly considering the two above components, the so-called manifest distribution is obtained. It is given by the marginal distribution of the response variables, once the latent variables have been integrated out.

The measurement model involves $\boldsymbol{\phi}_{obs}$ and it can be written as

$$\theta_{it}|U_{it} \sim p(\boldsymbol{\phi}_{obs}).$$

The parameters ϕ_{lat} of the Markov chain are the elements of the transition probability matrix

$$\mathbf{\Pi} = \{\pi_{(u|\bar{u})}\} \text{ , for } u, \bar{u} = \{1, \ldots, k\}$$

with

$$\pi_{(u|\bar{u})} = P(U_{it} = u | U_{i,t-1} = \bar{u}) \text{ , } u = \{1, \ldots, k\} \tag{3.7}$$

is the probability that area i visits state u at time t given that at time $t-1$ it was in state $\bar{u}$, and the initial probabilities

$$\boldsymbol{\pi} = (\pi_{(1)}, \ldots, \pi_{(u)}, \ldots, \pi_{(k)})$$

where

$$\pi_{(u)} = P(U_{i1} = u) \tag{3.8}$$

is the probability of being in state u at the initial time for $u = 1, \ldots, k$. For the sake of simplicity, in this work we consider homogeneous LMMs, i.e. LMMs where the transition probability matrix is constant as a function of time t, but the hidden Markov chain could also be considered non homogeneous and the transition probabilities time-varying.

The basic LMM, relying on a homogenous Markov chain, has several extensions based on parameterizations that allow to include hypotheses and constraints of interest. These parameterizations may concern the conditional distribution of the response variables given the latent process (measurement model) and/or the distribution of the latent process (latent model). Individual covariates could be included in the measurement or in the latent model. When the covariates are included in the measurement model (Bartolucci and Farcomeni, 2009), they affect the response variable and the latent process is seen as a way to account for the unobserved heterogeneity between areas. Differently, when the covariates are in the latent model (Vermunt and Magidson, 2002; Bartolucci, Pennoni, and Francis, 2007) they influence initial and transition probabilities of the latent process. We will consider then the former approach.

A Bayesian inference approach to LMMs is already available in the literature , e.g. in Marin, Mengersen, and Robert (2005), Spezia (2010), and Bartolucci and Pandolfi (2011). In the following section we illustrate how to incorporate a LMM into an area level SAE model.

3.3.3 LMMs SAE model specifications

In this section the methodology proposed is presented in more detail. The model presented is composed by two structures in a HB framework. At the first level, a sampling error model is assumed, then a LMM is used as linking model, that it is composed by

two equations, the measurement model and the latent model. The latent Markov SAE model has the following parametrization:

- Sampling Model

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i &\sim \boldsymbol{N}_T(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_i) \\ i = 1, \ldots, m; \boldsymbol{\theta}_i &= (\theta_{i1}, \ldots, \theta_{iT})^T \end{aligned} \tag{3.9}$$

- Linking Model
  - Measurement Model

$$\begin{aligned} \theta_{it}|U_{it}, \boldsymbol{x}_{it} &\sim N(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_{u(it)}, \sigma^2_{u(it)}) \\ i = 1, \ldots, m,\ t &= 1, \ldots, T \end{aligned} \tag{3.10}$$

  - Latent Model
    - Initial probabilities

$$\begin{aligned} P(U_{i1} = u) &= \pi_{(u)} \\ t = 1;\ u &= 1, \ldots, k \end{aligned} \tag{3.11}$$

    - Transition probabilities

$$\begin{aligned} P(U_{it} = u|U_{i,t-1} = \bar{u}) &= \pi_{(u|\bar{u})} \\ t = 2, \ldots, T;\ u, \bar{u} &= 1, \ldots, k. \end{aligned} \tag{3.12}$$

Here $\widehat{\boldsymbol{\theta}}_i$ is a row vector of dimension T of the direct estimates used as data, θ_{it} is the estimate which has to be produced for area i at time t, $\boldsymbol{x}_{it}^T$ is the vector $p \times 1$ of the auxiliary informations , where p is the number of auxiliary variables, $\boldsymbol{\beta}_{u(it)}$ is the $p \times 1$ vector of the regression coefficients for the latent state to which area i at time t belongs to, $\sigma^2_{u(it)}$ is the variance of the latent state to which area i at time t belongs to and $\boldsymbol{\Psi}_i$ are the sampling variances, which are assumed known. The matrix of transition probabilities $\boldsymbol{\Pi}$ has $\boldsymbol{\pi}_{(\cdot|\bar{u})}$ on the rows and, similarly, $\boldsymbol{\pi}_{(u|\cdot)} = (\pi_{(u|1)}, \ldots, \pi_{(u|k)})^T$ on the columns.

The model proposed is a very flexible modeling framework. It could be seen as a *matched Normal-Normal model* You and Rao (2002). However, this definitions is not totally appropriate in this case because a *matched model* is defined as a model which is obtained by combining the sampling and linking models with the aim to produce a relatively simple linear mixed model. Instead, in our proposal the latent model with the initial and the transition probabilities is present at the lowest hierarchical level.

Moreover, it has to be noticed that while in the linear mixed model heterogenity is continuous, in the considered context it is modelled with a discrete dynamic variable. As we can see from Fig. 3.2 our data has a skew distribution. However, the distribution is not far from a Normal density. D'Alo et al. (2012) show that the differences in estimates between adopting a Normal or a Binomial model are not as large as expected and Normal models are often used for unemployment rates estimation (You, Rao, and Gambino, 2003; Boonstra, 2014). Finally adopting the Normal distribution has computational advantages which will be clarified later in this section.

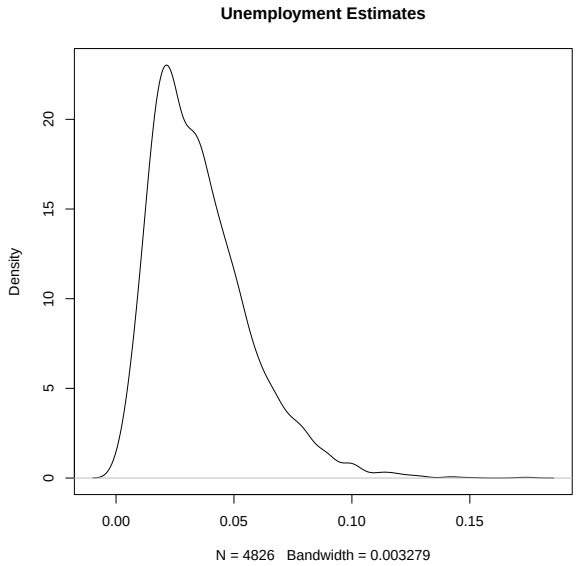


FIGURE 3.2: Direct estimates of unemployment rates density.

The parameters of interest in the model can be divided into three groups: the *small area parameters of interest*, the *measurement parameters of interest* and the *latent parameters of interest*. In particular, they are given as follows:

- *Small area*

$$\boldsymbol{\mu} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)^T; \tag{3.13}$$

- *Measurement*

$$\boldsymbol{\phi}_{obs} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \sigma_1^2, \ldots, \sigma_k^2\}; \tag{3.14}$$

- *Latent*

$$\boldsymbol{\phi}_{lat} = \{\boldsymbol{\pi}, \boldsymbol{\Pi}\}; \tag{3.15}$$

To complete the Bayesian specification of the proposed model, it is necessary to chose priors for the parameters of interest which appear into the model. *Small area parameters* do not need a specific prior because data direct estimates are available, so a set of priors has to be chosen for the *measurement* and the *latent* ones. Looking at $\boldsymbol{\phi}_{obs}$ diffuse

normal priors with mean $\boldsymbol{\eta}_0$ and variance $\sigma_u^2 \boldsymbol{\Lambda}_0^{-1}$ are assumed for the regression coefficients. These priors are conjugate, non-informative and computationally more convenient than the usually flat priors over the real line (see Rao, 2003, Chapter 10). In particular,

$$\boldsymbol{\beta}_u \sim \mathrm{N}_p(\boldsymbol{\eta}_0, \boldsymbol{\Sigma}_0), \text{ for } u = 1, \ldots, k, \ \boldsymbol{\Sigma}_0 = \sigma_u^2 \boldsymbol{\Lambda}_0^{-1}. \tag{3.16}$$

Variances σ_u^2, $u = 1, \ldots, k$, are unknown so it is necessary to set a prior on them, too. An inverse gamma distribution with shape parameter a_0 and scale parameter b_0, with $a_0, b_0 > 0$ is assumed. The positive constants a_0 and b_0 are set to be very small. In this way a large variance is assumed and the prior is considered non-informative:

$$\sigma_u^2 \sim IG(a_0, b_0), \text{ for } u = 1, \ldots, k. \tag{3.17}$$

For $\boldsymbol{\phi}_{lat}$, a system of Dirichlet priors is set on the initial probabilities and on the transition probabilities. The Dirichlet distribution is a conjugate prior for the multinomial distribution. This means that if the prior distribution of the multinomial parameters is Dirichlet then the posterior distribution is also a Dirichlet distribution. The benefit of this choice is that the posterior distribution is easy to compute and, in some sense, it is possible to quantify how much our beliefs have changed after collecting the data. Then,

$$\boldsymbol{\pi} = (\pi_{(1)}, \ldots, \pi_{(k)})^T \sim \text{Dirichlet}(\mathbf{1}_k) \tag{3.18}$$

$$\boldsymbol{\pi}_{(\cdot|\bar{u})} = (\pi_{(1|\bar{u})}, \ldots, \pi_{(k|\bar{u})}) \sim \text{Dirichlet}(\mathbf{1}_k), \text{ for } \bar{u} = 1, \ldots, k. \tag{3.19}$$

3.4 Model estimation

LMMs have been at the source of different methodological developments in computational Statistics. This work looks at the Data Augmentation method (Tanner and Wong, 1987; Liu, Wong, and Kong, 1994; Van Dyk and Meng, 2001) in which the latent states are introduced as missing data (Marin, Mengersen, and Robert, 2005; Germain, 2010). There are two main reasons for this choice. First of all, it's been showed that the performance of the marginal updating scheme is worse than that of the corresponding data augmantation approach (Boys and Henderson, 2003). Moreover, in this way we can simplify the process of sampling from the posterior distribution.

3.4.1 Data Augmentation Method

The goal of model inference is to estimate the set of small area, measurement and latent parameters and the latent component $\boldsymbol{u}$. In order to be able to calculate these estimates,

a data augmentation approach (Tanner and Wong, 1987) is applied. This approach associates at each observation $\hat{\theta}_{it}$ a latent multinomial variable $U_{it} \sim \text{Multi}_k(1; p_{it1}, \ldots, p_{itk})$ such that $\hat{\theta}_{it}|U_{it} \sim p(\theta_{it}|\boldsymbol{\phi})$. In this context there is the possibility of choosing priors which are conjugate to the form of the complete data likelihood, therefore sampling from the conditional posterior of the model parameters given the latent states (the so called complete data posterior) is straightforward. Moreover, because the state space is discrete and finite, sampling from the conditional posterior of the latent states given the model parameters is also possible. Then, it is possible to generate samples from the joint posterior distribution of the model parameters and latent states as follow. Let $\boldsymbol{\theta}$ be the matrix of realizations of the response variable.

$$\pi(\boldsymbol{\phi}, \boldsymbol{u}|\boldsymbol{\theta}) = \pi(\boldsymbol{\phi})p(\boldsymbol{u}|\boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{u})$$

$$\text{joint posterior} \;=\; \text{prior} \;\times\; \text{likelihood} \;\times\; \text{likelihood}.$$

Samples are computed by alternating between sampling $\boldsymbol{u}$ from the conditional posterior distribution $\pi(\boldsymbol{u}|\boldsymbol{\theta}, \boldsymbol{\phi})$ and drawing $\boldsymbol{\phi}$ from the conditional posterior distribution $\pi(\boldsymbol{\phi}|\boldsymbol{\theta}, \boldsymbol{u})$. When a priori independence is assumed between $\boldsymbol{\phi}_{obs}$ and $\boldsymbol{\phi}_{lat}$ the complete data posterior can be written as

$$\pi(\boldsymbol{\phi}|\boldsymbol{u}, \boldsymbol{\theta}) = \pi(\boldsymbol{\phi}_{obs}|\boldsymbol{u}, \boldsymbol{\theta})\pi(\boldsymbol{\phi}_{lat}|\boldsymbol{u}, \boldsymbol{\theta})$$

$$\text{complete data posterior} \;=\; \text{posterior} \;\times\; \text{posterior}.$$

and the MCMC sampling scheme leads to repeat for R iterations $r = 1, \ldots, R$ the following steps:

1. Simulate $\boldsymbol{u}^r$ from $\pi(\boldsymbol{u}|\boldsymbol{\phi}^r, \boldsymbol{\theta})$.
2. Simulate $\boldsymbol{\phi}^r$ from $\pi(\boldsymbol{\phi}|\boldsymbol{u}^{r-1}, \boldsymbol{\theta})$ where:
   (a) $\boldsymbol{\phi}^r_{lat}$ is simulated from $\pi(\boldsymbol{\phi}_{lat}|\boldsymbol{u}^{r-1})$,
   (b) $\boldsymbol{\phi}^r_{obs}$ is simulated from $\pi(\boldsymbol{\phi}_{obs}|\boldsymbol{u}^{r-1}, \boldsymbol{\theta})$.

