UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Spatio-temporal analysis of hospitalisations due to cerebro-cardiovascular diseases in Portugal - a disease mapping approach

Jéssica Filipa Vicente Marques

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Prof.ª Doutora Marília Antunes
Prof.ª Doutora Soraia Pereira

2022

"It always seems impossible until it´s done."

- Nelson Mandela

"Courage doesn't always roar...Sometimes it's the quiet voice at the end of the day whispering, I will try again tomorrow."

- Mary Anne Radmacher

Resumo

As doenças cérebro-cardiovasculares são todas as doenças que afetam o sistema circulatório, ou seja, o coração e os vasos sanguíneos (artérias, veias e vasos capilares). Segundo o Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA), as doenças cérebro-cardiovasculares são a maior causa de morte em Portugal. Ao longo dos últimos anos, foram desenvolvidos diversos projetos de investigação que visam a promoção da saúde e consciencialização da gravidade deste tipo de problemas. Para além disto, para que o diagnóstico seja mais precoce, existe uma oferta de análises, com o objetivo de diagnosticar este tipo de doenças ou possíveis fatores de risco. Em 2019, de acordo com relatórios do Instituto Nacional de Estatística (INE), as doenças cérebro-cardiovasculares representaram 29.9% dos óbitos registados, sendo que 9.8% do total de mortes foi devido a AVC e 6.4% devido a doenças isquémicas do coração e 3.8% devido a enfartes agudos do miocárdio. Estes valores são bastante preocupantes e foram o motor de arranque para a realização deste projeto.

Um dos objetivos deste projeto passa por caracterizar a distribuição espacial e temporal das admissões hospitalares devido às doenças cérebro-cardiovasculares e os seus fatores de risco. Para além disso, este projeto visa a deteção e avaliação de um padrão espacial e temporal deste tipo de doenças em Portugal Continental e identificar os regressores que melhor explicam esta variação espacial e temporal. Para esta finalidade, o principal interesse é analisar o consumo de recursos hospitalares sob a forma de taxa de admissão hospitalar. Neste estudo, foram considerados os dados agregados do número de admissões hospitalares por doenças cérebro-cardiovasculares, por área demográfica (distrito e município) e unidade temporal (ano). Em paralelo com estes dados, foram analisados dados agregados (à mesma escala territorial e temporal) relativos a características sociodemográficas, acesso a serviços de saúde e fatores de risco, que foram incluídos como covariáveis nas diversas análises. A nível do distrito considerou-se a proporção de residentes do sexo masculino, a proporção de residentes com 65 ou mais anos, a taxa de desemprego, a densidade populacional, a proporção de inscritos em federações e a taxa de mortalidade por 1000 habitantes devido à diabetes. A nível do município, foram consideradas todas as variáveis anteriormente mencionadas, exceto a proporção de inscritos em federações. No entanto, foi incluído o número de hospitais em cada município por 1000 habitantes.

Para detetar e avaliar padrões a nível espacial e temporal, recorreu-se ao mapeamento de doenças através de modelos bayesianos hierárquicos. Este tipo de modelos tem estado em constante crescimento, e reconhecimento por parte de estatísticos, por ser uma ferramenta bastante útil para análises espaciais e espaço-temporais. Como a variável de interesse é o número de admissões hospitalares, a distribuição subjacente aos dados escolhida foi a Poisson. Toda a inferência estatística foi feita recorrendo ao *package* INLA no *software* R.

Desde 2010 até 2016, o distrito mais afetado por uma elevada taxa de admissões hospitalares foi Castelo Branco. Em 2018, o distrito que se destacou foi Bragança, que é também o segundo distrito com uma taxa de admissão mais elevada em 2012 e 2014 e o terceiro em 2010 e 2016. Para além disso, com base neste estudo, concluiu-se que, a proporção de residentes do sexo masculino e a proporção de residentes com 65 ou mais anos são significativas para explicar a variação na taxa de admissões hospitalares nos distritos ao longo do tempo. Uma vez que o coeficiente, de ambas as variáveis, é negativo, é possível concluir que regiões com uma elevada proporção de homens e com uma elevada proporção de idosos tendem a ter uma baixa taxa de admissões hospitalares. Esta conclusão leva-nos a crer que as mulheres poderão ir com mais frequência ao hospital, e como tal, representam uma maior taxa de admissões hospitalares. Posto isto, suspeita-se que o género do paciente possa ser um *proxy* para um determinado comportamento face à saúde. Relativamente à proporção de residentes com mais de 65 anos, este resultado não foi o esperado, no entanto a interpretação desta conclusão deverá ser feita com cuidado, tendo em conta que o número de observações no modelo, ao nível do distrito, é pequeno. Para além disso, apesar da literatura afirmar que ser homem e ser idoso são fatores de risco para estas doenças, os resultados deste estudo não são diretamente comparáveis às conclusões dos estudos da literatura, uma vez que ser fator de risco para a propensão das doenças cérebro-cardiovasculares é diferente de ser fator de risco para a admissão hospitalar pelo mesmo tipo de doença.

Em relação aos municípios, é no distrito de Castelo Branco que se situam os municípios com maior taxa de admissão hospitalar desde 2010 até 2016. Em 2018, passam a ser os municípios localizados na região Médio Tejo que apresentam as taxas de admissão hospitalar mais elevadas. Com base neste estudo, concluiu-se que, apenas as variáveis taxa de desemprego e proporção de residentes com mais de 65 anos são significativas para explicar a variação na taxa de admissões hospitalares nos municípios ao longo do tempo. Deste modo, como os coeficientes são positivos, regiões com uma maior taxa de desemprego e uma maior proporção de residentes com mais de 65 anos tendem a ter uma maior taxa de admissão por doenças cérebro-cardiovasculares. O facto de um indivíduo estar desempregado pode gerar situações de *stress*, e tendo em conta a literatura, o *stress* é um fator de risco para este tipo de doença. Por outro lado, de acordo com a literatura, a idade avançada também é um fator de risco para estas doenças. Estas conclusões levam-nos a crer que estas variáveis são consideradas fatores de risco tanto para a taxa de admissão, como para a propensão de doenças cérebro-cardiovasculares, no entanto estas conclusões não são diretamente comparáveis, tal como foi explicado no caso dos distritos.

No entanto, quando os efeitos aleatórios (espacial, temporal e interação espaço-tempo) não são considerados, todas as variáveis, exceto a taxa de desemprego, tornam-se significativas para a taxa de admissão hospital a nível do distrito. Ao nível do município, todas as variáveis são significativas. De acordo com estes resultados, e com os resultados apresentados nos últimos dois parágrafos, concluiu-se que não são as próprias variáveis (variáveis não significativas no modelo com efeitos aleatórios e significativas no modelo sem efeitos aleatórios) que afetam diretamente a evolução da taxa de admissões hospitalares devido às doenças cérebro-cardiovasculares, mas sim a localização das áreas, caracterizadas por estes valores, ao longo do tempo. Um exemplo elucidativo deste comportamento, por parte das variáveis, pode ser dado analisando a taxa de mortalidade devido à diabetes. Uma vez que esta variável só é significativa quando os efeitos aleatórios não são considerados, conclui-se que não é o facto de ter diabetes num estado mais avançado (que poderá levar à morte) que se torna um fator de risco para a taxa de admissão hospitalar por doenças cérebro-cardiovasculares, mas sim o facto deste tipo de diabéticos viverem em áreas de risco para a taxa de admissão hospitalar por este tipo de doenças. Esta situação é

retratada quando se considera o facto do acesso aos cuidados de saúde ser mais limitado nas áreas rurais, e consequentemente, quando os doentes, que vivem nestas regiões, se deslocam ao hospital, apresentam um estado de saúde mais degradado, o que pode levar à morte. Seguindo esta linha de pensamento, não se está a descartar a associação entre a diabetes e a propensão para as doenças do aparelho circulatório, mas sim a ponderar que não é a gravidade da diabetes (aqui representada pela taxa de mortalidade) que está associada à taxa de admissões hospitalares por doenças cérebro-cardiovasculares, mas sim a zona de residência dos doentes.

Palavras-chave: Doenças cérebro-cardiovasculares; Mapeamento de doenças; Modelos espaço-temporais; Modelos bayesianos hirárquicos; INLA.

Abstract

Cerebro-cardiovascular diseases (CCD) are all diseases that affect the circulatory system, that is, the heart and blood vessels (arteries, veins and capillary vessels). According to the Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA), CCD are the biggest cause of death in Portugal. That said, one of the objectives of this project is to characterise the spatial and temporal distribution of hospital admissions due to CCD and their risk factors. In addition, this project aims to detect and assess a spatial and temporal pattern of this type of disease in mainland Portugal and identify the regressors that best explain this spatial and temporal variation. Thus, the hospital admission rate and their risk factors were analysed at spatial (district and municipality) and temporal (year) levels by disease mapping using Bayesian hierarchical models.

Based on this study, it can be concluded that regions with a high proportion of men and a high proportion of older residents tend to have a low hospital admission rate. This conclusion leads us to suspected that the patient's gender may be a proxy for a particular health-related behaviour. Regarding the proportion of residents aged 65 years or over, this result was not as expected, since, considering the literature, being older is a risk factor for circulatory system diseases. However, the results of this study should be interpreted with caution because the number of observations in the model at district level is small.

In addition, it was concluded that municipalities with a high unemployment rate and high proportion of residents aged 65 or over tend to have a high hospital admission rate for CCD. Considering the literature, stress and advanced age are risk factors for this type of disease. Furthermore, the findings of this study and the literature cannot be directly compared, since being a risk factor for the propensity for CCD is different from being a risk factor for hospital admission.

Regarding the remaining variables under study (non-significant variables in the model with random effects and significant in the model without random effects), it was concluded that it is not the variables that directly affect the evolution of the hospital admission rate due to CCD, but rather the location of the areas characterised by the values of these variables over time.

Keywords: Cerebro-cardiovascular diseases; Disease mapping; Spatio-temporal models; Bayesian hierarchical models; INLA.

Acknowledgements

I dedicate this project to my parents, for their unconditional support throughout my academic career and throughout my life. I take this opportunity to thank them for their commitment, and all the efforts they made to allow me to successfully complete my university trajectory and, consequently, go further. I also want to mention that my parents transmit to me love, protection, strength and the sense of togetherness. I am very grateful to have them by my side.

To my mother, Clara Vicente, I thank her for the values she instilled in me, for accompanying me every day of this journey and for her constant concern and availability in everything I needed. Despite the adversities that life put in our way, my mother never let me give up my dreams.

To my father, Pedro Marques, I thank him for everything he taught me about life, for being the person who most believes in my abilities, for all his love and for never quitting to give me the best education, despite the setbacks after the divorce with my mother. This is the father who never let me give up on my ambitions, who taught me that we should never say "never" because we are capable of anything - we just have to be persistent and run after our dreams. To my maternal great-grandparents (in memoriam), Alberto and Margarida Vicente, for having raised me in the best way and for having passed on to me the best ideals. Today and always, an unconditional love. I missed them so much, every minute of this journey, and every minute of my days.