If each $\boldsymbol{\theta}_i$ is assumed independent, $\boldsymbol{u}_i$ can be sampled individually using a Gibbs sampler from the posterior and can be drawn from

$$U_{it} \sim \text{Multi}_k(p(U_{it} = 1|\theta_{it-1}, \theta_{it+1}, \boldsymbol{\phi}), \ldots, p(U_{it} = k|\theta_{it-1}, \theta_{it+1}, \boldsymbol{\phi})). \tag{3.20}$$

The complete data posterior distribution is given by the Bayes Theroem as

$$\pi(\boldsymbol{\phi}|\boldsymbol{\theta}, \boldsymbol{u}) \propto \pi(\boldsymbol{\phi})p(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{\phi}).$$

If a priori independence is assumed between $\boldsymbol{\phi}_{obs}$ and $\boldsymbol{\phi}_{lat}$, given $\boldsymbol{u}$ these parameters remain conditionally independent a posteriori and the complete data posterior can be

decomposed in

$$\pi(\boldsymbol{\phi}_{obs}|\boldsymbol{\theta}, \boldsymbol{u}) \propto \pi(\boldsymbol{\phi}_{obs}) p(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{\phi}_{obs}),$$

and

$$\pi(\boldsymbol{\phi}_{lat}|\boldsymbol{\theta}, \boldsymbol{u}) \propto \pi(\boldsymbol{\phi}_{lat}) p(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{\phi}_{lat}).$$

The overall form of the complete data posterior distribution $\pi(\boldsymbol{\phi}_{obs}|\boldsymbol{\theta}, \boldsymbol{u})$ is specific to a particular latent Markov model. If the prior for a component of $\boldsymbol{\phi}_{obs}$ is conjugate to the form of the complete data likelihood, than the full conditional distribution belong to the same family of distribution as the prior and can be sampled directly with a Gibbs sampler.

3.4.2 Model estimation specification

Assuming priors presented in Section 3.3.3, Gibbs conditionals are given by :

$$[u_{it}|\boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \widehat{\theta}] \sim \text{Multi}_k(1; \boldsymbol{\pi}_{1,it}), \text{ for } i = 1, \ldots, m; t = 1, \ldots, T \tag{3.21}$$

$$[\boldsymbol{\pi}|\boldsymbol{u}, \boldsymbol{\Pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \widehat{\theta}] \sim \text{Dirichlet}(\mathbf{1}_k + \boldsymbol{n}_1) \tag{3.22}$$

$$[\boldsymbol{\pi}_{\bar{u}}|\boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \widehat{\theta}] \sim \text{Dirichlet}(\mathbf{1}_k + \boldsymbol{n}_{\bar{u},t}), \text{ for } t = 2, \ldots, T \tag{3.23}$$

$$[\boldsymbol{\beta}_u|\boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \widehat{\theta}] \sim \boldsymbol{N}_p(\boldsymbol{\eta}_{1,u}, \boldsymbol{\Sigma}_{1,u}) \tag{3.24}$$

$$[\sigma_u^2|\boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\beta}, \boldsymbol{\theta}, \widehat{\theta}] \sim \text{IG}(a_{1,u}, b_{1,u}) \tag{3.25}$$

$$[\theta_{it}|\boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \widehat{\theta}] \sim \text{N}(\widehat{\theta}_{it}^B(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2), \gamma_{it}\psi_{it}) \tag{3.26}$$

where

- $\boldsymbol{\pi}_{1,it}$ is different according to t:
  - for $t = 1$: $\boldsymbol{\pi}_{1,it} = \boldsymbol{\pi} * \boldsymbol{\pi}_{(u|\cdot)}$ with $u = u_{(i,t+1)}$
  - for $t = 2, \ldots, T-1$: $\boldsymbol{\pi}_{1,it} = \boldsymbol{\pi}_{(u|\cdot)} * \boldsymbol{\pi}_{(\cdot|\bar{u})}^T$ with $u = u_{(i,t+1)}, \bar{u} = u_{(i,t-1)}$,
  - for $t = T$: $\boldsymbol{\pi}_{1,it} = \boldsymbol{\pi}_{(\cdot|\bar{u})}^T$ wih $\bar{u} = u_{(i,t-1)}$

  and $*$ indicates the elementwise product;
- $\boldsymbol{\eta}_{1,u} = \boldsymbol{\Lambda}_{1,u}^{-1} \sum_{it} \boldsymbol{x}_{it}\theta_{it} I(U_{it} = u)$;
- $\boldsymbol{\Sigma}_{1,u} = \boldsymbol{\sigma}_u^2 \Lambda_{1,u}^{-1}$;

- $\boldsymbol{\Lambda}_{1,u} = \sum_{it} \boldsymbol{x}_{it}\boldsymbol{x}_{it}^T I(U_{it} = u) + \boldsymbol{\Lambda}_0$;
- $a_{1,u} = a_0 + N_u/2$ where N_u is the number of areas is state u;
- $b_{1,u} = b_0 + \frac{1}{2}(\sum_{it} \theta_{it}^2 I(U_{it} = u) + \boldsymbol{\eta}_0^T \boldsymbol{\Lambda}_0 \boldsymbol{\eta}_0 - \boldsymbol{\eta}_{1,u}^T \boldsymbol{\Lambda}_{1,u} \boldsymbol{\eta}_{1,u})$;
- $\boldsymbol{n}_1 = (n_{11}, \ldots, n_{1k})$ and n_{1u} is the number of areas in state u at time $t = 1$, $u = 1, \ldots, k$;
- $\boldsymbol{n}_{\bar{u},t} = (n_{\bar{u},t1}, \ldots, n_{\bar{u},tk})$ and $n_{\bar{u},t}$ is the number of areas in transit at time $t = 2, \ldots, T$; $u = 1, \ldots, k$.

The goal of SAE is to predict a good estimate of θ_{it} based on the model. This comes from:

$$\hat{\theta}_{it}^B(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \gamma_{it}\hat{\theta}_{it} + (1 - \gamma_{it})\boldsymbol{x}_{it}^T \boldsymbol{\beta}_{u(it)} \tag{3.27}$$

where $\gamma_{it} = \sigma_{u(it)}^2/(\sigma_{u(it)}^2 + \psi_{it})$.

A proof of (3.24), (3.25), (3.26) is sketched in Appendix B. Note that all the Gibbs conditionals have closed forms and hence the MCMC samples can be generated directly from the conditionals. Rao-Blackwell estimators of the posterior mean and the posterior variance of the estimates can be obtained.

In our application, many LMAs have zero sample size. As a consequence, direct estimates are not available for these areas. In our context it is possible to impute these missing values using a Gibbs sampler and sampling directly from its full conditional distribution. The full conditional for $\widehat{\theta}_{it}$ is given by:

$$[\widehat{\theta}_{it}|\boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \theta_{it}] \sim \mathrm{N}(\theta_{it}^D(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2), \gamma_{it}^* \psi_{it}) \tag{3.28}$$

where

- $\hat{\theta}_{it}^D(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \gamma_{it}^* \hat{\theta}_{i,t-1} + (1 - \gamma_{it}^*)\boldsymbol{x}_{it}^T \boldsymbol{\beta}_{u(i,t-1)}$
- $\gamma_{it}^* = \sigma_{u(i,t-1)}^2/(\sigma_{u(i,t-1)}^2 + \psi_{it})$

for areas i and time t for which the direct estimates are missing.

3.4.3 Model selection

The identification of the number of latent states is a fundamental step for model selection and parameter estimation. In the framework of LMMs the choice of the number of latent states of the unobserved Markov chain underlying the observed data corresponds to the model selection procedure. From a Bayesian perspective a crucial goal is to compute the marginal likelihood of the data for a given model . In this paper we use a model selection method based on the ma maximum marginal likelihood and to estimate this quantity we use the "Chib" estimator (Carlin and Chib, 1995). This method

is based on the estimation of the marginal likelihood of any available model from the output of the MCMC algorithm. It follows from noticing that for any parameter configuration, Bayes' rule implies that the marginal likelihood of the data $\boldsymbol{\theta}$ for model with m latent states satisfies

$$p(\boldsymbol{\theta}|k) = \frac{p(\bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_T, \overline{\phi}, \boldsymbol{\theta})}{p(\bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_T, \overline{\phi}|\boldsymbol{\theta})}, \tag{3.29}$$

where $\bar{\boldsymbol{u}}_t$, with $t = 1, \ldots, T$ is a $n \times 1$ fixed vector of latent states and $\overline{\phi}$ a vector of fixed parameters. Numerator of (3.29) can be computed immediately from parameters and data distributions. To compute the denominator of (3.29) the following decomposition is used

$$p(\bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_T, \bar{\boldsymbol{\phi}}|\boldsymbol{\theta}) = \left[p(\bar{\boldsymbol{u}}_1|\boldsymbol{\theta}) \prod_{t>2}^{T} p(\bar{\boldsymbol{u}}_t|\boldsymbol{\theta}, \bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_{t-1}) \right] p(\bar{\boldsymbol{\phi}}|\boldsymbol{\theta}, \bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_T).$$

Now each factor can be estimated from the Gibbs output. To estimate $p(\bar{\boldsymbol{u}}_1|\boldsymbol{\theta})$ we use

$$\hat{p}(\bar{\boldsymbol{u}}_1|\boldsymbol{\theta}) = \frac{1}{R} \sum_r p(\bar{\boldsymbol{u}}_1|\boldsymbol{\theta}, \boldsymbol{u}_2^{(r)}, \ldots, \boldsymbol{u}_T^{(r)}, \boldsymbol{\phi}^{(r)}),$$

where $^{(r)}$ denotes the value at the r^{th} iteration of the algorithm and R is the number of iterations. Regarding $p(\bar{\boldsymbol{u}}_t|\boldsymbol{\theta}, \bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_{t-1})$, the estimator is

$$\hat{p}(\bar{\boldsymbol{u}}_t|\boldsymbol{\theta}, \bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_{t-1}) = \frac{1}{R} \sum_r p(\bar{\boldsymbol{u}}_t|\boldsymbol{\theta}, \bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_{t-1}, \boldsymbol{u}_{t+1}^{(r)}, \ldots, \boldsymbol{u}_T^{(r)}, \boldsymbol{\phi}^{(r)}).$$

At last, $p(\overline{\phi}|\boldsymbol{\theta}, \bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_T)$ is a constant.

3.4.4 Label switching

A well-known problem occurring in Bayesian latent class and latent Markov modeling is label switching. This problem can be seen as the non-identifiability of the components due to the invariance of the posterior distribution to the permutation in the parameters labeling. This implies that the component parameters are not identifiable marginally, that is one component can not be distinguished from the others from the likelihood, because they are exchangeable.

In a Bayesian analysis, if the prior distribution does not distinguish the component parameters between each other, then the resulting posterior distribution will be invariant in the permutation of the labels, since it will be proportional to the product of a symmetric likelihood with a symmetric prior distribution.

Several solutions have been proposed, for a general review see Jasra, Holmes, and Stephens (2005). In this work, we impose an artificial identifiability constraint to the

MCMC sample. In such a case, the simulated MCMC output is permuted at every iteration according to the ordering of a specific parameter. Unluckly, this approach works well only when the selected constraint is able to separate well the symmetric posterior modes, which is rarely true. The easiest approach is probably to use relabeling techniques retrospectively, by post-processing the output (Marin, Mengersen, and Robert, 2005). However, in our case, we are interested to the prediction whose distribution parameters depend on the number of areas in each latent state. Then, we can not use the post-processing approach and we adopt the previous method using the mean of $\boldsymbol{\theta}$ as ordering constraint.

3.5 Results

We apply LMM SAE method to the data presented in Section 3.2. We run the algorithm with $k = 2, 3, 4$ latent states and we work with $k = 3$, following the proposed model selection aproach. In fact we obtain $p(\boldsymbol{\theta}|k = 2) = 24615.97$ and $p(\boldsymbol{\theta}|k = 3) = 25611.85$. With $k = 4$ the algorithm allocates areas in 3 groups and, thereby, it is not considered further. We run one Markov chain with 20000 iterations and then we consider a burn-in period of 10000 iterations.

Latent states can be seen associated to the severity of unemployment conditions, conditionally to the covariates. Tab. 3.4 reports the mode of the latent states classification for the data. We can see that areas are divided essentially into two groups, and the third is very small and quite constant in time. A temporal trend is present in the data. In fact we can see how group 1 decreases during time of observation, as group 2 increases. This temporal trend is also evident in the unemployment rate estimates (Fig. 3.3).

TABLE 3.4: Latent State Classification $k = 3$ (MCMC Mode).

u	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
1	578	568	562	558	552	549	543	539	527	523	520
2	31	41	46	50	56	59	65	69	81	86	89
3	2	2	3	3	3	3	3	3	3	2	2

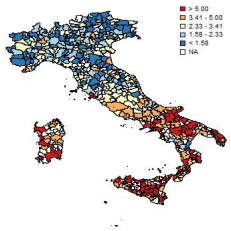
Estimated initial and transition probabilities are the following:

$$\hat{\boldsymbol{\pi}} = (0.930, 0.064, 0.005),$$

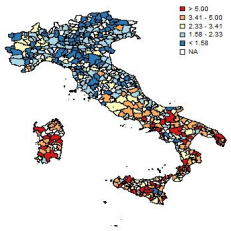
$$\hat{\boldsymbol{\Pi}} = \begin{pmatrix} 0.986 & 0.013 & 0.001 \\ 0.006 & 0.992 & 0.002 \\ 0.062 & 0.037 & 0.901 \end{pmatrix}.$$

The probability of changing latent states is very low. However it seems that the probability of moving to a worse state is higher than the the probability to move to a better condition. Informations regarding parameters estimates are in Appendix C.

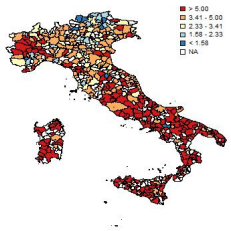
Fig. 3.3 shows the unemployment rate estimates in 2004, 2008 and 2014. Direct estimates, You Rao Gambino (YRG) (You, Rao, and Gambino, 2003) estimates with $\rho = 1$ and LMM estimates are compared. All the methods show how in 2008 the unemployment rate in Italy was lowest. At the beginning of the observation there is a quite clear division between north, center and south of Italy, with unemployment rates increasing going south. In 2014 unemployment rates are very high in all the country, according with the contemporary economic world crisis. Maps are built based on the quantiles of the direct estimates distribution in 2004. The whole set of maps is provided in Ap-



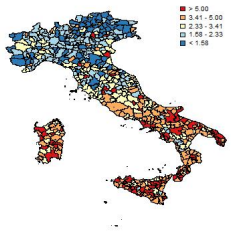
(a) Direct estimates 2004



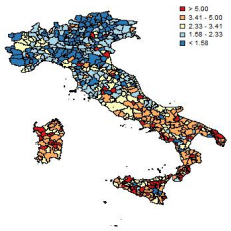
(b) Direct estimates 2008



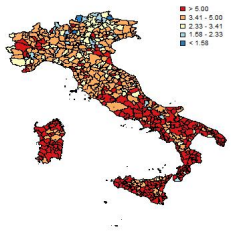
(c) Direct estimates 2014



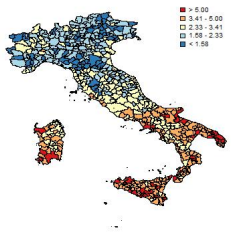
(d) YRG estimates 2004



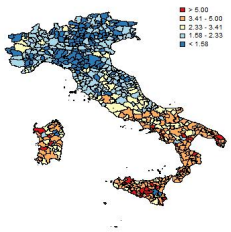
(e) YRG estimates 2008



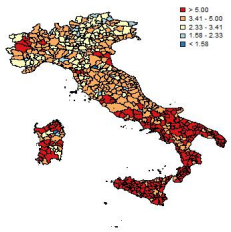
(f) YRG estimates 2014



(g) LMM estimates 2004



(h) LMM estimates 2008



(i) LMM estimates 2014

FIGURE 3.3: Unemployment rates estimates 2004-2008-2014.

pendix D.

One of the basic properties that we require for model-based SAE estimates is that they should have coefficient of variations lower than the coefficient of variations of corresponding direct estimates. In Tab.3.5 we reported the CVs for the LMM estimates classified following the Statistic Canada suggestion already use in Tab. 3.3 for direct estimates. For LMM estimates areas are classified mostly in the first two classes. As expected, when estimates are larger, as in the most recent years, CVs are smaller. We report also the CVs classification for the YRG model. Values in the last column from Tab. 3.6 are decreasing with time, and this provides evidence of temporal trend that is well captured in the Random walk model.

TABLE 3.5: Number of small areas with values of CV less than 16.6%, between 16.6% and 33.3% and over 33.3% for 3-state LMM estimator (from 2004 to 2014).

year	< 16.6	16.6-33.3	> 33.3
2004	149	380	82
2005	145	367	99
2006	126	357	128
2007	119	349	143
2008	135	357	119
2009	138	376	97
2010	163	381	67
2011	195	374	42
2012	251	340	20
2013	292	301	18
2014	307	287	17
2004-2014	2020	3869	832

3.5.1 Diagnostics for LMM SAE estimates

The aim of this diagnostics procedure is to validate the reliability of the LMM SAE estimates versus direct survey estimates (Srivastava, Sud, and Chandra, 2007; Brown et al., 2001).

If the model-based estimates are close to the small area values of interest, then unbiased direct estimators should behave like random variables whose expected values correspond to the values of the model-based estimates (Brown et al., 2001). The bias diagnostic is used to assess the deviation of the LMM estimates from the direct survey estimates. In fact, LMM estimates are expected to be biased predictors of the direct estimates. The LMM estimator will be biased if the relationship between the variable of interest and the auxiliary variables has been misspecified or misestimated. When the relationship has not been misspecified or misestimated, a linear relationship of the type

TABLE 3.6: Number of small areas with values of CV less than 16.6%, between 16.6% and 33.3% and over 33.3% for YRG estimator (from 2004 to 2014).

year	< 16.6	16.6-33.3	> 33.3
2004	128	377	106
2005	169	360	82
2006	196	343	72
2007	215	317	79
2008	219	318	74
2009	241	317	53
2010	269	296	46
2011	314	267	30
2012	371	224	16
2013	395	206	10
2014	370	231	10
2004-2014	2887	3256	578

$y = x$ is expected between the direct estimates and the model-based LMM estimates. Fig. 3.4 reports scatterplots of LMM estimates vs direct estimates for every year. Coefficients estimates of the intercept and the slope for the regression of direct estimates vs LMM estimates together with their standard errors are in Tab. 3.7. From the slope estimates we can notice that direct estimates for each year are sistematically higher than the LMM estimates but usually these differences increase when direct estimates are larger. This diagnostic method has to be taken carefully. When the variable of interest is a proportion as in our case, the denominator of the direct estimator is a random variable and so the proportion is a ratio estimator and hence possibly biased. We can compare the direct and model-based estimators of the proportion but we have to accept that the resulting ratio bias may slightly distort the interpretation of the diagnostic. The lower the coefficient of variation of the denominator of the small area proportion is, the lower the risk of bias in the direct estimate of the proportion is (Brown et al., 2001). Finally, this diagnostic procedure provides a way of looking for bias due to model misspecification but can not discover any bias in the direct estimates. Furthermore, LMM estimates always fall in the corresponding direct estimates 95% confidence interval.

We want LMM estimates to be close to the direct estimates when the direct estimates are good. For this reason, as a test for unconditional bias in the model-based LMM estimates we use a Wald goodness of fit statistic to test whether there is a significant difference between the expected values of the direct estimates and the LMM estimates. To evaluate this we compute the squared difference between the model estimates and the direct estimate which are weighted inversely by their variance and summed over all the areas. This statistic is compared with the quantiles of the χ^2 distribution with degrees of freeedom equal to the number of small areas. The goodness of fit statistics are reported in Tab. 3.8. None of the statistics shows evidence to reject

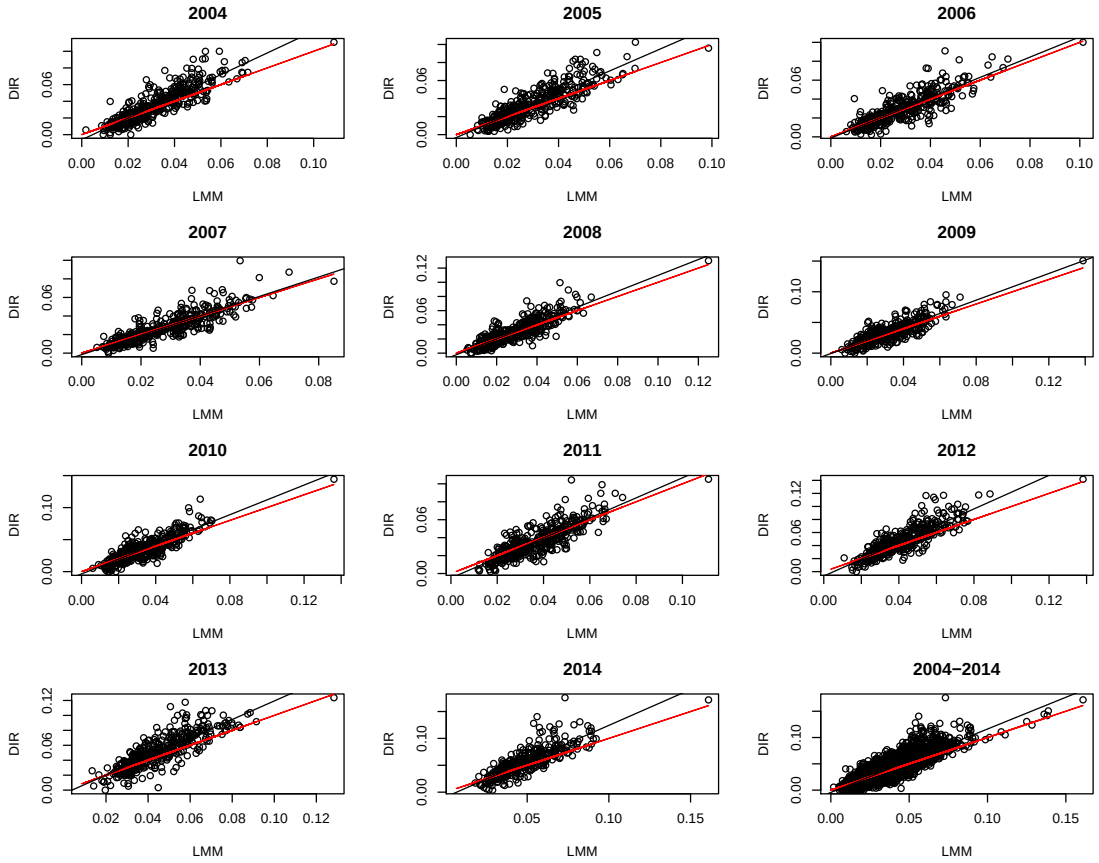


FIGURE 3.4: Bias scatterplots: direct estimates vs LMM estimates $45°$ line (red) and regression fitted line (black).

TABLE 3.7: OLS regression parameters (standard error) from bias scatterplots: direct estimates vs LMM estimates.

year	Intercept	Slope
2004	-0.006 (0.001)	1.301 (0.032)
2005	-0.003 (0.001)	1.230 (0.032)
2006	-0.001 (0.000)	1.071 (0.030)
2007	-0.002 (0.001)	1.045 (0.026)
2008	-0.002 (0.001)	1.012 (0.029)
2009	-0.000 (0.001)	1.086 (0.031)
2010	-0.003 (0.001)	1.157 (0.032)
2011	-0.005 (0.001)	1.117 (0.033)
2012	-0.007 (0.001)	1.287 (0.036)
2013	-0.005 (0.002)	1.240 (0.038)
2014	-0.009 (0.002)	1.333 (0.042)
2004-2014	0.010(0.001)	0.643 (0.004)

the hypothesis that the two sets of estimates are significally different.

Another diagnostic procedure is the comparison of estimates with data from the 15^{th} Italian Census. We consider unemployment rate from the 15^{th} Italian Census as the

TABLE 3.8: Goodness of fit statistic values with p-value: direct estimates vs LMM estimates.

year	Statistic	p-value
2004	173.227	0.99
2005	163.513	0.99
2006	126.382	≈ 1
2007	101.931	≈ 1
2008	133.988	≈ 1
2009	153.935	≈ 1
2010	159.443	≈ 1
2011	163.389	0.99
2012	205.301	0.99
2013	230.586	0.98
2014	260.370	0.98
2004-2014	1872.064	1

TABLE 3.9: AARE compared with the 15^{th} Census 2011.

Method	OBS	NO OBS	ALL
DIR	0.401	na	0.401
LMM	0.373	0.391	0.378
YRG	0.349	0.491	0.389

true value and we evaluate the difference between direct estimates and LMM estimates in 2011 and the Census value. To do that we consider the Average Absolute Relative Error (AARE). AARE is computed as

$$AARE = \sum_{i=1}^{m} \frac{|\hat{\theta}_i - Cens_i|}{Cens_i}.$$

for each estimates. Values are reported in Tab. 3.9. The model-based estimates have smaller AARE values than the direct estimates. It is also interesting to look at the distribution of AARE. In particular, we look at the higher quantiles of the AARE distribution (Fig. 3.5), which are associated with higher errors. The values for LMM are much smaller than those for direct estimates for these observed areas. We want to investigate also the YRG behaviour in the higher quantile of AARE distributon and compare it with LMM. They are reported in Fig. 3.6. We can see that the two methods have a similar behaviour. In addition, LMM has an overall smaller AARE than YRG, and this is particularly due to the better performances of LMM in out of sample areas (0.391 vs 0.491).

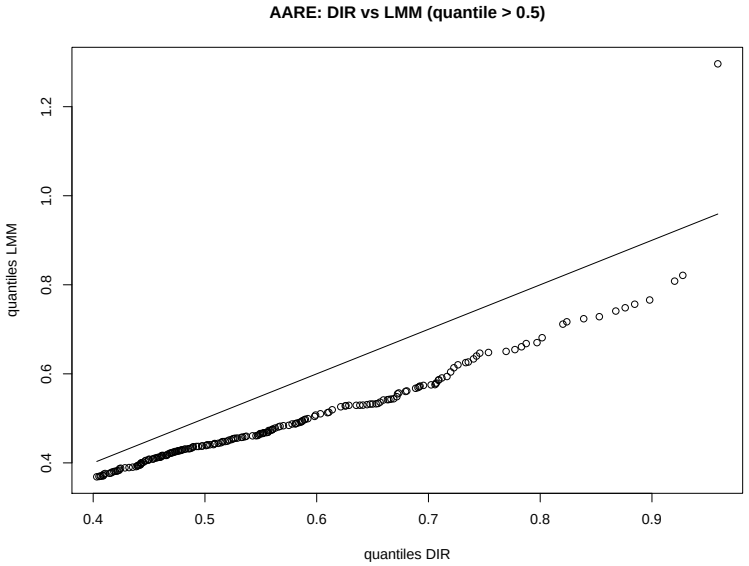


FIGURE 3.5: AARE compared with the 15^{th} Census 2011: LMM estimates vs direct estimates (quantile>0.5).