A special thanks to my thesis supervisors, Professor Marília Antunes and Professor Soraia Pereira, always accessible, dedicated and patient towards me. Their constructive criticism and advice simplified the inconveniences that arose throughout this time.

A word of thanks to my professors at the Department of Statistics and Operational Research of the Faculty of Sciences, who contributed to my academic career by transmitting me knowledge and adding value to my work.

To my friend Vanessa Lopes, I thank her for our friendship, for the affection, for all the hours of laughter and despair, for what we have already lived and for what we will still live together. She allowed my academic education not to be a lonely journey and I owe her every minute of happiness, crying and companionship. Indeed, there are moments that deserve to be repeated. To my friend Diana Montrond, I thank her for all the help, the laughs and for telling me "If you can't do it, no one can" when I had several university assignments at the same time.

A huge thank you to my best friends who, despite the distance, managed to accompany me and contribute to my academic and personal growth.

Finally, my thanks also go to the Administração Central do Sistema de Saúde, I.P (ACSS) for providing data.

February of 2022,
Jéssica Marques

Contents

List of Figures

List of Tables

List of acronyms and abbreviations

ACSS	Administração Central do Sistema de Saúde
CCD	Cerebro-cardiovascular diseases
CPO	Conditional predictive ordinate
DGS	Direção Geral da Saúde
DIC	Deviance information criterion
FCS	Fully conditional specification
GLMM	Generalized linear mixed models
GMRF	Gaussian Markov Random Field
ICD-10-CM	International Classification of Diseases, Tenth Revision
ICD-9-CM	International Classification of Diseases, Ninth Revision
INE	Instituto Nacional de Estatística
INLA	Integrated Nested Laplace Approximation
JM	Joint modeling
LGM	Latent Gaussian models
MC	Markov chain
MCMC	Markov chain Monte Carlo
MI	Multiple imputation
MICE	Multivariate Imputation by Chained Equations
PIT	Probability integral transform
RW	Random walk

1. Introduction

Cerebro-cardiovascular diseases (CCD) are all those that affect the circulatory system, that is, the heart and blood vessels (arteries, veins and capillary vessels). According to the Instituto Nacional de Saúde Doutor Ricardo Jorge (Bourbon et al. (n.d.)), diseases of the circulatory system "may be of various types, the most concerning being the coronary artery disease (heart arteries) and the brain artery disease".

According to Ferreira et al. (2016), the cerebro-cardiovascular diseases are the main cause of death in Portugal. In 2019, the Instituto Nacional de Estatística (INE) advance that the CCD represented 29.9% of the total number of deaths, being that 9.8% of the total deaths in the country were due to strokes, 6.4% due to isquemic heart diseases and 3.8% due to acute myocardial infarction. These values justify the prioritisation of CCD in health planning and motivate the analysis of the impact of these diseases on Portuguese citizens. Furthermore, the best way to prevent these types of diseases is to control or prevent the associated risk factors, which may be modifiable or non-modifiable (Bourbon et al., n.d., Gulbenkian Descobrir and Maratona da Saúde, 2016 and Médis, 2018):

Modifiable risk factors:

- Diabetes
- High cholesterol
- Overweight
- Smoking
- Alcoholism
- Stress
- Sedentarism

Non-modifiable risk factors:

- Age (according to the Ferreira et al. (2016) there is a higher incidence in the 65 or over age group)
- Gender (according to Bourbon et al. (n.d.) and Ferreira et al. (2016), in Portugal, the CCD affecting mostly men)
- Genetics

The initial database provided by the Administração Central do Sistema de Saúde, I.P (ACSS) for this project had all the hospital admissions (all diseases) for the regions of mainland Portugal for the

years 2010-2019. The database to be worked on will include only the main coded diagnoses belonging to the group of CCD according to the 9th revision (ICD-9-CM) for the years between 2010 and 2016 or according to the 10th revision (ICD-10-CM) for the years from 2017 onwards (Pires (2018)). The aim of this study is to analyse how the number of hospital admissions is distributed in space and time (spatio-temporal analysis) and to identify which regressors best explain the spatial and temporal variation. The spatial analysis will take into account districts and municipalities (see the Figure 1.1 below and the Figure 17 in Attachments, respectively). Thus, the database at district and municipality level is composed of variables specifying the patient's area, the number of hospital admissions due to CCD at area level and, in addition, variables provided by INE, which were considered risk factors for CCD hospital admissions, were also introduced.

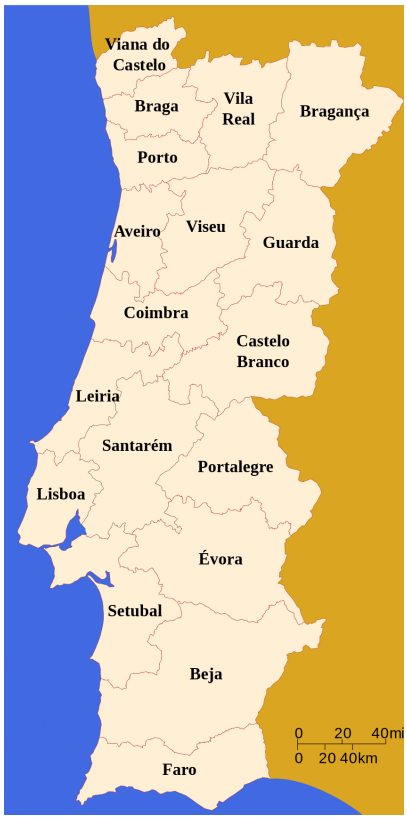

Figure 1.1: Districts of mainland Portugal.

In this type of studies, the regions are not independent, so it is normal that closer regions are more similar to each other compared to more distant regions and the same is true for years. That said, it is necessary to consider a possible spatial and/or temporal correlation. As the data are aggregated and the number of hospital admissions due to CCD (response variable) belongs to a Poisson distribution (a distribution that belongs to the exponential family), the Generalised Linear Mixed Models (GLMM) would be a good approach (Cadima (2015), M. Antônia Amaral Turkman and Silva (2000), Dobson (2002), Kleinschmidt et al. (2001), Oliveira (2018) and Kleinman et al. (2004)). This type of models allows the incorporation of fixed effects associated with socio-demographic and clinical variables and random effects that explain spatial and/or temporal influences, characterised by districts/municipalities and years.

For this purpose, both classical and Bayesian approaches can be used, however the latter is more flexible and has been more explored in recent times (Bermudez (2021), Lawson (2008) and Juárez (2018)). Therefore, the Bayesian approach under Integrated Nested Laplace Approximation (INLA) will be used to analyse the data of this project. According to Lindgren and Rue (2015), this methodology "is designed for latent Gaussian models, a very wide and flexible class of models ranging from Generalized Linear Mixed to spatial and spatio-temporal models".

Chapter 2 will describe all the methodologies studied and applied to the data under study. Chapter 3 describes the cleaning of the initial data provided by ACSS, the introduction of the potential risk factors of INE and finally the set of variables that will be analysed in this project. Furthermore, Chapter 3 will show the explanatory analysis of the data. Chapter 4 will show the results of applying the methodologies described in Chapter 2. Finally, Chapter 5 will present the discussion and the main conclusions.

2. Methods

2.1 Missing data

Missing values are very common in epidemiological studies. Rubin (1976) classified the pattern of missing values into three categories:

- **Missing completely at random (MCAR)** - the probability of a value being missing is the same in all cases, that is, the cause of omission of information is not associated with the data.
- **Missing at random (MAR)** - the probability of missing information depends on the values of the observed variables.
- **Not missing at random (NMAR)** - the values are not omitted at random.

In the case of this study, the pattern of missing values is not NMAR since there is no specific reason why the missing values correspond to the respective municipalities. Furthermore, the pattern of missing values should not be MCAR, since there is a higher probability that data were not collected in less developed and less populated municipalities. Thus, the pattern of missing values is assumed to be MAR.

According to Katitas (2019), there are three ways to address these missing values:

- **Remove rows containing missing data:** If the amount of missing data is very small, this may be the best option to ensure that the analysis is not biased. However, deleting data means that the user may not have access to important information, especially if the sample size is small.
- **Replace the missing values with the median or mean of the data:** As the missing values are substituted by a constant (mean or median), this option may cause bias since it decreases the variance.
- **Multiple imputation:** The observed data/variables distribution is used to estimate plausible values, which will replace the missing values.

Multiple imputation (MI) was the method chosen to impute missing values from the raw data, as this method has become a very popular tool in replacing missing values in recent years.

2.1.1 Multiple Imputation

Multiple imputation is a method that works on databases with incomplete data, where missing values may occur in one or more variables. There are two approaches for imputing multivariate data (Stuart et al. (2009) and Buuren (2007)):

- **Joint modeling (JM)** - assumes that incomplete variables belong to a specific multivariate distribution (usually a multivariate normal distribution) and consequently imputations are extracted from their conditional distributions by Markov chain Monte Carlo (MCMC). If the imputed values do not belong to the same distribution as the non-missing values, they are incorrect and this is the main problem with this approach.
- **Fully conditional specification (FCS)**, also known as **Multivariate Imputation by Chained Equations (MICE)** - impute the missing values using univariate conditional distributions to each incomplete variable, given all the others. From an initial imputation, this approach extracts imputations iteratively over the conditional densities.

The MICE was used to impute missing values. According to Buuren and Groothuis-Oudshoorn (2011), in the multiple imputation there are three steps:

1. **Imputation** - The imputation process starts with an incomplete dataset, that is, a dataset with missing values. In this step several imputed versions of the data (m) are created, where the missing values are replaced by plausible imputed values. The missing values belong to a specific distribution and the imputed values are drawn according to this distribution. Thus, these datasets are copies of the original dataset, but in place of the initial missing values are now the imputed values. In the example of Figure 2.1, three datasets ($m = 3$) were created, each with the missing values replaced by plausible imputed values. Once the imputed values are generated, there is uncertainty on the actual value of those missing values.
2. **Analysis** - In this step the analysis of each set of imputed data is performed, and this analysis must be equal to the analysis that would be performed if the data were complete. Thus, m different analyses and m different coefficients of determination result from this step, since each data set has different imputed values.
3. **Pooling** - In this final step, the m results found in previous step are pooled to obtain the final analysis of data.

All this steps are represented in the figure below, based on Buuren and Groothuis-Oudshoorn (2011).

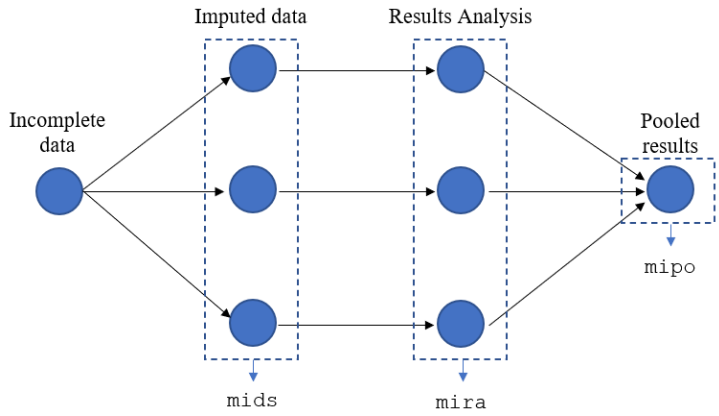

Figure 2.1: Steps of multiple imputation.