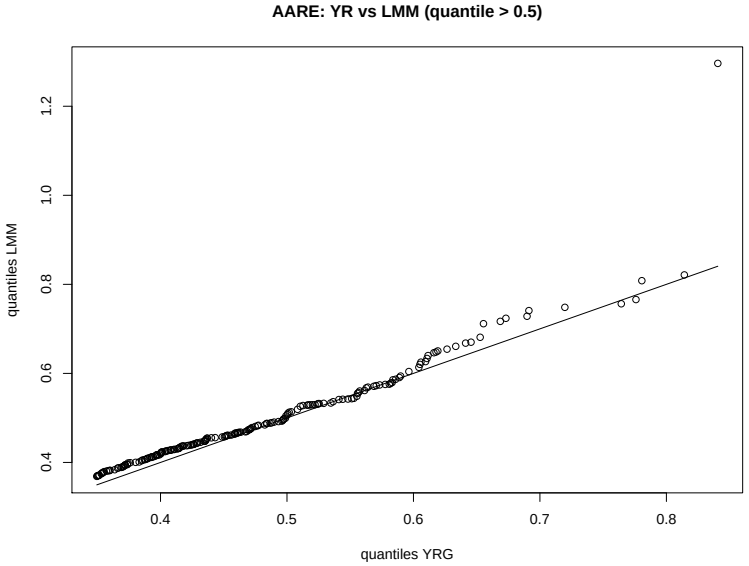


FIGURE 3.6: AARE compared with the 15^{th} Census 2011: LMM estimates vs YRG estimates (quantile>0.5).

3.6 Conclusion

A new time-series SAE area level model has been developed and applied to italian LFS from 2004 to 2014 in order to estimate unemployment rates. Area-level SAE models consider a sampling and a linking model. In our context, a LMM is used as the linking model. In a hierarchical Bayesian framework the sampling model is introduced as the highest level of hierarchy. Under the assumption of normality for the response variable,

the model is estimated using an augmented Gibbs sampling.

The realiability of the LMM estimates versus direct estimates has been validated through several diagnostic procedures. After that we compare direct estimates and LMM estimates with unemloyment rates from the 15^{th} Italian Census and evaluate the differences. LMM has an overall smaller AARE than YRG, especially thanks of the better performances of LMM in out of sample areas.

The model-based method has been found to be effective for developing LMAs level estimates of unemployment rates and for most of the areas the reduction in the coefficient of variation is quite evident. LMM estimator seems to be more accurate than direct and YRG estimator compared to census data. An advantage of this methodology is that it provide an area classification.

It could be interesting to fit the model to quarterly estimates and aggregate them at the end of the procedure, like for direct estimates. In fact, accounting for all available data could actually improve model-based estimates precision.

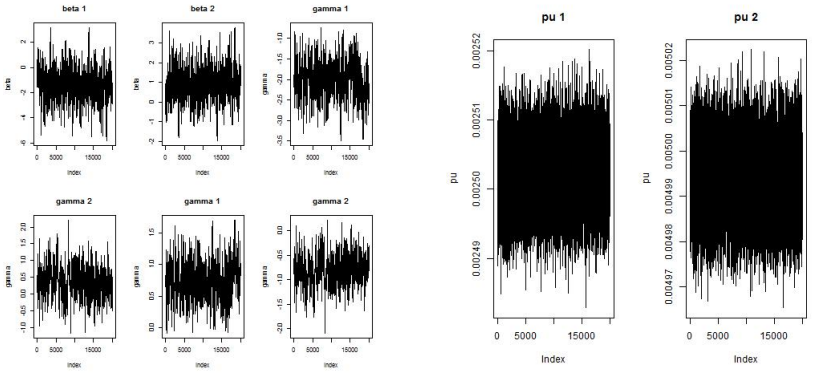
Moreover, due to the fact that a temporal trend is evident in the data, we could think of inserting in the model a temporal trend covariate to better estimate the rates and the latent states. Further investigation is required to understand where it is more feasible to insert such a covariate, if in the measurement or in the latent model. In addition, area spatial structure could be inserted in the latent model following the approach developed in Chapter 2. Finally, a more effective approach to label switching has to be developed, in order to better estimate the parameters of the model and to obtain a better latent states classificaton.

The proposed model provides a very flexible modeling framework. It could be extended also in a cross-sectional framework, using area spatial correlation informations, and it could consider different distributions for the manifest variables, like Poisson, Binomial and Multinomial responses. In this last scenario we could fit unmatched sampling and linking models. The univariate model proposed can account for measurement error, but the extension to multivariate framework could be also possible, taking into account the conditional independence problem.

Appendix A

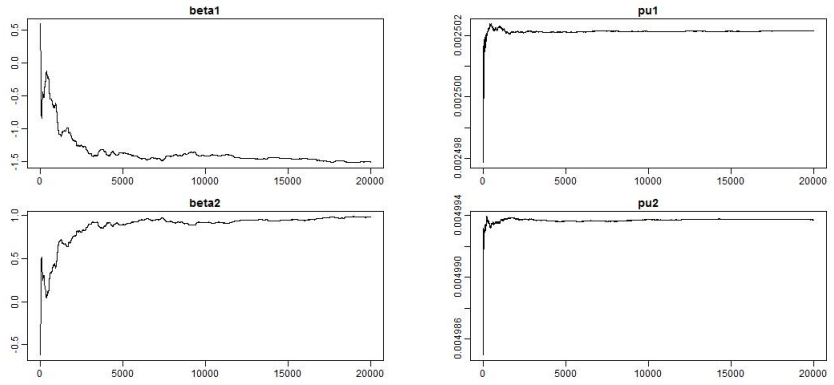
MCMC chains

A.1 Scenario 1a



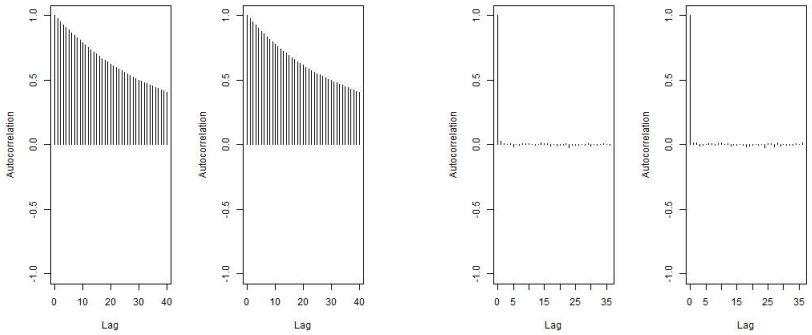
(a) Latent model parameters (b) Measurement model parameters

FIGURE A.1: Scenario 1a: Trace plots.



(a) β (b) Measurement model parameters

FIGURE A.2: Scenario 1a: β and measurement model parameters mean plots.



(a) $\boldsymbol{\beta}$ (b) Measurement model parameters

FIGURE A.3: Scenario 1a: $\boldsymbol{\beta}$ and measurement model parameters autocorrelation plots.

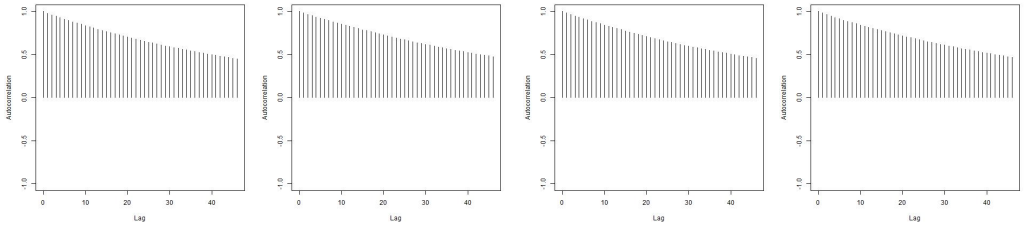
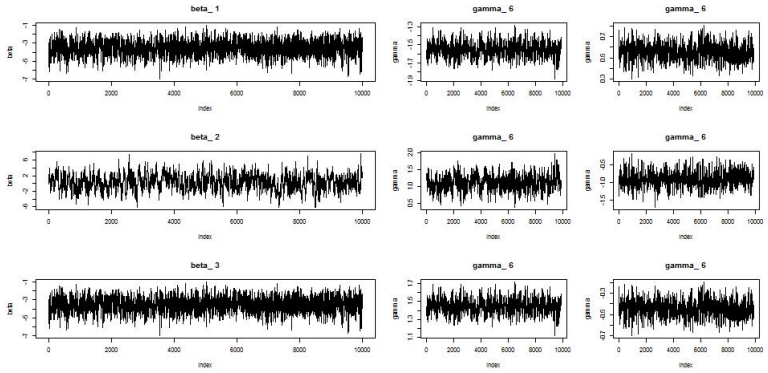
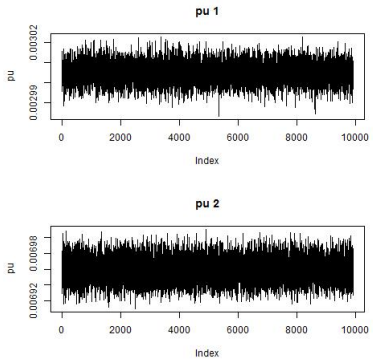


FIGURE A.4: Scenario 1a: $\boldsymbol{\Gamma}$ autocorrelation plot.

A.2 Scenario 2a

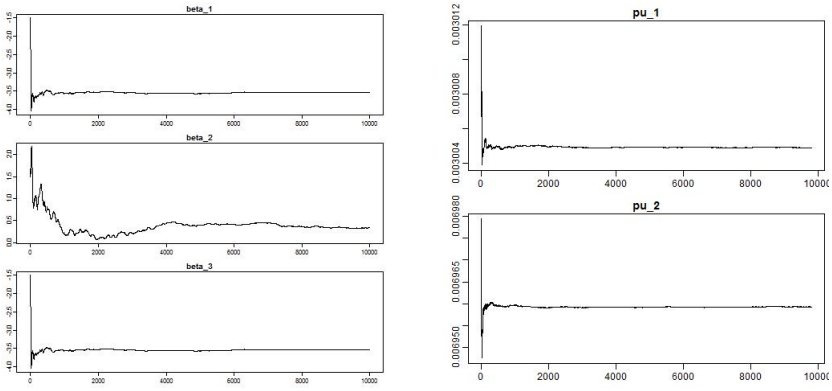


(a) Latent model parameters



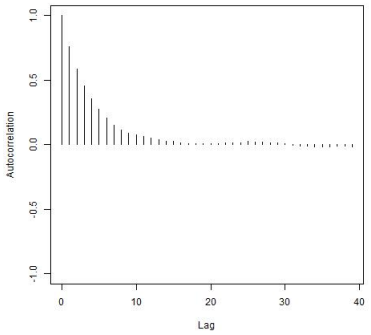
(b) Measurement model parameters

FIGURE A.5: Scenario 2a: Trace plots parameters.

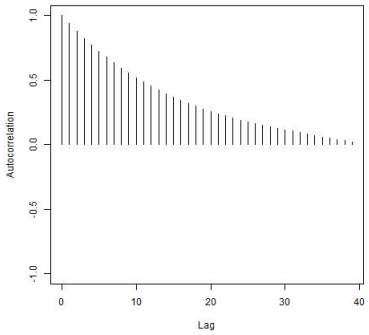


(a) β (b) Measurement model parameters

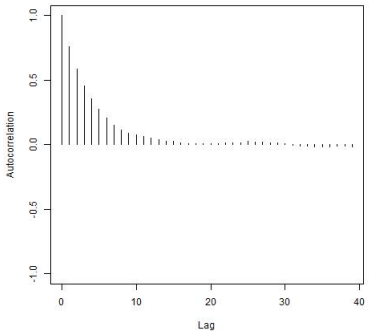
FIGURE A.6: Scenario 2a: β and observed succes probabilities.



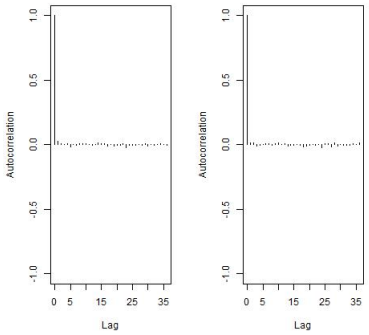
(a) β_{01}



(b) β_{11_1}

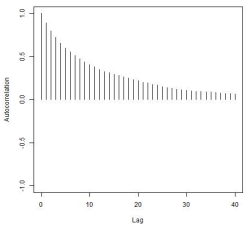


(c) β_{11_2}

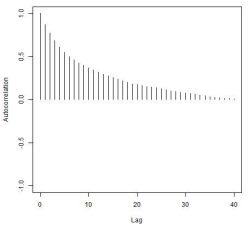


(d) Measurement model parameters

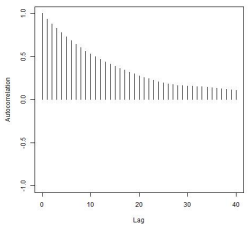
FIGURE A.7: Scenario 2a: $\boldsymbol{\beta}$ and observed succes probabilities autocorrelation plots.



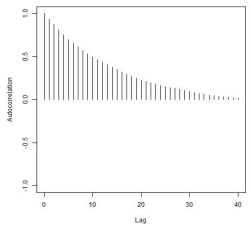
(a) γ_{012}



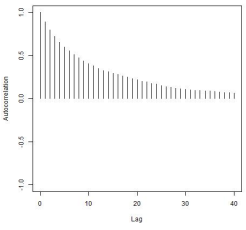
(b) γ_{021}



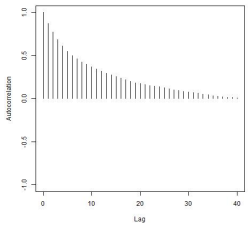
(c) γ_{21_1}



(d) γ_{21_2}



(e) γ_{12_1}



(f) $\gamma\gamma_{12_2}$

FIGURE A.8: Scenario 2: $\boldsymbol{\Gamma}$ autocorrelation plots.

Appendix B

Full Conditional Distributions

The Gibbs full conditional $[\boldsymbol{\beta}_u|\boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}]$ can be written as

$$f(\boldsymbol{\beta}_u|\boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \frac{f(\boldsymbol{\beta}_u, \boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\int f(\boldsymbol{\beta}_u, \boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) d\boldsymbol{\beta}} \propto f(\boldsymbol{\beta}_u, \boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) \quad \text{(B.1)}$$

because the denominator is constant with respect to $\boldsymbol{\beta}$. This posterior distribution is proportional to the product of likelihood function and prior distribution. Considering only terms involving $\boldsymbol{\beta}$ and σ^2 we have

$$\begin{aligned}
f(\boldsymbol{\beta}_u, \boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) \propto & \frac{1}{(\sigma_u^2)^{N_u/2}} \exp\Big(-\frac{1}{2\sigma_u^2}(\boldsymbol{\theta}_u - \boldsymbol{X}\boldsymbol{\beta}_u)^T(\boldsymbol{\theta}_u - \boldsymbol{X}\boldsymbol{\beta}_u)\Big) \times \\
& \times \frac{1}{(\sigma_u^2)^{(k+1)/2}|\boldsymbol{\Lambda}_0|^{-1/2}} \exp\Big(-\frac{1}{2\sigma_u^2}(\boldsymbol{\beta}_u - \boldsymbol{\eta}_{0,u})^T\boldsymbol{\Lambda}_{0,u}(\boldsymbol{\beta}_u - \boldsymbol{\eta}_{0,u})\Big) \times \\
& \times \frac{b_{0,u}^{a_{0,u}}}{\Gamma(a_{0,u})} \frac{1}{(\sigma_u^2)^{a_{0,u}+1}} \exp\Big(-\frac{b_{0,u}}{\sigma_u^2}\Big) = \\
& = const \frac{1}{(\sigma_u^2)^{(k+1)/2}} \exp\Big(-\frac{1}{2\sigma_u^2}(\boldsymbol{\beta}_u - \boldsymbol{\eta}_{1,u})^T\boldsymbol{\Lambda}_{1,u}(\boldsymbol{\beta}_u - \boldsymbol{\eta}_{1,u})\Big) \times \\
& \times \frac{1}{(\sigma_u^2)^{N_u/2+a_{0,u}+1}} \exp\Big(-\frac{b_{1,u}}{\sigma_u^2}\Big) \propto \\
& \mathrm{N}(\boldsymbol{\beta}; \boldsymbol{\eta}_{1,u}, \sigma_u^2\Lambda_{1,u}) \times \mathrm{InvGamma}(\sigma_u^2; a_{1,u}, b_{1,u})
\end{aligned}$$

with

- $\boldsymbol{\eta}_{1,u} = \boldsymbol{\Lambda}_{1,u}^{-1} \sum_{it} \boldsymbol{x}_{it}\theta_{it} I(U_{it} = u) + \boldsymbol{\Lambda}_{0,u}\boldsymbol{\eta}_{0,u}$
- $\boldsymbol{\Lambda}_{1,u} = \sum_{it} \boldsymbol{x}_{it}\boldsymbol{x}_{it}^T I(U_{it} = u) + \boldsymbol{\Lambda}_{0,u}$
- $a_{1,u} = a_0 + N_u/2$
- $b_{1,u} = b_0 + \frac{1}{2}(\sum_{it} \theta_{it}^2 I(U_{it} = u) + \boldsymbol{\eta}_0^T\boldsymbol{\Lambda}_0\boldsymbol{\eta}_0 - \boldsymbol{\eta}_{1,u}^T\boldsymbol{\Lambda}_{1,u}\boldsymbol{\eta}_{1,u})$

The full conditional distribution of $\boldsymbol{\beta}_u$ and σ_u^2 are proportional to the joint posterior. By ignoring factors not depending on the parameter of interest, it can be seen, that the full conditionals of $\boldsymbol{\beta}_u$ and σ_u^2 are proportional to multivariate and the inverse gamma

distributions which make the joint posterior.

Retaining terms involving only $\boldsymbol{\theta}$

$$
\begin{aligned}
f(\theta_{it}|\boldsymbol{\beta}_u, \boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\sigma}_u^2, \widehat{\theta}) \propto \exp\Big(-\frac{1}{2}\Big(\frac{\theta_{it}\theta_{it}}{\gamma_{it}\phi_{it}} - 2\frac{\theta_{it}\widehat{\theta_{it}}}{\phi_{it}} - 2\frac{\theta_i\widehat{\theta}_i}{\phi_i} - 2\frac{\theta_{it}\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{u(it)}}{\sigma_u^2}\Big)\Big) = \\
= \exp\Big(-\frac{1}{2\gamma_t\phi_{it}}\Big[\gamma_{it}\hat{\theta}_{it} + (1-\gamma_{it})\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{u(it)}\Big]\Big) \sim \\
\sim \mathrm{N}(\widehat{\theta}_{it}^B(\boldsymbol{u}, \boldsymbol{\beta}_{u(it)}, \sigma_u^2), \gamma_{it}\psi_{it})
\end{aligned}
$$

Appendix C

Parameter Estimates

In the following, we report β MCMC output estimates for each latent states u={1,2,3}. We notice (Fig.C.1 - Fig. C.9) that some label switching problem still affects the β parameters estimates even after a burn-in period of 10000 iterations, mostly for less numerous groups. It is necessary to work on this point to obtain an event more better latent states classification. We also report the MCMC output of four LMA during the observation period. We choose four areas with different behaviour. In the first group (Fig. C.10) direct estimates are computed for each year and their behaviour does not seem to be influenced by the world economic crisis. Then we present the results for a LMA whose unemployment rate increases (Fig. C.11). Finally, two LMAs with missing values are reported. The one in Fig. C.12 has no direct estimates in the first three years of observation, the last (Fig. C.13) has never been sampled. It seems that label switching does not influence the estimates, except when direct estimates are very high as in the last three periods Fig. C.11.

Auxiliary variables available for these data are the following:

- population rates: Females $\times$ 14 age classes (0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34,35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65+)$\rightarrow$ from β_1 to β_{14};
- population rates: Males $\times$ 14 age classes (0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34,35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65+) $\rightarrow$ from β_{15} to β_{28};
- *Cultural Vocation* : a qualitative nominal variable with five categories (insert as dummy variables)
  - Great beauty (70 LMAs that are both artistic and naturalistic centers and which have a manufacturing based on cultural connotation);
  - Potential heritage (138 LMAs that are artistic and naturalistic centers but where the entrepreneurial dimension is less developed) $\rightarrow \beta_{29}$;
  - Cultural activity (138 LMAs that have a manufacturing base to cultural connotation even if they are not artistic or naturalistic centers) $\rightarrow \beta_{30}$;

  - Tourism (194 LMAs that have a poorly developed cultural dimension or industry but they are a tourist destination)→ β_{31};
  - Cultural remotness (71 LMAs that are under the EU standard in any cultural or naturalistic classes) → β_{32};

- *Prevalent Specialization*: a qualitative nominal variable with four levels (insert as dummy variables):
  - not specialized,
  - not manufacturing → β_{33};
  - made in Italy → β_{34};
  - industrial → β_{35}.

Due to the identifiability problem, the intercept of the model and the first dummy variables created from *Cultural Vocation* and *Prevalen Specialization* are not in the model.

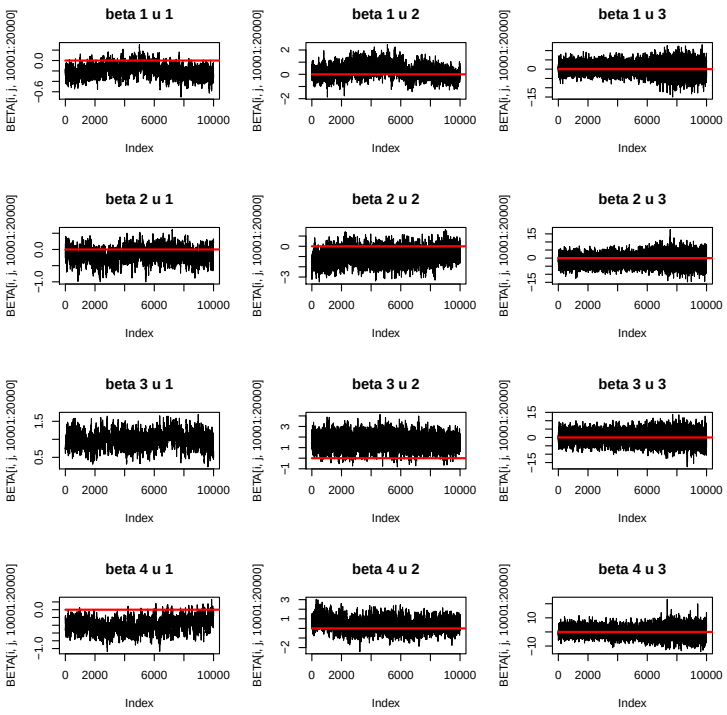


FIGURE C.1: $\beta_1 - \beta_4$ MCMC output.

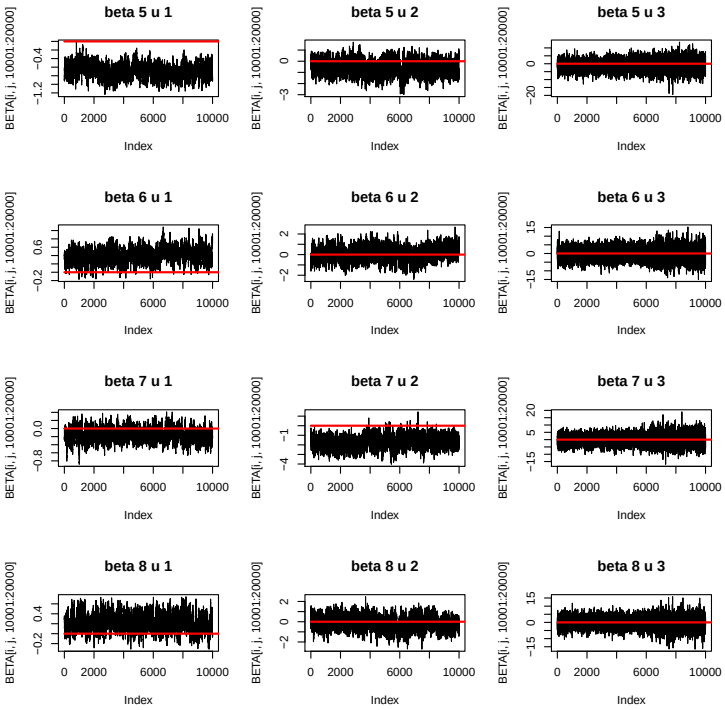


FIGURE C.2: $\beta_5 - \beta_8$ MCMC output.

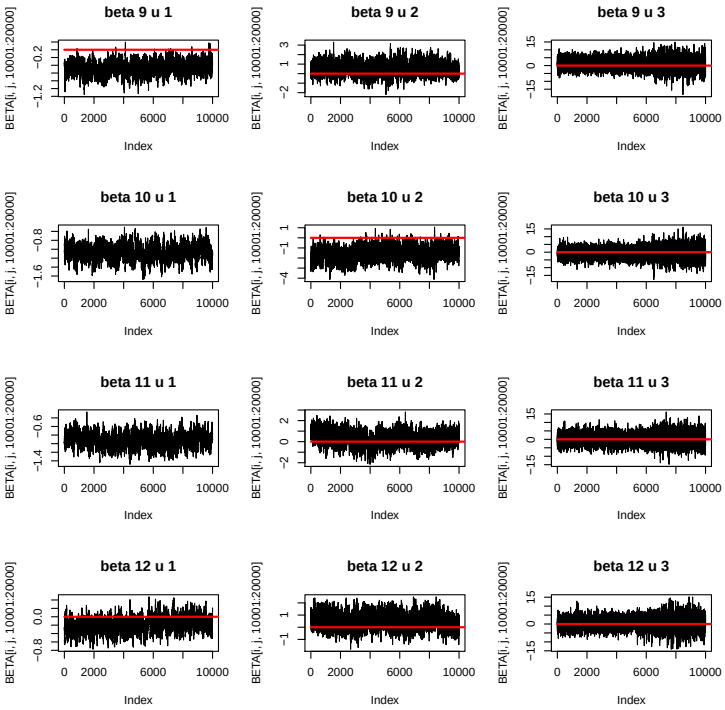


FIGURE C.3: $\beta_9 - \beta_{12}$ MCMC output.

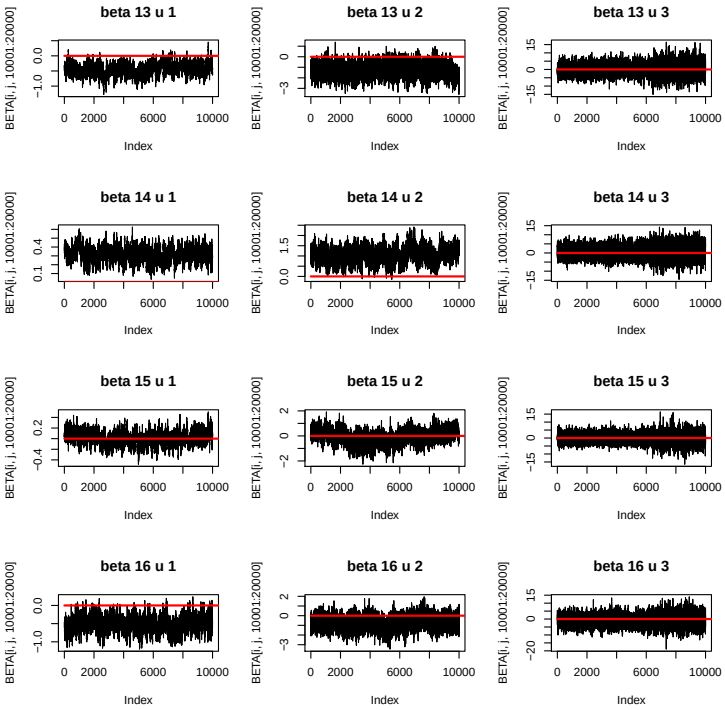


FIGURE C.4: $\beta_{13} - \beta_{16}$ MCMC output.