More details on the multiple imputation procedure are in Buuren and Groothuis-Oudshoorn (2011) and Buuren (2007).

2.1.1.1 Tools for Multiple imputation

In software R, the package used for multiple imputation is `mice`. For the first step, the function used to generate imputed values from a distribution specifically modelled for each missing entry is `mice()`. This function has several parameters:

1. `data` is the dataset with missing values;
2. `m` is the number of complete datasets created;
3. `maxit` is the number of iterations. According to Buuren and Groothuis-Oudshoorn (2011) to obtain the good values, in general, it is only necessary to select between 10 and 20 iterations;
4. `method` is the method used to generate values;
5. `predictorMatrix` is the predictor matrix. The rows of the predictor matrix represent the incomplete variables and the columns represent all the variables in the original dataset. This matrix is filled with 0 and 1, and as mentioned in Buuren and Groothuis-Oudshoorn (2011), "the value 1 indicates that a column variable is used as a predictor to impute the row variable, and a 0 means that it is not used. The diagonal is 0, since a variable cannot predict itself."

Firstly, the `mice()` function is used only with argument `data` and then is extracted the `predictorMatrix` of this function. After the extraction of the predictor matrix, the `mice()` function is used again but now the parameters `m`, `maxit`, `method` and `predictorMatrix` are included in order to create imputations. The imputed values are stored in an object of class `mids` - *multiply imputed dataset*. An important step in MI is to assess whether the imputed values are plausible. The imputations should be values that could have been obtained if they not were missing. To ensure that the final dataset is plausible, the imputed values that are clearly impossible (for example negative counts for this specified study) should not be included in the domain of the plausible values which could be used in MI. The command `scripplot(imp)` is useful to visualize the observed and imputed values and to conclude if is necessary to filter the domain of imputations. Finally, through function `complete()` it is possible to see the datasets with imputed and observed values, that is, the complete datasets.

In the second step, the analysis of `m` datasets is made through the command `with()`. The coefficient of determination R^2 is calculated through function `pool.r.squared()`. The results of this step are stored in class `mira` - *multiply imputed repeated analysis*.

In the third step, the `m` regression models are aggregated in the only model through command `pool()`. The argument of function `pool()` is the object created in the last step with function `with()`. The results are stored as a *multiple imputed pooled outcomes* object - `mipo`.

For more details on the `mice` package, see Buuren and Groothuis-Oudshoorn (2011) and Buuren (2007).

2.2 Bayesian inference

Statistical inference is the science that allows users to draw conclusions about a population from a sample taken from that same population. According to M. Antónia Amaral Turkman and Paulino (2015), statistical inference can be performed using two main approaches: classical approach and Bayesian approach.

In the classical approach, the repeated sampling principle is used, that is, it is necessary to repeat an event A several times to obtain the probability of its occurrence. Furthermore, only the information obtained through observation of the sample data is used, that is, all external information is ignored. Summarily, the data $\boldsymbol{y}$ are observed, the model is chosen - $p(\boldsymbol{y}|\theta)$ - and, finally, the parameter θ is often estimated by maximum likelihood estimation. Thus, the model parameters are unknown and fixed, and the data are random (Bermudez, 2021). In the Bayesian approach, the probability of occurrence of an event A corresponds to measure of degree of credibility that the user has in this event. In this approach all information is considered necessary, so information from the sample can be combined with information already available or with information obtained through expert opinion or past experience (Bermudez, 2021). Thus, in Bayesian inference there are three important concepts: *likelihood, prior* and *posterior*. Statisticians' beliefs or past experiences about the parameter distribution are called *prior* information, but first, to understand this concept it is necessary to know Bayes' theorem in its simple form. Let A and B be two events with $P(B) \neq 0$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

2.2.1 *Likelihood* function and *prior* and *posterior* distributions

Considering a random variable Y (discrete or continuous), the distribution of data under the parameter θ (mass function or the distribution function, respectively) is:

$$p(\boldsymbol{y}|\theta) \tag{2.2}$$

and this is called *likelihood function*, which here is denoted in terms of the sampling distribution. As θ is the unknown parameter, its *prior* distribution has to be specified with the aim to obtain the *posterior* distribution. This parameter is specified before the knowledge of the data, so the *prior* distribution is represented by $p(\theta)$. Let Θ the support of distribution, then:

$$\Theta = \{\theta : p(\theta) > 0\}. \tag{2.3}$$

Note that $p(\theta)$ reflects the *prior* belief. In turn, the parameters that allow explaining the parameter/vector of parameters of interest are called hyperparameters. This dependence structure between data and parameters and, consequently, between parameters and hyperparameters is called hierarchical structure. Thus, in the presence of hierarchical structure, e.g., spatial and/or temporal dependence in the data, the knowledge about θ is expressed through $\boldsymbol{\psi}$ (vector of hyperparameters) and, consequently, the *prior* distribution of the hyperparameters will be $p(\boldsymbol{\psi})$:

$$p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}|\boldsymbol{\psi})p(\boldsymbol{\psi})d\boldsymbol{\psi}. \tag{2.4}$$

Finally, in order to update the *prior* information with the information from the data, it is necessary to resort to Bayes' theorem:

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{p(\boldsymbol{y})} \tag{2.5}$$

and this is the *posterior* distribution of θ (combines the *prior* information about the parameter vector $\boldsymbol{\theta}$ contained in the distribution $p(\theta)$ with the information about the data $\boldsymbol{y}$ contained in the likelihood function $p(y|\theta)$). As the denominator is the marginal distribution of the data, it will be a constant, and as

it does not depend on θ, it is possible to remove it from the equation and obtain an equation proportional to 2.5 (Besag et al. (1991) and Blangiardo and Cameletti (2015)):

$$p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)p(\theta). \tag{2.6}$$

However, the formulation of $p(\boldsymbol{y})$ depends on the nature of θ, that is, if θ assumes discrete values in Θ:

$$p(\boldsymbol{y}) = \sum_{\theta \in \Theta} p(\boldsymbol{y}|\theta)p(\theta). \tag{2.7}$$

On the other hand, if θ is a continuous variable:

$$p(\boldsymbol{y}) = \int_{\theta \in \Theta} p(\boldsymbol{y}|\theta)p(\theta)d\theta. \tag{2.8}$$

When we work with complex models, it is not possible to obtain the analytical expression of the *posterior* distribution. However, according to Blangiardo and Cameletti (2015), it is possible to obtain some information from the *posterior* distribution (empirical measures calculated on the modelled values of the *posterior* distributions), such as:

- ***Posterior* mean**

  For continuous θ :

  $$E(\theta|\boldsymbol{y}) = \int_{\theta \in \Theta} \theta p(\theta|\boldsymbol{y})d\theta \tag{2.9}$$

  For discrete θ :

  $$E(\theta|\boldsymbol{y}) = \sum_{\theta \in \Theta} \theta p(\theta|\boldsymbol{y}) \tag{2.10}$$

- **Posterior median**

  $$p(\theta \leq \theta_{0.5}|\boldsymbol{y}) = 0.5 \text{ and } p(\theta \geq \theta_{0.5}|\boldsymbol{y}) = 0.5 \tag{2.11}$$

- $(1-\alpha)\times 100\%$ **credibility interval (CI)**

  $$p(\theta \leq \theta_{\alpha/2}|\boldsymbol{y}) = \alpha/2 \text{ and } p(\theta \geq \theta_{1-\alpha/2}|\boldsymbol{y}) = \alpha/2 \tag{2.12}$$

Choice of *prior* distribution

When the researcher chooses the *prior* distribution, the nature of the parameters and of the hyper-parameters, which allows estimating θ, must be taken into account. About the nature of parameters, their support should be taken into account, that is, if the parameter represents a proportion (for instance, probability of death by stroke) the distribution under this parameter should have the support between 0 and 1. In the other hand, if the parameter represents the average body mass index (BMI) in the Portuguese citizens, the support of *prior* distribution should be between 0 and $+\infty$.

Given the *likelihood* $p(\boldsymbol{y}|\theta)$, if the *posterior* $p(\theta|\boldsymbol{y})$ (Equation 2.5) belongs to the same family as the *prior*, then $p(\theta)$ is a conjugate *prior*. Thus, it is said that the *prior* is conjugated to the *likelihood*. This property is very useful when the functional form of the *posterior* distribution is known and its hyperparameters as well. However, in many cases it will not be possible to find a conjugate *prior* and in these cases it is necessary to resort to a computational method, as will be seen in Section 2.2.2.

Informative or noninformative *prior*

The *prior* distribution parameters should be defined according to the type of existing knowledge about them: informative or noninformative *prior*. For example, when the researcher has information derived from previous experiments, on the problem at hand, this is called the informative *prior* and this information should be included in the model.

When parameter information is missing, a non-informative *prior* is often used to let the data speak for themselves. One procedure that allows the construction of noninformative *priors* was proposed by Jeffrey (1946). As reported by Bermudez (2021), the Jeffrey's *prior* are invariant under injective transformations. So, in the case where θ is a scalar ($k = 1$), the Jeffrey's *prior* is:

$$p(\theta) \propto I(\theta)^{1/2}, \tag{2.13}$$

where

$$I(\theta) = E_{X|\theta}\left[\left(\frac{d \log p(y|\theta)}{d\theta}\right)^2\right]. \tag{2.14}$$

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, so the *prior* distributions of Jeffrey is:

$$p(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}, \tag{2.15}$$

where $|I(\boldsymbol{\theta})|$ is the determinant of the expected value of the Fisher information matrix, whose element (i, j) is given by:

$$I_{ij}(\boldsymbol{\theta}) = -E_{X|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(y|\boldsymbol{\theta})\right]. \tag{2.16}$$

If the integral of this *prior* is different from 1 (it is not a probability distribution), the *prior* is called improper *prior*. If the *posterior* is also improper, there are complex problems that will lead to an incorrect analysis of the data.

Other way to build a noninformative *prior* distribution is to resort to the so-called *vague prior*. In this case, the *prior* distribution function is flat in the local where the *likelihood* function reaches the maximum. According to Blangiardo and Cameletti (2015), a Normal distribution with mean 0 and large variance, for instance, Normal$(0, 10^6)$, can be used as *prior* for a mean or regression parameter. A similar case is the Beta(1,1) or Gamma(0.01,0.01) for the inverse of variance. The Figure 2.2 is inspired in Blangiardo and Cameletti (2015) and show an example of a noninformative *prior*.

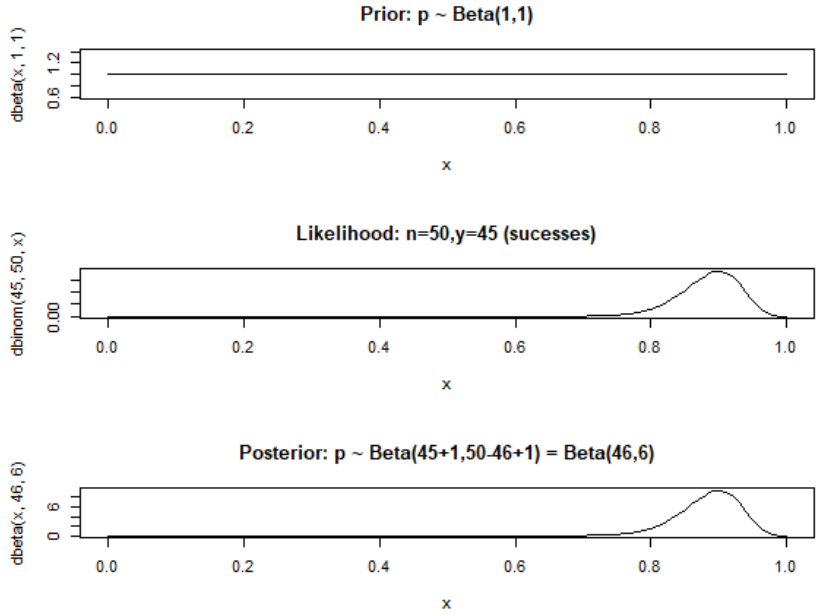

Figure 2.2: Example of noninformative *prior*.

2.2.2 Bayesian computing

In the "Choice of *prior* distribution" part of Section 2.2 conjugate models were presented, however the functional form is frequently unknown (when the conjugacy is not appropriate), not being possible to manipulate the *posterior* distribution analytically. However, there are simulation methods that can be used to overcome this obstacle.

The first two methods of simulation presented generate values from the *posterior* distribution and are called Monte Carlo (MC) and Markov chain Monte Carlo (MCMC). The MC assumes that the *posterior* distribution has a known form, via conjugate models, and generates independent values through this distribution. The MCMC combines the MC approach and the Markov chains, where the *posterior* distribution is not known and generates approximate values, by means of two methods: Gibbs sampler or Metropolis-Hastings algorithm. However, MC and MCMC are algorithms with limitations in terms of computational load, and given that the dataset is currently quite extensive and taking into account the spatial and spatio-temporal dependencies, it becomes too complex for this approach. According to Lindgren and Rue (2015), INLA (Integrated Nested Laplace Approximation) is a computationally efficient method and an alternative to MC and MCMC capable of overcoming computational limitations. As this method provides fast results, it will be used in this analysis.

2.2.2.1 The Integrated Nested Laplace Approximation

As mentioned in the previous paragraph, INLA is a methodology used to overcome the computational flaws present in the other methodologies. This approach allows the user to obtain an approximation of the *posterior* marginal distribution for each element of the vector of the parameters of interest with the implementation of Latent Gaussian Models (LGM).

According to Blangiardo and Cameletti (2015), the class of LGM can be represented by a three-level hierarchical structure:

1. **Data|Parameters:** Identify the distribution of observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ through *likelihood* function:

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{n} p(y_i|\theta_i, \boldsymbol{\psi}), \tag{2.17}$$

being that the distribution is characterized by a parameter γ_i. This parameter is linked to structured additive predictor η_i through a link function $g(.)$, that is, $g(\gamma_i) = \eta_i$, where η_i is defined as:

$$\eta_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ji} + \sum_{k=1}^{K} f_k(z_{ki}), \tag{2.18}$$

where β_0 is the intercept; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ is the vector with coefficients that quantify the linear effects of the covariates $\boldsymbol{x}$; $\boldsymbol{f} = (f_1(.), \ldots, f_K(.))$ are the functions of covariates $\boldsymbol{z} = (z_1, \ldots, z_K)$ that can assume different forms: smooth, nonlinear effects of covariates, time and seasonal trends and temporal or spatial random effects. By this enumeration, it can be stated that the LGM are very flexible and encompass a diverse range of models (including spatial and spatio-temporal models).

2. **Parameters|Hyperparameters**

Thus, the set of parameters composed by the latent components of interest for the inference is defined as $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$. As said in Section 2.2.1, $\boldsymbol{\theta}$ is explained by means the hyperparameters, presented in the vector $\boldsymbol{\psi}$, so:

$$\boldsymbol{\theta}|\boldsymbol{\psi} \sim \text{Normal}(0, \boldsymbol{Q^{-1}}(\boldsymbol{\psi})), \tag{2.19}$$

and the density function is:

$$p(\boldsymbol{\theta}|\boldsymbol{\psi}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{Q}(\boldsymbol{\psi})|^{\frac{1}{2}} \exp\left(-\frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{Q}(\boldsymbol{\psi}) \boldsymbol{\theta} \right). \tag{2.20}$$

$\boldsymbol{Q}(\boldsymbol{\psi})$ is a sparse precision matrix, since the componentes of $\boldsymbol{\theta}$ are supposed to be conditionally independent. Thus, $\boldsymbol{\theta}$ is modelled by a Gaussian Markov Random Field (GMRF).

3. **Hyperparameters**

In turn, the vector of the L hyperparameters is characterized by $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_L)$ and

$$\boldsymbol{\psi} \sim p(\boldsymbol{\psi}). \tag{2.21}$$

Therefore, the joint distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is given by product of *prior* distribution of hyperparameters (2.21), of GMRF density (2.20) and of *likelihood* function (2.17):

$$
\begin{aligned}
p(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y}) &\propto p(\boldsymbol{\psi}) \times p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\psi}) \\
&\propto p(\boldsymbol{\psi}) \times |\boldsymbol{Q}(\boldsymbol{\psi})|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{Q}(\boldsymbol{\psi})\boldsymbol{\theta}\right) \times \prod_{i=1}^{n} p(y_i|\theta_i,\boldsymbol{\psi}) \\
&\propto p(\boldsymbol{\psi}) \times |\boldsymbol{Q}(\boldsymbol{\psi})|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{Q}(\boldsymbol{\psi})\boldsymbol{\theta}\right) \times \prod_{i=1}^{n} \exp(\log(p(y_i|\theta_i,\boldsymbol{\psi}))) \\
&\propto p(\boldsymbol{\psi}) \times |\boldsymbol{Q}(\boldsymbol{\psi})|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{Q}(\boldsymbol{\psi})\boldsymbol{\theta} + \sum_{i=1}^{n} \log(p(y_i|\theta_i,\boldsymbol{\psi}))\right).
\end{aligned}
\tag{2.22}
$$

In turn, the aim of INLA is to obtain analytical approximations for the marginal *posterior* distributions for the parameters:

$$
p(\theta_i|\boldsymbol{y}) = \int p(\theta_i,\boldsymbol{\psi}|\boldsymbol{y})d\boldsymbol{\psi} = \int p(\theta_i|\boldsymbol{\psi},\boldsymbol{y})p(\boldsymbol{\psi}|y)d\boldsymbol{\psi}, \tag{2.23}
$$

and for the hyperparameters:

$$
p(\psi_l|\boldsymbol{y}) = \int p(\boldsymbol{\psi}|\boldsymbol{y})d\psi_{-l}, \tag{2.24}
$$

where ψ_{-l} represents the vector $\boldsymbol{\psi}$ without the element ψ_l.

Note that both the marginals distributions depend on $p(\boldsymbol{\psi}|\boldsymbol{y})$. For this reason it is necessary to specify this component:

$$
p(\boldsymbol{\psi}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y})}{p(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{y})}. \tag{2.25}
$$

Considering the first line of the demonstration of $p(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y})$ in Equation 2.22:

$$
\begin{aligned}
p(\boldsymbol{\psi}|\boldsymbol{y}) &\propto \frac{p(\boldsymbol{\psi}) \times p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\psi})}{p(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{y})} \\
&\approx \left.\frac{p(\boldsymbol{\psi}) \times p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\psi})}{\tilde{p}(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{y})}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\psi})}
\end{aligned}
\tag{2.26}
$$

where $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{y})$ is the Gaussian approximation of $p(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{y})$, given by Laplace method and $\boldsymbol{\theta}^*(\boldsymbol{\psi})$ is the mode for a given $\boldsymbol{\psi}$.

The expression $p(\theta_i|\boldsymbol{\psi},\boldsymbol{y})$ can be approximated in three ways:

- Approximate $p(\theta_i|\boldsymbol{\psi},\boldsymbol{y})$ directly as the marginals from $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{y})$ by using the Normal distribution, where the precision matrix is defined through the Cholesky decomposition. This procedure is very fast, however the approximation is not very good.
- Considering $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i})$ and using the Laplace approximation (more developments of this method in Blangiardo and Cameletti, 2015):

$$
\begin{aligned}
p(\theta_i|\boldsymbol{\psi},\boldsymbol{y}) &= \frac{p((\theta_i,\boldsymbol{\theta}_{-i})|\boldsymbol{\psi},\boldsymbol{y})}{p(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})} \\
&\propto \frac{p(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y})}{p(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})} \\
&\approx \frac{p(\boldsymbol{\theta},\boldsymbol{\psi}|y)}{\tilde{p}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})}\Bigg|_{\boldsymbol{\theta}_{-i}=\boldsymbol{\theta}^*_{-i}(\theta_i,\boldsymbol{\psi})}
\end{aligned}
\tag{2.27}
$$

where $\tilde{p}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})$ is the Laplace approximation of $p(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})$ and $\boldsymbol{\theta}^*_{-i}(\theta_i,\boldsymbol{\psi})$ is its mode. As usually the random variables $\theta_i|\boldsymbol{\theta}_{-i},\boldsymbol{\psi},\boldsymbol{y}$ belong to a Normal distribution, the approximation shown in this item works well, however is very expensive in terms of computation.

- Use the *simplified Laplace approximation* which is based on Taylor's series expansion of Equation 2.27. This approach is reasonable in many cases and uses a short computational time.

Finally, considering the approximation of $p(\boldsymbol{\psi}|\boldsymbol{y})$ (Equation 2.26) and the approximation of $p(\theta_i|\boldsymbol{\psi},\boldsymbol{y})$ by one of the procedures explained above, the *posterior* marginal distribution equation for each element of the parameter vector (Equation 2.23) is:

$$
\tilde{p}(\theta_i|\boldsymbol{y}) \approx \int \tilde{p}(\theta_i|\boldsymbol{\psi},\boldsymbol{y})\tilde{p}(\boldsymbol{\psi}|\boldsymbol{y})d\boldsymbol{\psi}. \tag{2.28}
$$

More details on the development of these approaches are available in Blangiardo and Cameletti (2015). In software R, to implement this approach, the `INLA` package was used.

2.2.3 Bayesian Hierarchical Models

As explained in Section 2.2.1, Bayesian models with a hierarchical structure are called Bayesian hierarchical models, as is the case for models with spatial dependence and/or temporal dependence.

2.2.3.1 Spatial Models

Several areas of science, such as epidemiology and ecology, have increasingly used models that encompass the geographical location of the data under study. Thus, spatial data present a spatial structure between regions and are described as realizations of stochastic processes indexed by space:

$$
Y(s) \equiv \{y(s), s \in \mathcal{D}\}, \tag{2.29}
$$

where $\mathcal{D}$ is a fixed subset of $\mathbb{R}^d$.