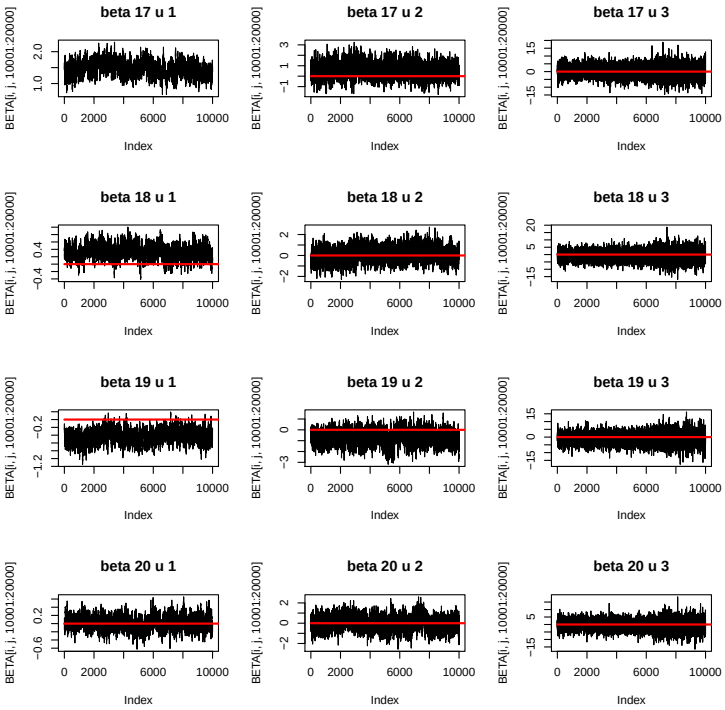


FIGURE C.5: $\beta_{17} - \beta_{20}$ MCMC output.

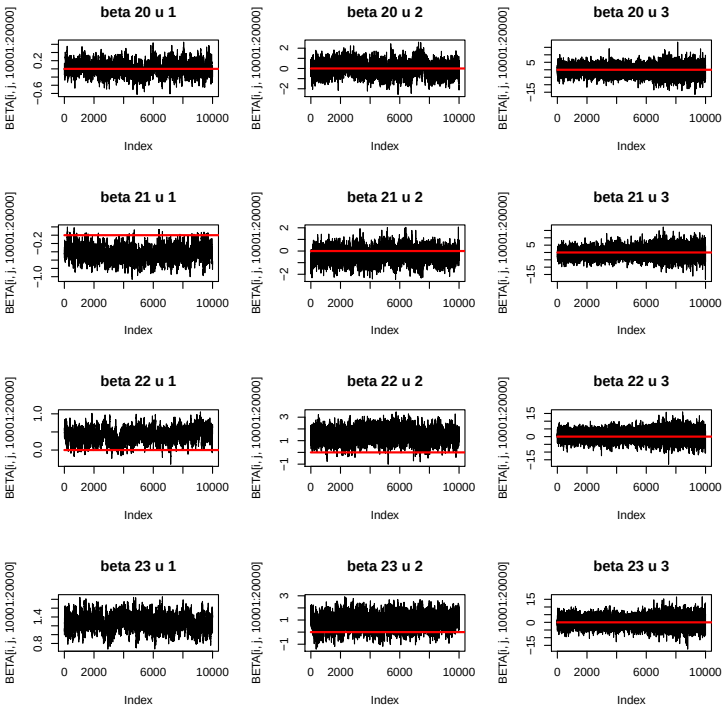


FIGURE C.6: $\beta_{21} - \beta_{24}$ MCMC output.

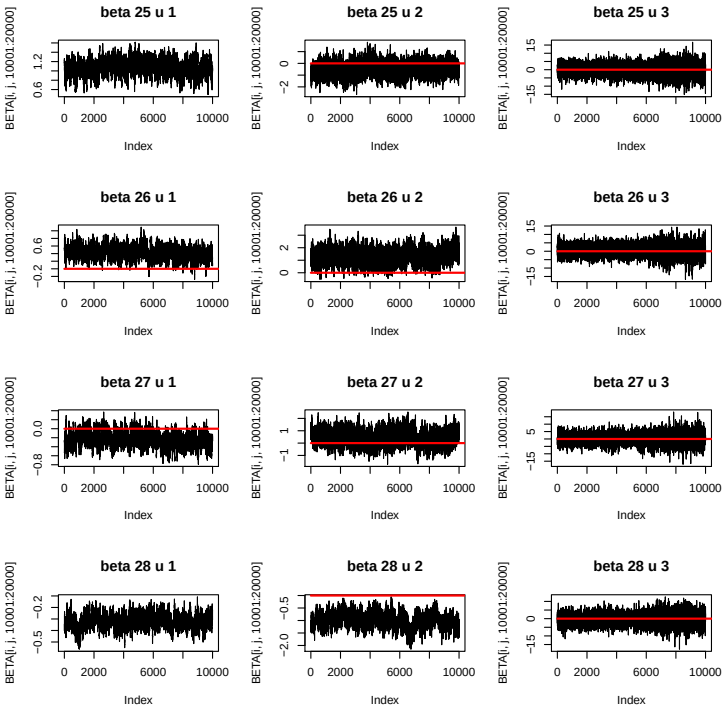


FIGURE C.7: $\beta_{25} - \beta_{28}$ MCMC output.

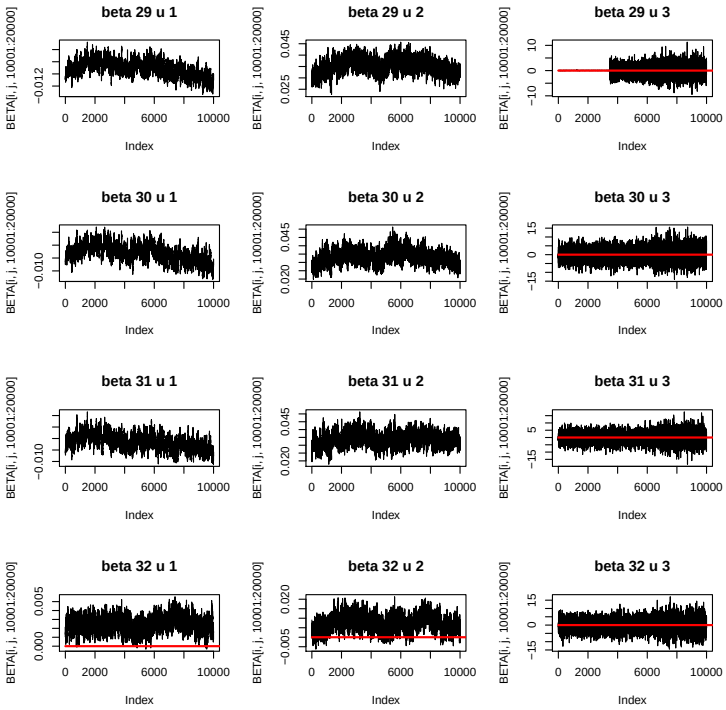


FIGURE C.8: $\beta_{29} - \beta_{32}$ MCMC output.

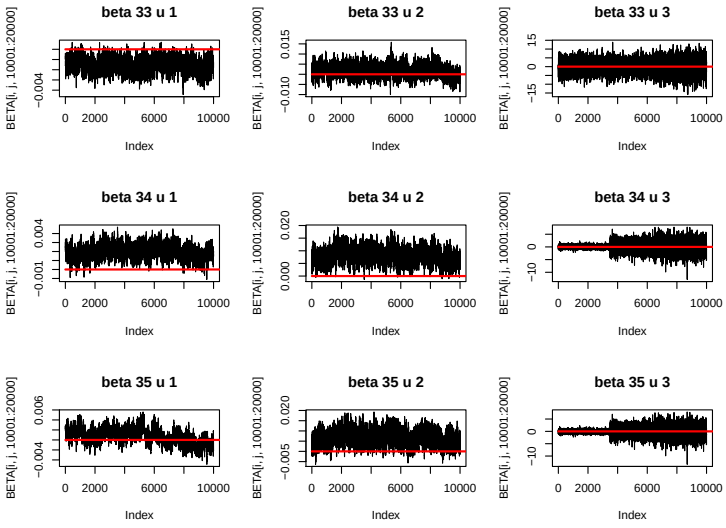


FIGURE C.9: $\beta_{32} - \beta_{35}$ MCMC output.

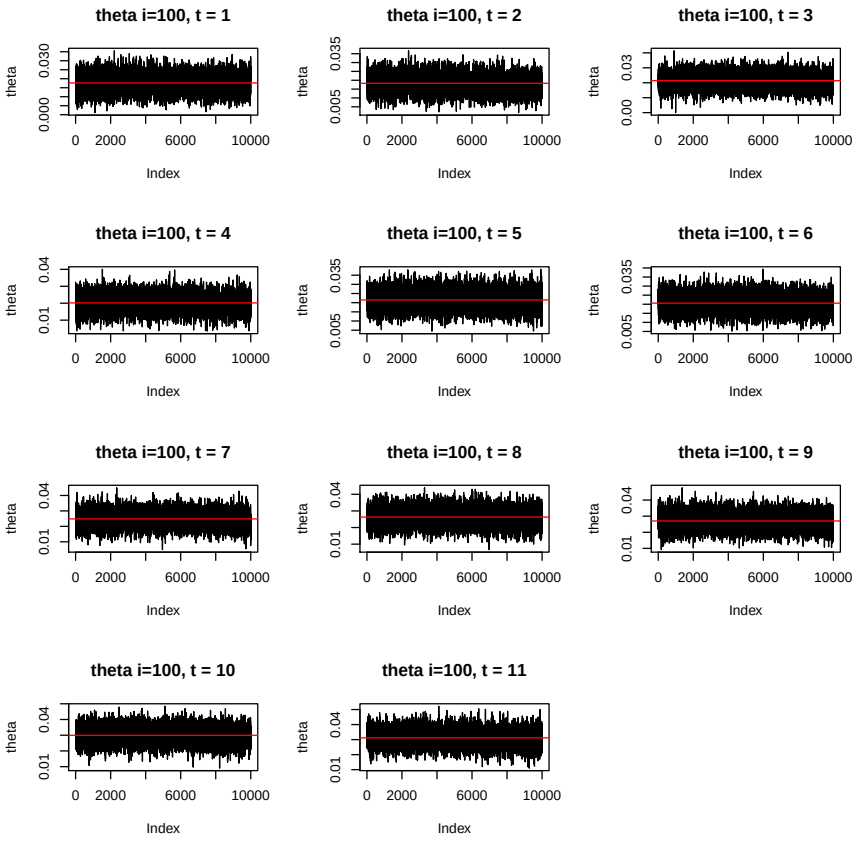


FIGURE C.10: LMA with no missing direct estimates and where direct estimates are quite constant.

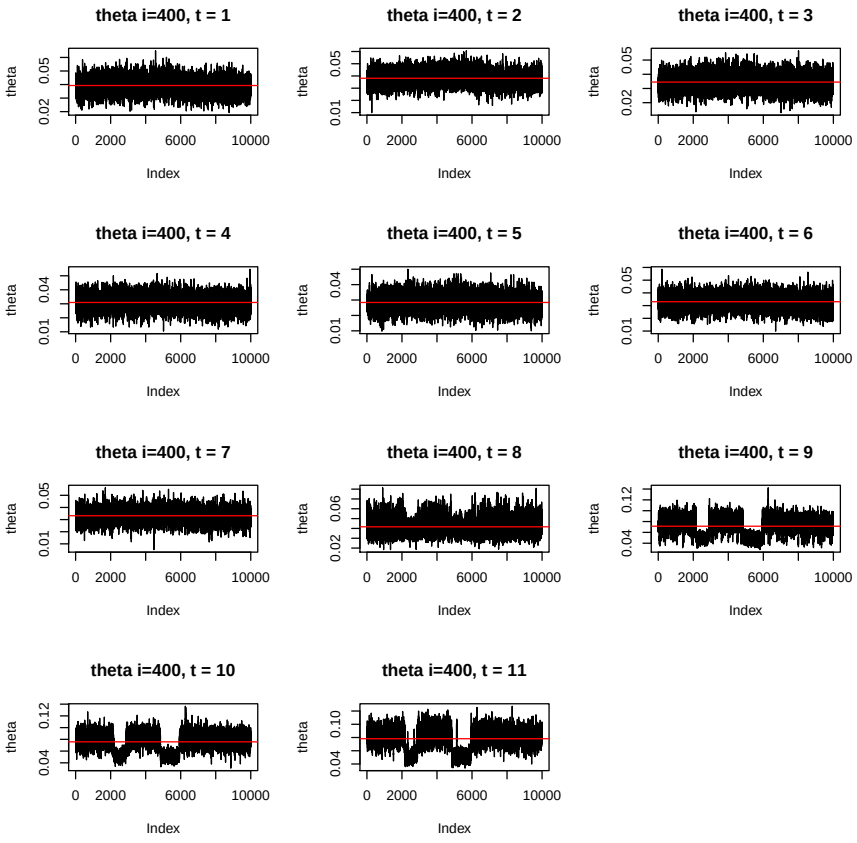


FIGURE C.11: LMA with no missing direct estimates and a strong temporal trend.