According to Banerjee et al. (2004), Cressie (1993), and Gelfand et al. (2010) there are three types of spatial data:

1. *Area data:* $\mathcal{D}$ is a fixed subset well defined by the area unit s through its boundaries.
2. *Point-referenced (or geostatistical) data:* $y(s)$ is the random outcome at a specific location s, where s can vary continuously in the fixed domain $\mathcal{D}$.
3. *Spatial point patterns:* $y(s)$ is equal to 1 if $s \in \mathbb{R}^d$ and 0 otherwise, that is, if the event occurs or not.

This project considers area data, since district and municipality level records of each patient, who is admitted to the hospital due to CCD, will be modelled. Therefore, only area data will be presented in this section, however further explanations on this topic are given in the Blangiardo and Cameletti (2015).

Area data models

When the data under study are spatial data, all areas have neighbours, therefore the spatial dependence between observations is taken into account. According to Zuur et al. (2009), in area data models, this spatial dependence is introduced through random effects based in the neighbourhood structure.

Based on Blangiardo and Cameletti (2015), considering the area i, there are first-order neighbours, that share borders with the area i and there are second-order neighbours that share borders with first-order neighbours of the area i.

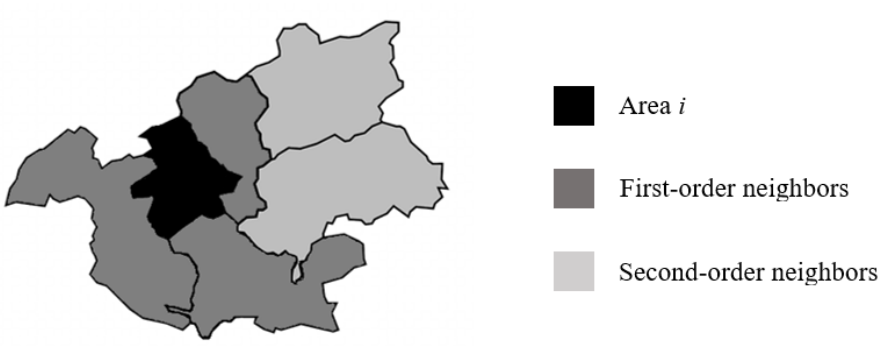

Figure 2.3: Neighboring structure: first-order neighbours and second-order neighbours.

One approach using area data is disease mapping, whose aim is to assess the spatial pattern of a specific disease and identify areas of high or low relative risk (Lawson (2008), Lawson and Williams (2001)). In most cases, the data represent illness counts, admissions to hospitals/other health areas or deaths and therefore the nature of the data is discrete.

Considering the study data, let $Y = (y_1, \ldots, y_n)$ be the number of hospital admissions due to cerebro-cardiovascular diseases in areas of Portugal and E_i the number of exposed in the area i. Thus, the aim of the study is to assess the spatial pattern of hospital admissions. To this end, since the data represent counts, the Poisson model will be used:

$$Y_i \sim \text{Poisson}(\lambda_i),\ i = 1, \ldots, n. \tag{2.30}$$

The interest of this study falls on the incidence rate in each region, since Y represents raw counts, made under different population sizes (each region has a different number of people likely to have the disease under study). Thus, the parameter λ_i can be written as $p_i E_i$, where E_i represents the exposed individuals in the area i and p_i is the incidence rate in the area i. Thus,

$$\lambda_i = p_i E_i \iff p_i = \frac{\lambda_i}{E_i}.$$

Note that, this formulation is typically used when the parameter of interest is relative risk.

The linking of the explanatory variables and the response variable is done through a link function. In this case, as the parameters of $\boldsymbol{\lambda}$ are strictly positive, the link function used is the logarithm (Fox and Weisberg, 2015):

$$\log(\lambda_i) = \eta_i. \tag{2.31}$$

The logarithm of the expected number of cases for each region ($\log(E_i)$) is the offset and represents the exposed population in the region i. Thus, the offset is included in the linear predictor η_i, and the parameters can be interpreted on the logarithm relative risk scale:

$$\log(\lambda_i) = b_0 + \sum_{j=1}^{n_\beta} \beta_j x_{ji} + u_i + v_i + \log(E_i), \tag{2.32}$$

where b_0 is the intercept and represents the average rate across all study regions when the explanatory variables are null; $\beta = (\beta_1, \ldots, \beta_{n_\beta})$ is the vector with the linear effects of the regressors x; v_i is the area-specific effect modelled as exchangeable (the random effect associated with regions without taking into account the spatial dependence, that is, not take into account that what happens in the region i depends on what happens in its neighbourhood); u_i is another area-specific effect modelled as spatially structured (is the random effect associated to the regions, taking into account the spatial dependence, that is, taking into account what happens in the neighbourhood).

The assumptions of this model are:

- $v_i \sim \text{Normal}(0, \sigma_v^2)$
- u_i is modelled as an *intrinsic conditional autoregressive* (iCAR) process, presented by Besag et al. (1991) with the following specification:

$$u_i | \boldsymbol{u}_{-i} \sim \text{Normal}\left(\frac{1}{\mathcal{N}_i} \sum_{j=1}^{n} a_{ij} u_j, s_i^2\right). \tag{2.33}$$

  where $\boldsymbol{u}_{-i}$ is the vector with all the elements of $\boldsymbol{u}$ except the i^{th} element; $a_{ij} = 1$ if the areas i and j are neighbours and $a_{ij} = 0$ otherwise; $s_i^2 = \sigma_u^2 / \mathcal{N}_i$ is the variance for the area i, where $\mathcal{N}_i = \#\mathcal{N}(i)$. According to Blangiardo and Cameletti (2015), s_i^2 informs that in the presence of powerful spatial correlation, the more neighbours an area has, the more information there is in the data about the value of its random effect u_i and σ_u^2 controls the quantity of variance between partially structured random effects.

The random spatial effect under the iCAR process associated with the exchangeable random spatial effect (v_i) gives rise to the *Besag-York-Mollièe* (BYM) model. In Sections 4.1 and 4.2 it is possible to see models with the unstructured random effect v_i, modelled as exchangeable (i.i.d model), models with the structured random effect u_i as an iCAR (Besag model) and finally models with the unstructured random effect v_i and with the structured random effect u_i as an iCAR (BYM model).

2.2.3.2 Spatio-temporal Models

Models suitable for investigating a spatial pattern over time are spatio-temporal models. Spatial models are simply extended to spatio-temporal models and are used in disease mapping combined with

surveillance studies, that is, the aim is to identify the spatial and temporal pattern of diseases. Updating the example from the previous section, the focus is now on evaluating the spatial pattern of hospital admissions due to cerebro-cardiovascular diseases in areas of Portugal for the years 2010-2019. Hence, the model of Section 2.2.3.1 was updated where y_{it} is the number of observed cases in area i and time t and E_{it} is the number of expected cases in area i and time t:

$$Y_{it} \sim \text{Poisson}(\lambda_{it}) \text{ with } \lambda_{it} = E_{it} p_{it},\ i = 1, \ldots, n \text{ and } t = 1, ..., T; \tag{2.34}$$

$$\begin{aligned} \log(\lambda_{it}) &= \eta_{it} \\ \log(\lambda_{it}) &= b_0 + \sum_{j=1}^{n_\beta} \beta_j x_{jit} + u_i + v_i + \text{Temporal}_t + \log(E_{it}), \end{aligned} \tag{2.35}$$

where Temporal_t may represent several terms, depending on the type of analysis, as shown below. The logarithm of the expected number of cases for region i and year t ($\log(E_{it})$) is the offset and represents the exposed population in region i in year t.

Starting with a simple model and assuming that spatial and temporal effects are present but are independent of each other (there is no spatio-temporal relationship), the linear predictor can be written as:

$$\eta_{it} = b_0 + \underbrace{u_i + v_i}_{\text{spatial effect}} + \underbrace{\gamma_t + \phi_t}_{\text{temporal effect}}, \tag{2.36}$$

where u_i (specified in Equation 2.33) and γ_t are structured effects and v_i and ϕ_t are unstructured effects.

Thus, there is an impact of space and an impact of time, and when the two are joint the result is additive. As γ_t represents the temporal correlation between the years, it can be specified using *random walk* (RW) of order 1 or 2 (Blangiardo and Cameletti, 2015):

- RW of order 1: $\gamma_t | \gamma_{t-1} \sim \text{Normal}(\gamma_{t-1}, \sigma^2)$
- RW of order 2: $\gamma_t | \gamma_{t-1}, \gamma_{t-2} \sim \text{Normal}(2\gamma_{t-1} + \gamma_{t-2}, \sigma^2)$

In turn, the temporal unstructured effect ϕ_t is specified by means of a Gaussian exchangeable *prior*, $\phi_t \sim \text{Normal}\left(0, \frac{1}{\tau_\phi}\right)$ (Blangiardo and Cameletti, 2015).

If the interaction between area and time is considered, the model described in Equation 2.36 will contain one more component:

$$\eta_{it} = b_0 + \underbrace{u_i + v_i}_{\text{spatial effect}} + \underbrace{\gamma_t + \phi_t}_{\text{temporal effect}} + \underbrace{\delta_{it}}_{\text{interaction}}, \tag{2.37}$$

- $\delta_{it} \sim \text{Normal}\left(0, \frac{1}{\tau_\delta \boldsymbol{R}_\delta}\right)$
- τ_δ is an unknown scalar
- $\boldsymbol{R}_\delta$ is the structure matrix, that identify the type of spatial and/or temporal dependence between the elements of δ.

Thus, $\boldsymbol{R}_\delta$ results from calculating the Kronecker product of the interacting random effects matrices. As Blangiardo and Cameletti (2015) show, there are four types of interaction taking into account what happens with space and time:

Type I interaction: assumes that the unstructured effect v_i and the unstructured effect ϕ_t interact:

$$\boldsymbol{R}_\delta = \boldsymbol{R}_v \otimes \boldsymbol{R}_\phi = \boldsymbol{I} \otimes \boldsymbol{I} = \boldsymbol{I} \tag{2.38}$$

Thus, there is no a spatial or temporal structure on the interaction, so the elements of δ_{it} are *iid* and $\delta_{it} \sim \text{Normal}\left(0, \frac{1}{\tau_\delta}\right)$. Thus, the matrix $\boldsymbol{R}_\delta$ has a rank of nT.

Type II interaction: assumes that the unstructured effect v_i (there is no spatial structure) and the structured effect γ_t (there is temporal structure) interact:

$$\boldsymbol{R}_\delta = \boldsymbol{R}_v \otimes \boldsymbol{R}_\gamma, \tag{2.39}$$

where $\boldsymbol{R}_v = \boldsymbol{I}$ and $\boldsymbol{R}_\gamma$ is the neighborhood structure specified using a first or second-order random walk. Thus, the matrix $\boldsymbol{R}_\delta$ has a rank of $n(T-1)$ for a first-order RW and $n(T-2)$ for a second-order RW. In this case, the temporal dependence structure for each area is independent of the temporal dependence structure of other areas.

Type III interaction: assumes that the structured effect u_i (there is spatial structure) and the unstructured effect ϕ_t (there is no temporal structure) interact:

$$\boldsymbol{R}_\delta = \boldsymbol{R}_u \otimes \boldsymbol{R}_\phi, \tag{2.40}$$

where $\boldsymbol{R}_\phi = \boldsymbol{I}$ and $\boldsymbol{R}_u$ is the neighboring structure defined through the CAR specification. Thus, the matrix $\boldsymbol{R}_\delta$ has a rank of $T(n-1)$. In this case, the spatial dependence structure for each time point is independent of the spatial dependence structure of other time points.

Type IV interaction: assumes that the structured effect u_i (there is spatial structure) and the structured effect γ_t (there is temporal structure) interact:

$$\boldsymbol{R}_\delta = \boldsymbol{R}_u \otimes \boldsymbol{R}_\gamma, \tag{2.41}$$

where $\boldsymbol{R}_u$ is the neighbouring structure defined by applying the iCAR specification and $\boldsymbol{R}_\gamma$ is specified using a first or second-order random walk. Thus, the matrix $\boldsymbol{R}_\delta$ has a rank of $(n-1)(T-1)$ for a first-order RW and $(n-1)(T-2)$ for a second-order RW. In summary, in the present case, the temporal dependence structure for each area depends on the temporal dependence structure of other areas.

This synthesised information is presented in the following table (Blangiardo and Cameletti, 2015):

Table 2.1: Summary of types of interaction.

Interaction	Parameter interaction	Rank
I	v_i and ϕ_t	nT
II	v_i and γ_t	$n(T-1)$ for RW1 $n(T-2)$ for RW2
III	u_i and ϕ_t	$(n-1)T$
IV	u_i and γ_t	$(n-1)(T-1)$ for RW1 $(n-1)(T-2)$ for RW2

2.2.4 Model checking and selection

In Bayesian modelling, according to Blangiardo and Cameletti (2015), it is crucial to check the plausibility and fit of the model. To this end, the evaluation of the variables that should be included in the model to define the best fit for the data in question, the comparison of models with different variables and the distribution of parameters and hyperparameters are highlighted. However, in practice, there are two commonly used methods to check models: methods based on the predictive distribution and methods based on the deviance.

2.2.4.1 Methods based on the predictive distribution

In this approach, the sample $\boldsymbol{y}$ is divided in two groups, so that $\boldsymbol{y} = (\boldsymbol{y}_a, \boldsymbol{y}_b)$. The first group $\boldsymbol{y}_a$ is used to fit the model and to estimate the *posterior* distribution of the parameters, and the second group $\boldsymbol{y}_b$ is used to perform the model criticism. Furthermore, there are two procedures to create these two groups, to assess the plausibility of the model assumptions and to detect the presence of outliers: cross-validation and *posterior* predictive check.

Cross-validation

In the cross-validation procedure, each observation will belong to one of the groups $\boldsymbol{y}_a$ or $\boldsymbol{y}_b$. After obtaining the two groups, cross-validation is based on two quantities, in order to assess the quality of the model (considering that $\boldsymbol{y}_a = y_{-i}$ and $\boldsymbol{y}_b = y_i$):

1. the *conditional predictive ordinate* (CPO):

$$\text{CPO}_i = p(y_i^* | \boldsymbol{y}_a) = p(y_i^* | \boldsymbol{y}_{-i}).$$

   If the sum of the log of the values of CPO is a large number, it means that the quality of the predictive model is good. Thus, when analysing competitive models, the one with the highest value shows a better fit.

2. the *probability integral transform* (PIT):

   $\text{PIT}_i = p(y_i^* \leq y_i|\boldsymbol{y}_a) = p(y_i^* \leq y_i|\boldsymbol{y}_{-i})$ if $\boldsymbol{y}$ comes from continuous distributions,

   $\text{PIT}_i^{adj} = \text{PIT}_i + 0.5 \times p(y_i^* = y_i|\boldsymbol{y}_a) = \text{PIT}_i + 0.5 \times p(y_i^* = y_i|\boldsymbol{y}_{-i})$ if $\boldsymbol{y}$ comes from discrete distributions.

   In relation to the PIT, if the empirical distribution is Uniform (see the histogram of the PIT), it means that the predictive distribution is consistent with the data.

***Posterior* predictive check**

The *posterior* predictive check highlights that $\boldsymbol{y}_a = \boldsymbol{y}_b = \boldsymbol{y}$, so all the observations are used to fit the model and to estimate the *posterior* distribution of the parameters and to perform the model criticism. In this approach are considering two quantities:

1. the *posterior predictive distribution:*

$$p(y_i^*|\boldsymbol{y}) = \int p(y_i^*|\theta_i)p(\theta_i|\boldsymbol{y})d\theta_i.$$

   If $p(y_i^*|\boldsymbol{y})$ is a small value, the observations are located in the tails of the distribution and can be classified as outliers. In addition to this, if there are many small values, the model is not adequate for the data in question.

2. the *posterior predictive p-value*:

$$p(y_i^* \leq y_i|\boldsymbol{y}).$$

   If the values of $p(y_i^* \leq y_i|\boldsymbol{y})$ were close to 0 or 1 it means that the model does not fit the data adequately.

On top of that, there are two summary indices that allows assessing of the goodness of fit of the model:

1. mean squared error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2.$$

2. R squared (R^2):

$$\text{R}^2 = \frac{\sum_{i=1}^{n}(y_i^* - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

2.2.4.2 Methods based on the deviance

A alternative to the above mentioned methods is the method based on the deviance. Firstly, it is necessary to understand how to calculate the deviance:

$$D(\theta) = -2\log\left(p(\boldsymbol{y}|\boldsymbol{\theta})\right). \tag{2.42}$$

Considering the example presented in Blangiardo and Cameletti (2015), let $\boldsymbol{y}$ be a sample such that:

$$Y_i \sim \text{Poisson}(\lambda).$$

So the *likelihood* of this model is:

$$p(\boldsymbol{y}|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!},$$

and the deviance is equal to:

$$D(\lambda) = -2\left(\sum_{i=1}^{n} y_i \log(\lambda) - n\lambda - \sum_{i=1}^{n} \log(y_i!)\right).$$

Deviance information criterion (DIC)

Blangiardo and Cameletti (2015) explain that the deviance information criterion is the most frequently used measure of model fit. The DIC is a generalisation of the well-known Akaike information criterion (AIC), which is used for model comparison. Specifically, the DIC results from the sum of two components: the first corresponds to the *posterior* expectation of the deviance $D(\boldsymbol{\theta})$ and the second corresponds to the *effective number of parameters*. So, the DIC is given by:

$$DIC = \bar{D} + p_D, \tag{2.43}$$

where

$$p_D = E_{\boldsymbol{\theta}|\boldsymbol{y}}(D(\boldsymbol{\theta})) - D(E_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta})) = \bar{D} - D(\bar{\boldsymbol{\theta}}).$$

Thus, the model with the lowest DIC fits the data better.

2.2.5 Tools for Bayesian inference

In software R, the package used for Bayesian inference using the INLA approach is `INLA`. A brief description of the use of this package will be given below (according to Blangiardo and Cameletti (2015) and Blangiardo, Cameletti, et al. (n.d.)). The model formula is built in the same way as the equation for regression models, however in Bayesian inference it is possible to include random effects for both area and time through the function *f(.)*. Considering two covariates x_1 and x_2 and the random effect z_1, an general form of the formula is:

```
formula <- y~x1+x2+f(z1,model = "")
```

By default, the argument `model` is equal to `iid` and it should be applied when the random effect `z1` is independent and Gaussian distributed. The list of the other alternatives are present in Blangiardo and Cameletti (2015).

Then, through the function `inla()`, it is run the INLA algorithm:

```
model <- inla(formula, family="", offset, data)
```

The argument `family` is a string that specifies the distribution of the data and in the case of this study, the distribution of the number of hospital admissions is Poisson.

Then, with the aim of choosing the model that better explains the data in study, the DIC is observed in the output of `summary(model)` (a model with the lowest DIC is the best model). This output has also the posterior mean, the standard deviation and the quartiles of the fixed effects (in this case is β_0, β_1 and β_2) and the random effects. The posterior mean for the response variable is obtained through the command `model$summary.fitted.values$mean` .

Finally, with the aim of checking the plausibility and fit of the model, some methods can be used. The approach used in this analysis was the posterior predictive check. The *posterior predictive distribution* is represented by the scatter plot of the posterior mean for the predictive distributions against the observed values and *posterior predictive p-values* are represented in a histogram. Thus, this two representations are obtained through the following commands, respectively:

```
plot(data$y,model$summary.fitted.values$mean,
     xlab="Observed values",ylab="Mean Post. Pred. Distr.")

predicted.p.value<-c()
for(i in (1:n)) {
  predicted.p.value[i] <- inla.pmarginal(q=data$y[i],
                                         marginal=model$marginals.fitted.values[[i]])
}
hist(predicted.p.value,main="",xlab="Posterior predictive p-value")
```

Note that, if the distribution of the points in the scatter plot is similar to a straight line, it is possible to conclude that, on average, the prediction is very close to the observed values. On the other hand, if there is a high number of areas with p-values close to 0.5 (in the middle of the histogram) and few areas whose p-value is very low or high, it is possible to conclude that the model fits the data well.

3. Exploratory data analysis

3.1 The data

The data used in this project relate to hospital admissions records in mainland Portugal, provided by the Administração Central do Sistema de Saúde, I.P (ACSS) of the Ministério da Saúde. To obtain a database with only the necessary information, a process of cleaning and organising the available data was carried out:

1. The Diagnosis and Episodes files of the various periods of each year were aggregated, and duplicate observations were eliminated (some periods overlap and, therefore, repeated observations were discarded).
2. Filter the CCD by principal diagnosis code according to the 9th revision (ICD-9-CM) for years between 2010 and 2016 or the 10th revision (ICD-10-CM) from 2017 onwards.
3. Remove individuals whose diagnosis by CCD is not the principal diagnosis.
4. Remove individuals whose district does not belong to mainland Portugal.
5. Remove individuals whose district code is "99" (the proportion of these cases is less than 0.001).
6. Remove the individuals whose fictitious patient number was negative.

After cleaning and organising the data, it was found that the sample sizes were very distinct. The years 2011, 2013, 2015, 2017 and 2019 had about half of the observations compared to the other years. As it was not possible to solve this matter in time for the delivery of this project, only the years 2010, 2012, 2014, 2016 and 2018 are plausible for the analysis. Considering only these five years, the sample size is 701 786, and it is divided as follows:

- **2010:** $n = 144\,533$
- **2012:** $n = 145\,618$
- **2014:** $n = 146\,254$
- **2016:** $n = 132\,631$
- **2018:** $n = 132\,750$

From these databases, the number of hospital admissions due to CCD at the area level was extracted. In addition, the district code, the district name, the municipality code and the municipality name were also extracted in order to specify the patient's area.