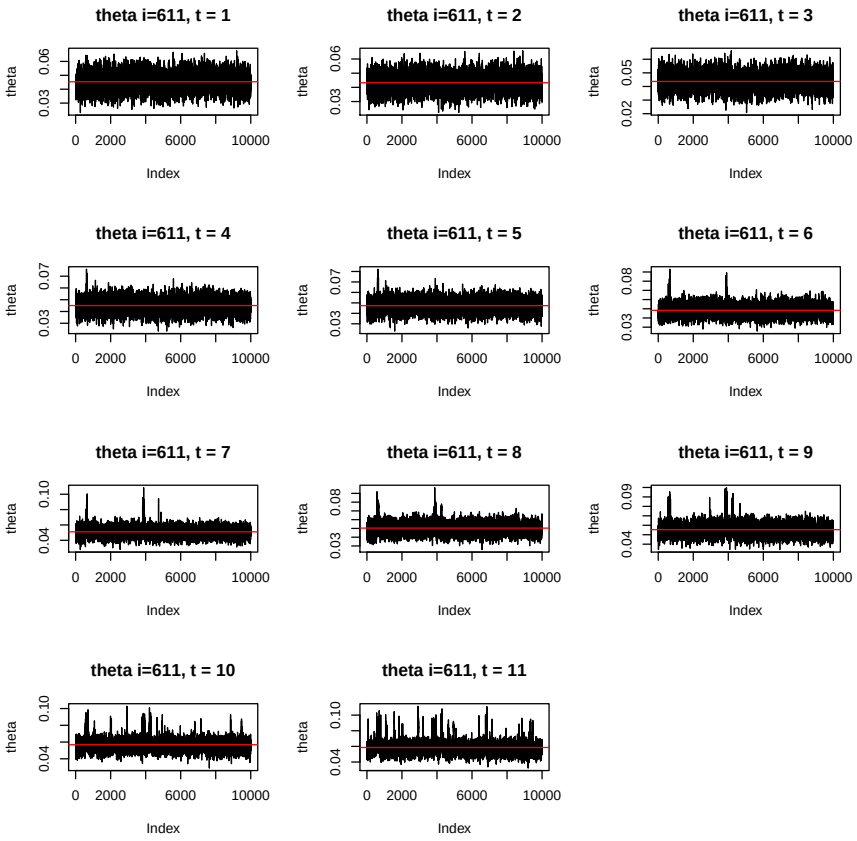


FIGURE C.12: LMA with missing direct estimates for the first 3 years.

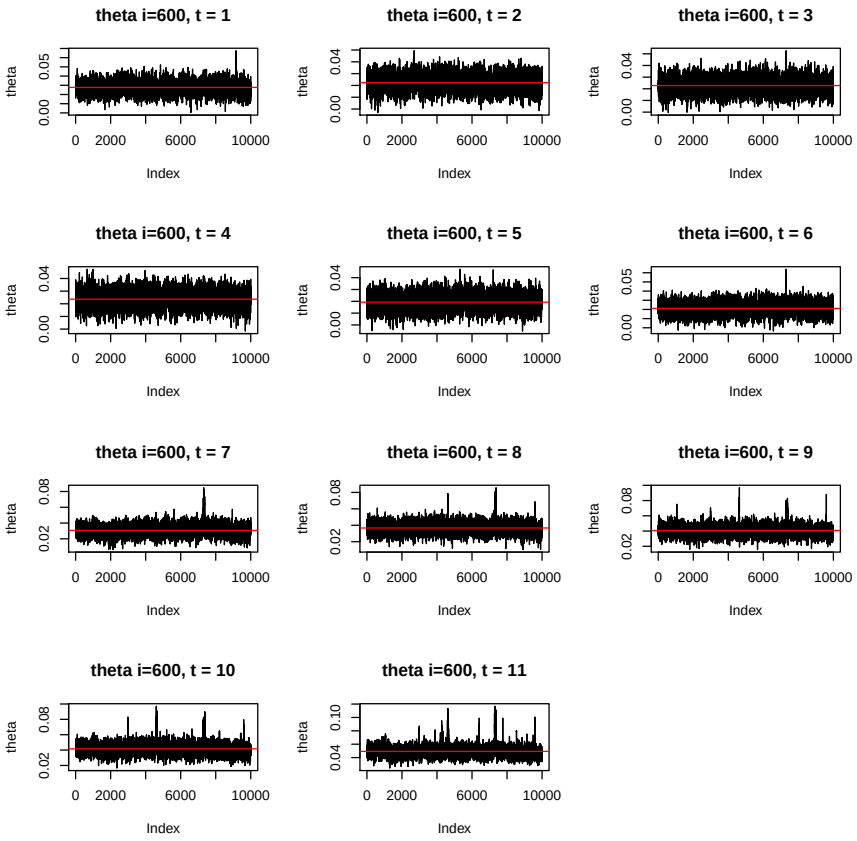
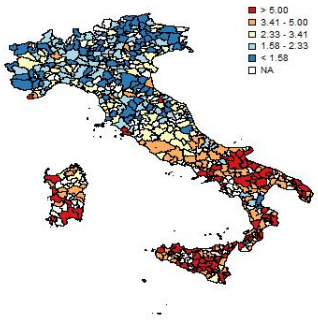


FIGURE C.13: LMA without any observations.

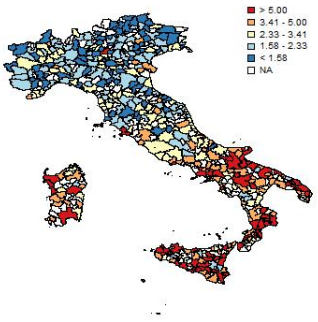
Appendix D

Unemployment estimates maps

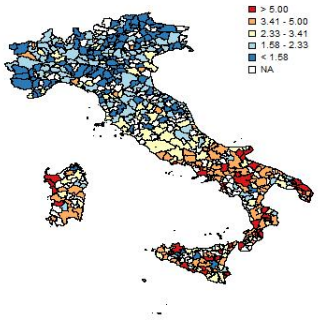
In this appendix we report the Unemployment rate estimates maps for each year in the period of observation. They are compared based on the quantile of direct survey estimates at 2004. Direct estimates, You Rao Gambino estimates and LMM SAE with $k = 3$ estimates are reported. A temporal trend is evident. As expected, model-based estimates partially smooth direct estimates.



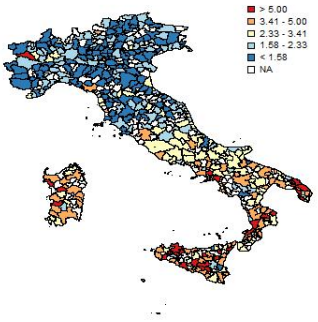
(a) Direct estimates 2004



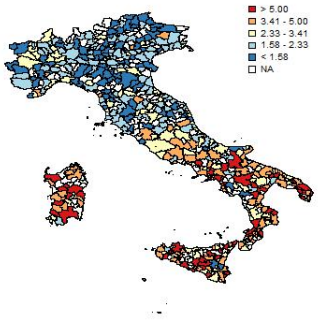
(b) Direct estimates 2005



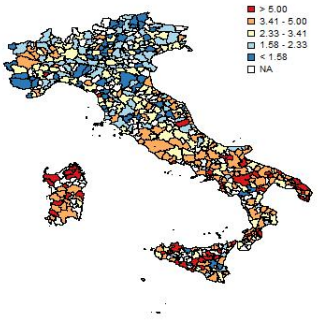
(c) Direct estimates 2006



(d) Direct estimates 2007

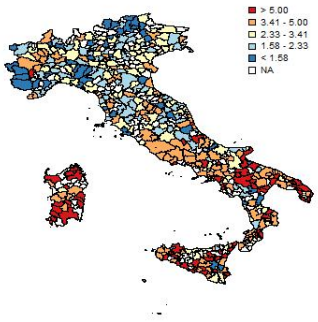


(e) Direct estimates 2008

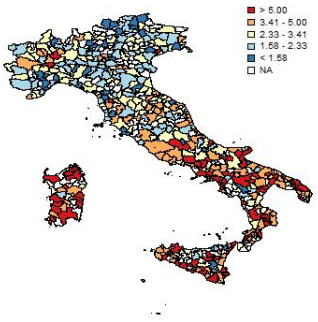


(f) Direct estimates 2009

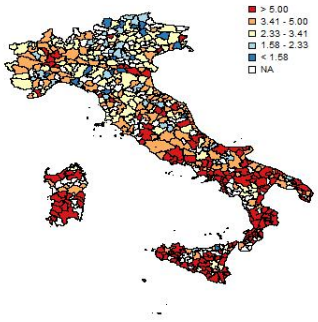
FIGURE D.1: Unemployment direct estimates from 2004 to 2009.



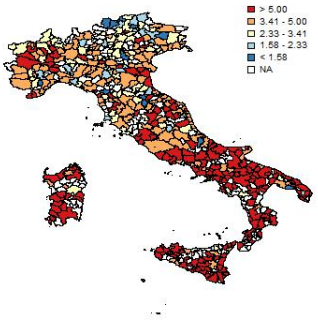
(a) Direct estimates 2010



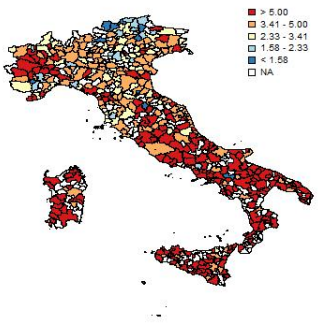
(b) Direct estimates 2011



(c) Direct estimates 2012

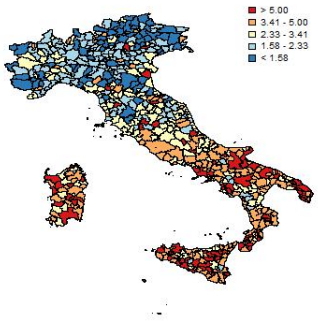


(d) Direct estimates 2013

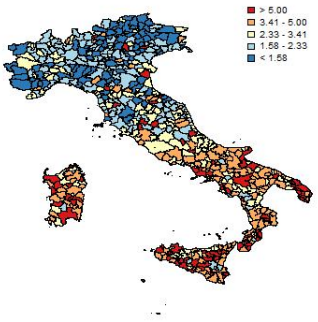


(e) Direct estimates 2014

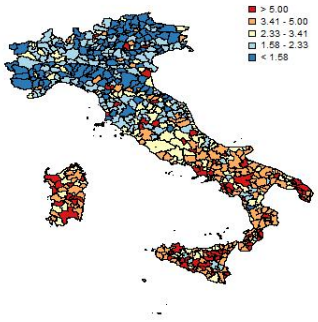
FIGURE D.2: Unemployment direct estimates from 2010 to 2014.



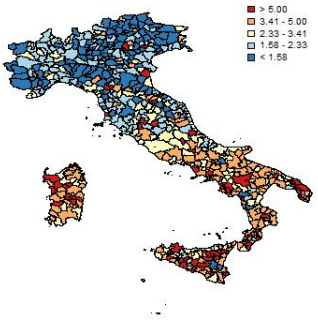
(a) YRG estimates 2004



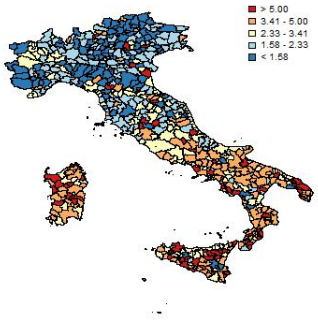
(b) YRG estimates 2005



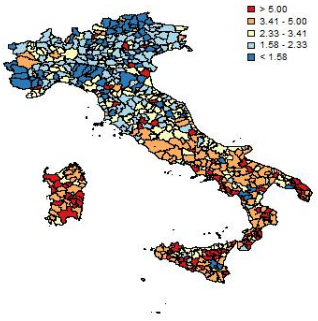
(c) YRG estimates 2006



(d) YRG estimates 2007

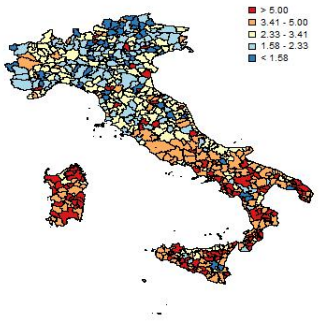


(e) YRG estimates 2008

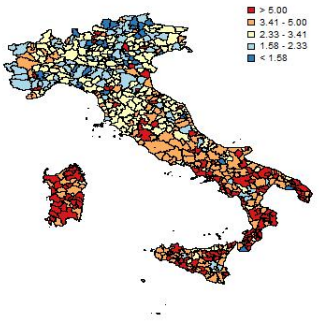


(f) YRG estimates 2009

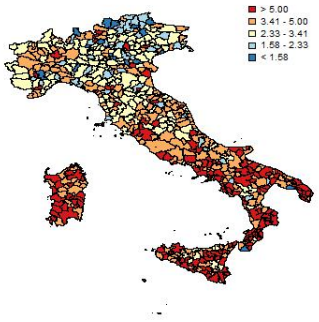
FIGURE D.3: Unemployment YRG estimates from 2004 to 2009.