In addition to these variables, variables related to risk factors for CCD and some factors affecting good hospital functioning at area level were introduced. These variables were obtained from INE at municipality level. To build the database at district level, these variables were aggregated at district level. Furthermore, these variables are classified as follows:

- **Socio-demographic risk factors:** proportion of residents aged 65 or over, proportion of male residents, population density and unemployment rate;
- **Clinical risk factors:** mortality rate per 1000 inhabitants due to diabetes.

At district level, the proportion of members in federations (proportion of individuals who attend sports academies, that is, it is an indicator variable of the sedentarism of individuals) was also extracted from INE and considered as a risk factor, as well as the number of hospitals per 1000 inhabitants, at municipality level. The total population was also extracted from INE to be used as an offset in the models used later.

Thus, the database, with the combination of variables from the databases provided by ACSS and the variables extracted from INE, is composed by the following variables:

- **Socio-demographic variables:** district code and district name (for the database at district level) or municipality code and municipality name (for the database at municipality level), proportion of residents aged 65 or over, proportion of male residents, population density, unemployment rate, proportion of members in federations (district level only) and total population.
- **Clinical variables:** number of admissions due to cerebro-cardiovascular diseases, number of hospitals per 1000 inhabitants (municipality level only) and mortality rate per 1000 inhabitants due to diabetes.

As already mentioned, all these variables are aggregated by demographic area (districts and municipalities) and by time unit (year) and will be included as covariates in the models described and analysed below. Note that the size of the final database at district level is 90 (18 districts $\times$ 5 years) and the size of the database at municipality level is 1390 (278 municipalities $\times$ 5 years). These data will be used to study cerebro-cardiovascular diseases and their risk factors.

3.1.1 Multiple imputation

The variable mortality rate per 1000 inhabitants due to diabetes was created using the number of deaths due to diabetes per municipality, available in INE. Since it had missing values, the following boxplots show the distribution of values of the remaining variables in the group of municipalities with missing values and in the one without missing values.

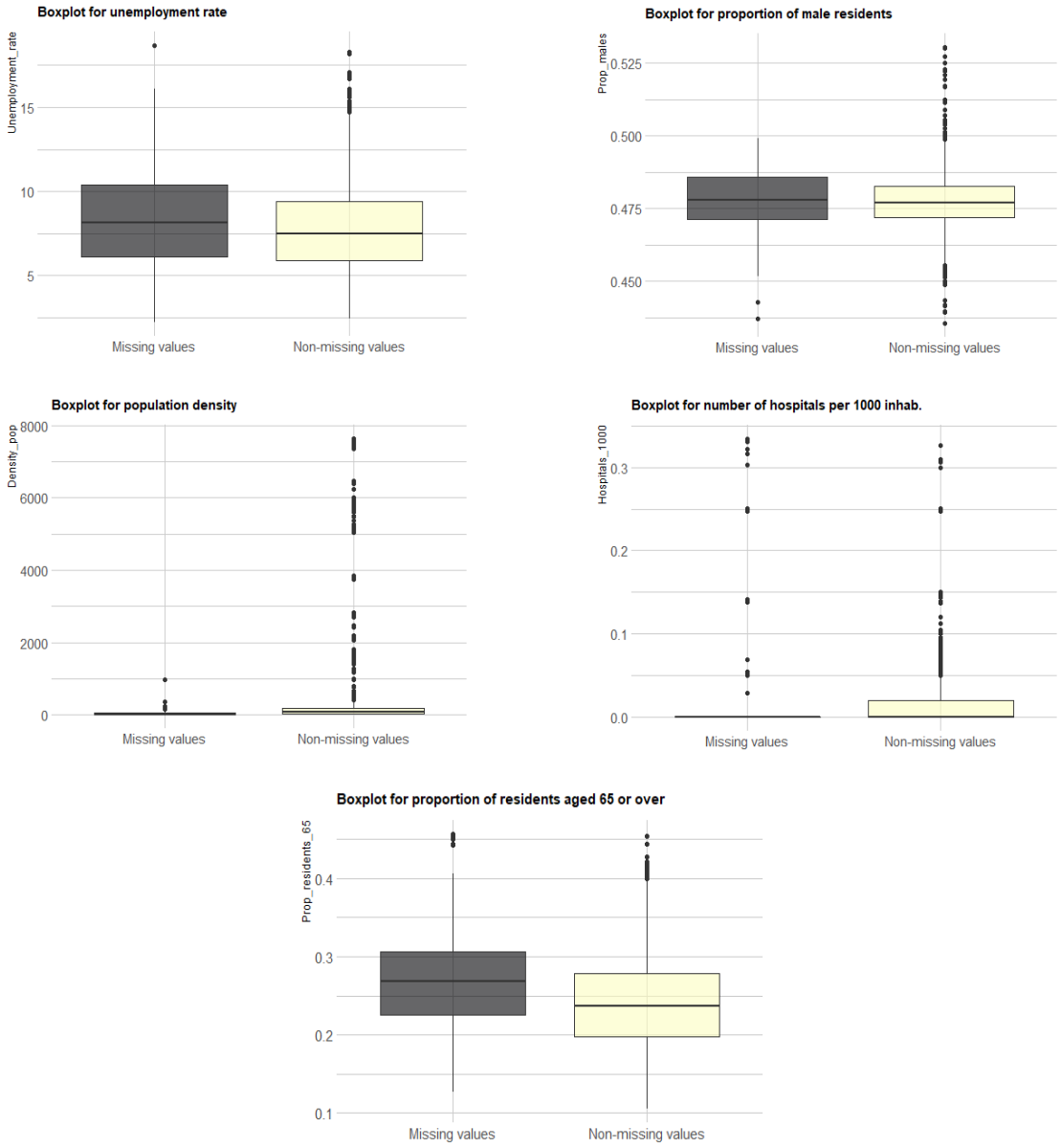

Figure 3.1: Boxplots of the remaining variables taking into account the missing and non-missing values of the variable number of deaths due to diabetes.

When analysing the boxplots, it can be concluded that municipalities, whose number of deaths due to diabetes is missing, have a higher unemployment rate, a lower population density, a lower number of hospitals per 1000 inhabitants and a higher proportion of residents aged 65 or over. In Section 2.1, the suspicion that the missing values corresponded mostly to municipalities in the interior of Portugal was mentioned, and the conclusion drawn from the boxplots analysis, on the characteristics of these municipalities, is in accordance with this suspicion.

Since there is only one variable (candidate explanatory variable) with missing values, whose percentage is very low (about 5%), it seems wise to use multiple imputation. Therefore, the missing values of the variable number of deaths due to diabetes were imputed using multiple imputation. It should be noted that, as this method is not compatible with Bayesian models, the second (analysis) and third (pooled) steps of multiple imputation were not implemented. Thus, the final data with imputed values were obtained by the average of the five imputed datasets (m parameters chosen).

3.2 Descriptive analysis of the study variables

In this chapter, the explanatory analysis of all the variables, described in Section 3.1, was done at year, district and municipality level. Tables of the descriptive values of the hospital admission rate and INE variables at annual level are presented in Appendice A. The variation of the hospital admission rate due to CCD and INE variables, over time, are expressed in the following graphs:

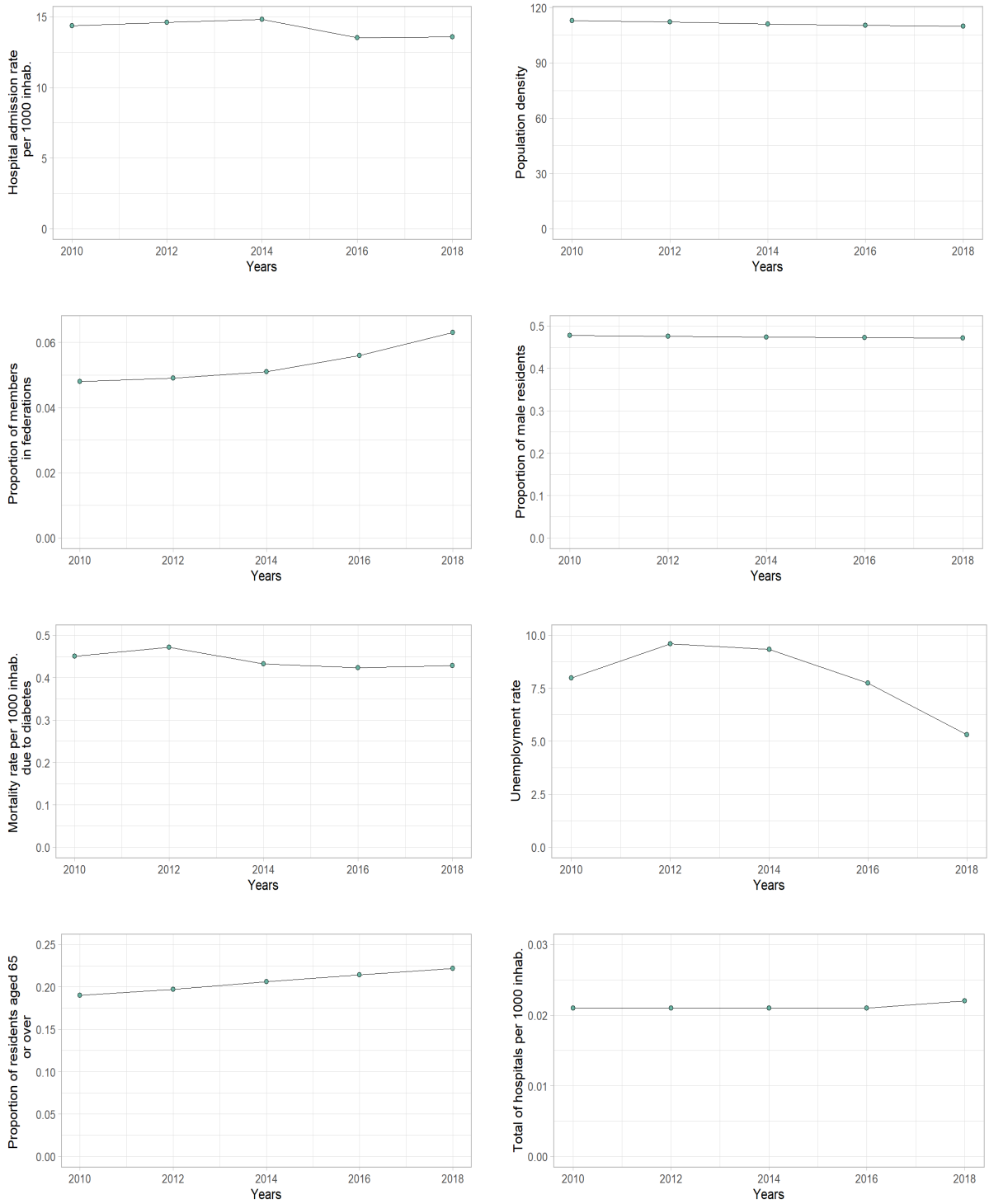

Figure 3.2: Temporal evolution of the study variables at year level.