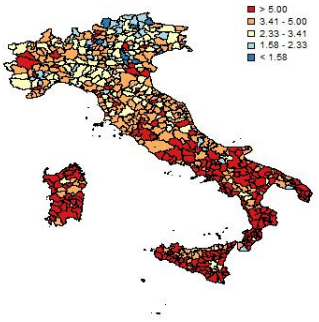
(a) YRG estimates 2010



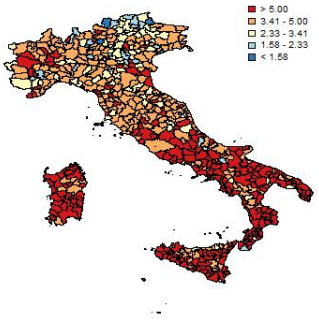
(b) YRG estimates 2011



(c) YRG estimates 2012

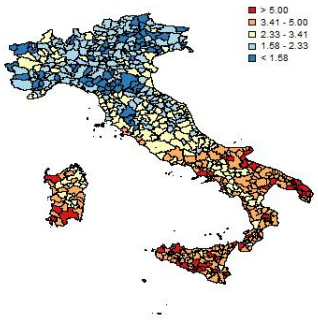


(d) YRG estimates 2013

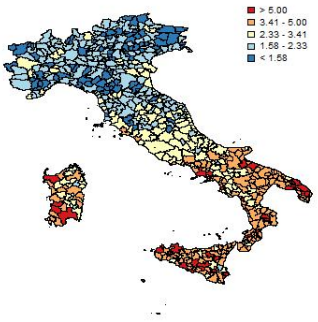


(e) YRG estimates 2014

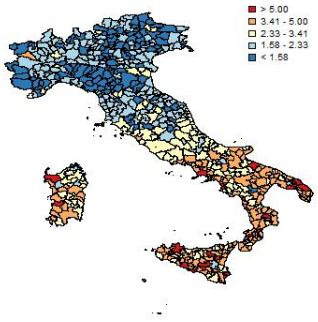
FIGURE D.4: Unemployment YRG estimates from 2010 to 2014.



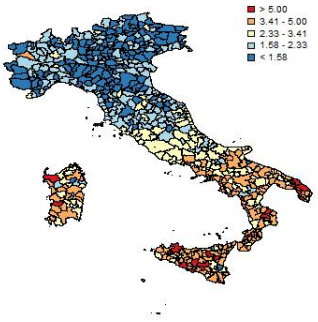
(a) LMM estimates 2004



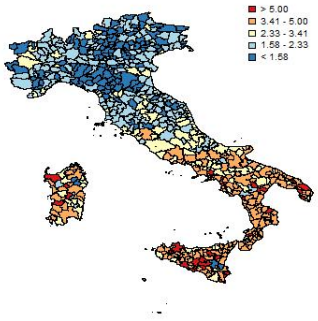
(b) LMM estimates 2005



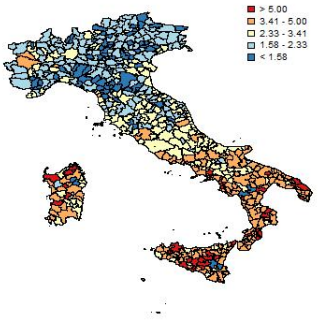
(c) LMM estimates 2006



(d) LMM estimates 2007

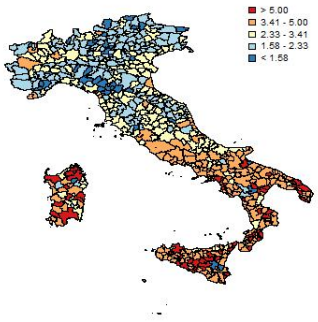


(e) LMM estimates 2008

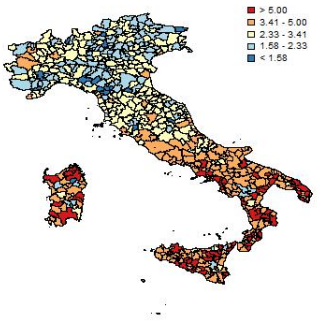


(f) LMM estimates 2009

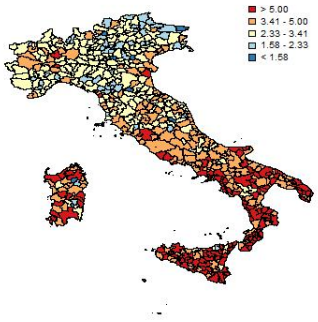
FIGURE D.5: Unemployment LMM estimates from 2004 to 2009.



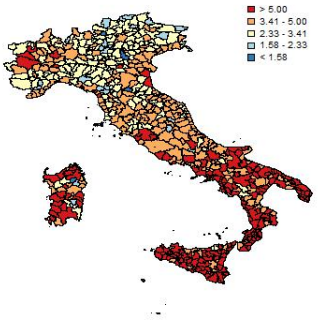
(a) LMM estimates 2010



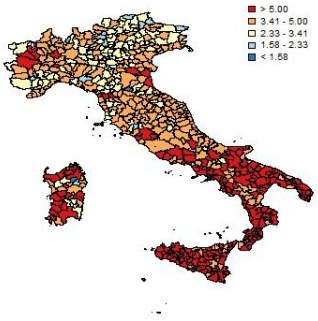
(b) LMM estimates 2011



(c) LMM estimates 2012



(d) LMM estimates 2013



(e) LMM estimates 2014

FIGURE D.6: Unemployment LMM estimates from 2010 to 2014.

Bibliography

Alfó, Marco, Luciano Nieddu, and Donatella Vicari (2009). "Finite mixture models for mapping spatially dependent disease counts". In: *Biometrical Journal* 51.1, pp. 84–97.

Bartolucci, Francesco and Alessio Farcomeni (2009). "A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure". In: *Journal of the American Statistical Association* 104.486, pp. 816–831.

Bartolucci, Francesco, Alessio Farcomeni, and Fulvia Pennoni (2014). "Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates". In: *Test* 23.3, pp. 433–465.

Bartolucci, Francesco, Monia Lupparelli, and Giorgio E Montanari (2009). "Latent Markov model for longitudinal binary data: an application to the performance evaluation of nursing homes". In: *The Annals of Applied Statistics*, pp. 611–636.

Bartolucci, Francesco and Silvia Pandolfi (2011). "Bayesian inference for a class of latent Markov models for categorical longitudinal data". In: *arXiv preprint arXiv:1101.0391*.

Bartolucci, Francesco, Fulvia Pennoni, and Brian Francis (2007). "A latent Markov model for detecting patterns of criminal activity". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170.1, pp. 115–132.

Besag, Julian, Jeremy York, and Annie Mollié (1991). "Bayesian image restoration, with two applications in spatial statistics". In: *Annals of the institute of statistical mathematics* 43.1, pp. 1–20.

Boonstra, Harm Jan (2014). "Time-series small area estimation for unemployment based on a rotating panel survey". In:

Boys, RJ and DA Henderson (2003). *Data augmentation and marginal updating schemes for inference in hidden Markov models*. Tech. rep. Technical report, Univ. Newcastle.

Brown, Gary et al. (2001). "Evaluation of small area estimation methods-An application to unemployment estimates from the UK LFS". In: *Proceedings of Statistics Canada Symposium*.

Carlin, Bradley P and Siddhartha Chib (1995). "Bayesian model choice via Markov chain Monte Carlo methods". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 473–484.

Chib, Siddhartha and Ivan Jeliazkov (2001). "Marginal likelihood from the Metropolis–Hastings output". In: *Journal of the American Statistical Association* 96.453, pp. 270–281.

Clayton, David and Luisa Bernardinelli (1992). "Bayesian methods for mapping disease risk". In: *Geographical and environmental epidemiology: methods for small-area studies*, pp. 205–220.

D'Alo, M et al. (2012). "Use of spatial information in small area models for unemployment rate estimation at sub-provincial areas in Italy". In: *Journal of the Indian Society of Agricultural Statistics* 66.1, pp. 43–53.

Datta, Gauri S et al. (1999). "Hierarchical Bayes estimation of unemployment rates for the states of the US". In: *Journal of the American Statistical Association* 94.448, pp. 1074–1082.

Devine, OJ (1992). "Empirical Bayes and Constrained Empirical Bayes Methods for Estimating Incidence Rates in Spatially Aligned Areas, unpublished Ph. D". PhD thesis. dissertation, Division of Biostatistics, Emory University.

Eurostat (2015). *Labour force survey in the EU, candidate and EFTA countries. 201' ed.* Eurostat Statistical Working Papers.

Fabrizi, Enrico, Giorgio E. Montanari, and M. Giovanna Ranalli (2015). "A hierarchical latent class model for predicting disability small area counts from survey data". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Fabrizi, Enrico et al. (2011). "Hierarchical Bayes multivariate estimation of poverty rates based on increasing thresholds for small domains". In: *Computational Statistics & Data Analysis* 55.4, pp. 1736–1747.

Fay, Robert E and Roger A Herriot (1979). "Estimates of income for small places: an application of James-Stein procedures to census data". In: *Journal of the American Statistical Association* 74.366a, pp. 269–277.

Germain, Sarah Elizabeth (2010). "Bayesian spatio-temporal modelling of rainfall through non-homogenous hidden Markov models". PhD thesis. University of Newcastle Upon Tyne.

Ghosh, Malay, Narinder Nangia, and Dal Ho Kim (1996). "Estimation of median income of four-person families: a Bayesian time series approach". In: *Journal of the American Statistical Association* 91.436, pp. 1423–1431.

Goodman, Leo A (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models". In: *Biometrika* 61.2, pp. 215–231.

Green, Peter J and Sylvia Richardson (2002). "Hidden Markov models and disease mapping". In: *Journal of the American statistical association* 97.460, pp. 1055–1070.

Hubert, Lawrence J (1973). "The use of orthogonal polynomials for trend analysis". In: *American Educational Research Journal*, pp. 241–244.

Istat (2006). "La rilevazione sulle forze di lavoro – Contenuti, metodologie, organizzazione." In: *Metodi e norme, n.32*.

– (2014). "I Sistemi Locali del Lavoro 2011". In: *Rapporto Annuale 2014*.

ISTAT (2015). "Rapporto Annuale 2015: la situazione del paese." In:

Jasra, Ajay, CC Holmes, and DA Stephens (2005). "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling". In: *Statistical Science*, pp. 50–67.

Knorr-Held, Leonhard, Günter Raßer, and Nikolaus Becker (2002). "Disease Mapping of Stage-Specific Cancer Incidence Data". In: *Biometrics* 58.3, pp. 492–501.

Knorr-Held, Leonhard and Sylvia Richardson (2003). "A hierarchical model for space–time surveillance data on meningococcal disease incidence". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52.2, pp. 169–183.

Lawson, AB et al. (2000). "Disease mapping models: an empirical evaluation. Disease Mapping Collaborative Group". In: *Statistics in medicine* 19.17-18, pp. 2217–41.

Lawson, Andrew B and Hae-Ryoung Song (2010). "Bayesian hierarchical modeling of the dynamics of spatio-temporal influenza season outbreaks". In: *Spatial and spatio-temporal epidemiology* 1.2, pp. 187–195.

Lazarsfeld, Paul F (1950). "The logical and mathematical foundation of latent structure analysis". In: *Measurement and prediction* 4.

Lazarsfeld, Paul Felix, Neil W Henry, and Theodore Wilbur Anderson (1968). *Latent structure analysis*. Houghton Mifflin Boston.

Liu, Jun S, Wing Hung Wong, and Augustine Kong (1994). "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes". In: *Biometrika* 81.1, pp. 27–40.

MacDonald, Iain L and Walter Zucchini (1997). *Hidden Markov and other models for discrete-valued time series*. Vol. 110. CRC Press.

Marhuenda, Yolanda, Isabel Molina, and Domingo Morales (2013). "Small area estimation with spatio-temporal Fay–Herriot models". In: *Computational Statistics & Data Analysis* 58, pp. 308–325.

Marin, Jean-Michel, Kerrie Mengersen, and Christian P Robert (2005). "Bayesian modelling and inference on mixtures of distributions". In: *Handbook of statistics* 25.16, pp. 459–507.

Mollié, Annie (1999). "Bayesian and empirical Bayes approaches to disease mapping". In: *Disease mapping and risk assessment for public health*, pp. 15–29.

Rao, JNK and Mingyu Yu (1994). "Small-area estimation by combining time-series and cross-sectional data". In: *Canadian Journal of Statistics* 22.4, pp. 511–528.

Rao, John NK (2003). *Small area estimation*. Wiley Online Library.

Richardson, Sylvia et al. (1995). "Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain". In: *Statistics in medicine* 14.21-22, pp. 2487–2501.

Robertson, Colin et al. (2010). "Review of methods for space–time disease surveillance". In: *Spatial and Spatio-temporal Epidemiology* 1.2, pp. 105–116.

Spezia, Luigi (2010). "Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes". In: *Journal of Time Series Analysis* 31.1, pp. 1–11.

Srivastava, AK, UC Sud, and H Chandra (2007). "Small area estimation-An application to national sample survey data". In: *Journal of the Indian Society of Agricultural Statistics* 61.2, pp. 249–254.

Tanner, Martin A and Wing Hung Wong (1987). "The calculation of posterior distributions by data augmentation". In: *Journal of the American statistical Association* 82.398, pp. 528–540.

Van Dyk, David A and Xiao-Li Meng (2001). "The art of data augmentation". In: *Journal of Computational and Graphical Statistics* 10.1.

Vermunt, Jeroen K and Jay Magidson (2002). "Latent class cluster analysis". In: *Applied latent class analysis* 11, pp. 89–106.

Wiggins, Lee M (1973). "Panel analysis: Latent probability models for attitude and behavior processes." In:

You, Yong and JNK Rao (2002). "Small area estimation using unmatched sampling and linking models". In: *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pp. 3–15.

You, Yong, JNK Rao, and Jack Gambino (2003). "Model-based unemployment rate estimation for the Canadian Labour Force Survey: a hierarchical Bayes approach". In: *Survey Methodology* 29.1, pp. 25–32.