Analysing the Figure 3.2, it can be concluded that the hospital admission rate increased slightly from 2010 to 2014, where it reached a maximum of 14.818 admissions per 1000 inhabitants. From 2014 there was a decrease until 2016, where it reached a minimum of 13.521 admissions per 1000 inhabitants, and then remained practically constant until 2018. In relation to mainland Portugal, the population density decreased slightly from 2010 to 2018 from 112.9 to 109.8 inhabitants per km^2. The proportion of members in federations appears to have increased exponentially from 2010 to 2018, where it reached the maximum of 0.063. The proportion of male residents decreased slightly from 2010 to 2018, where it reached the minimum of 0.472, however the values were very similar in all years. The mortality rate per 1000 inhabitants due to diabetes suffered an increase from 2010 to 2012, where it reached a maximum of 0.472 deaths per 1000 inhabitants. After 2012, there was a decrease until 2016, where it reached a minimum of 0.423 deaths per 1000 inhabitants. Then remained practically constant until 2018. The unemployment rate experienced an increase from 2010 to 2012, where it peaked at 9.602 and then there was a decrease until 2018, where it reached a low of 5.307. The proportion of residents aged 65 or over increased at a certain rate until 2018, where it reached a maximum of 0.222. Finally, the number of hospitals did not change from 2010 to 2016, where there were 0.021 hospitals per 1000 inhabitants. However, in 2018, there were 0.022 hospitals per 1000 inhabitants.

At district level, the tables of the descriptive values of the hospital admission rate and INE variables are presented in Appendice A. The maps of the study variables from 2010 to 2016 are presented in Appendice B. In addition, the scatter plots of the same set of variables versus variable of interest (hospital admission rate) are also presented in Appendice B, in order to analyse the relationship between each candidate explanatory variable and the response variable. The same scheme was used at municipality level.

Considering the year 2018 (the year used in the spatial analysis), Table 6 (presented in Appendice A) and Figures 3.3 to 3.6 are considered. Note that, 2018 was the year chosen since it is the most recent year and as such is the most interesting to analyse. The maps below represent the data at district level:

3. EXPLORATORY DATA ANALYSIS

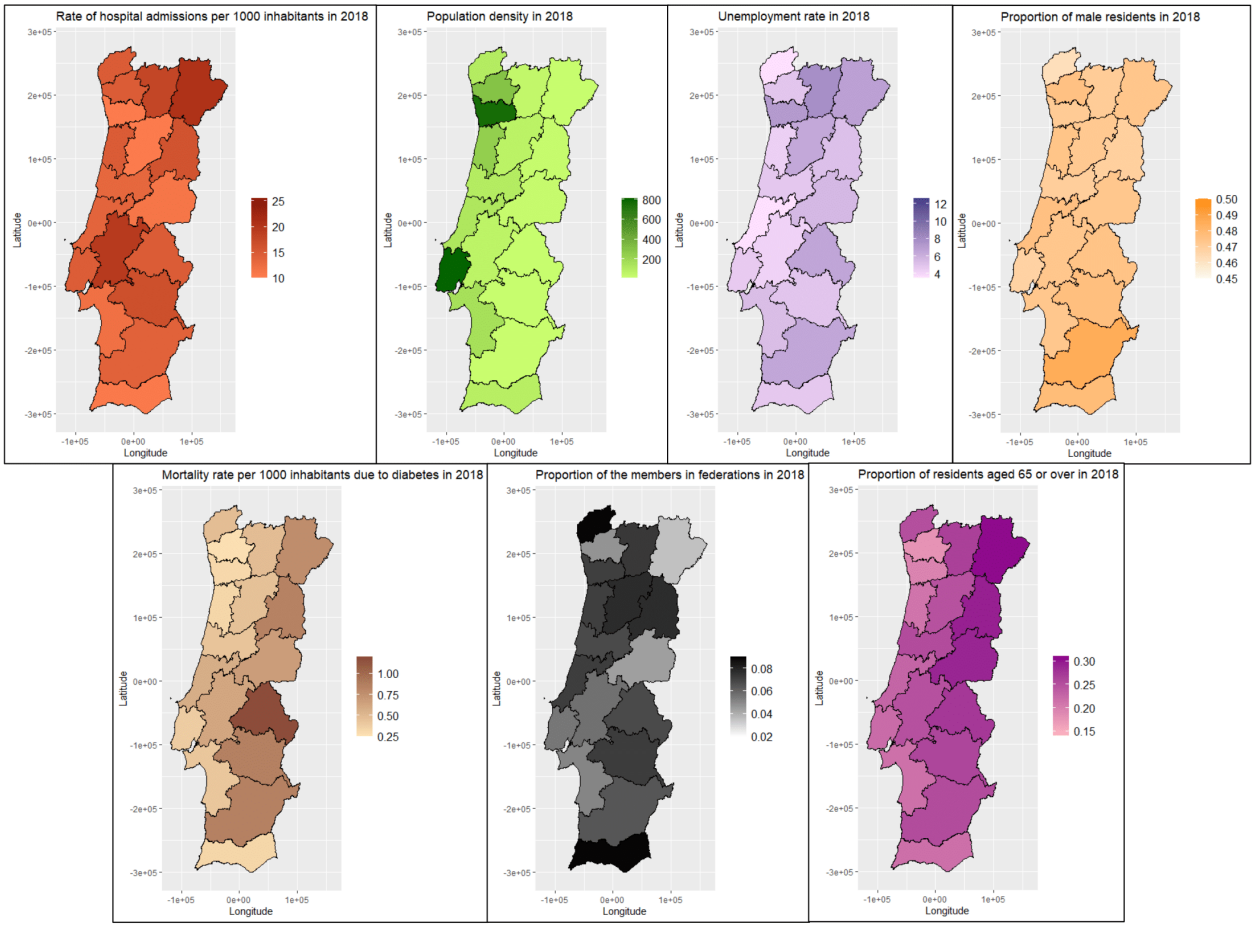

Figure 3.3: Maps at district level of the hospital admission rate and the variables from INE in 2018.

In 2018, the districts with the lowest hospital admission rate were Porto, Viseu and Faro. Castelo Branco was the district with the highest hospital admission rate in previous years, but in 2018 Bragança becomes the district with the highest rate, followed by Santarém and Vila Real. Regarding the variables from INE, Lisboa and Porto were the two districts with the highest population density in all years. The unemployment rate was the lowest in Viana do Castelo, Leiria and Santarém (the second district with the highest admission rate) and the highest in Vila Real, Bragança (the third and first districts with the highest admission rate) and Porto (the district with the lowest admission rate). The proportion of male residents was practically constant throughout all districts (approximately half of the residents), and this happens in all years. However, Beja stood out for having the highest proportion of men in all years. The mortality rate due to diabetes was the lowest in Braga and Porto (the district with the lowest admission rate) and the highest in Portalegre. The proportion of members in federations was the lowest in Bragança (the district with the highest admission rate) and the highest in Viana do Castelo, Faro and Viseu (the last two being the third and second districts with the lowest admission rate). Considering the population aged 65 or over, Braga and Porto (the district with the lowest admission rate) were the districts with the lowest proportion of older residents and Bragança was the district with the highest proportion of older residents and the district with the highest admission rate.

The scatter plots at district level are represented below:

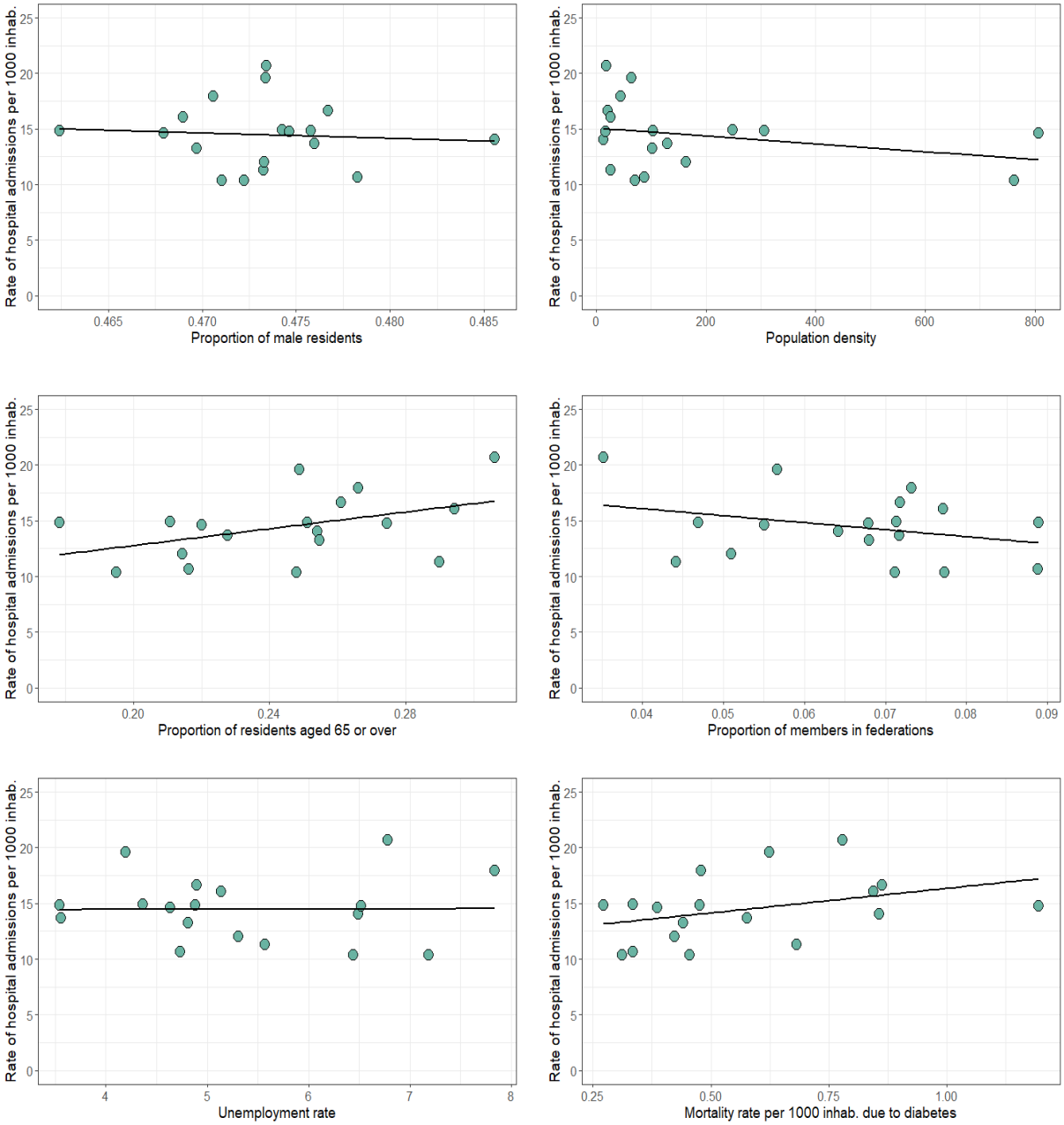

Figure 3.4: Scatter plots of the hospital admission rate vs all the variables at district level for 2018 and the respective regression line of the GLM model.

Observing the graphs in Figure 3.4, there seems to be a positive correlation between the variable hospital admission rate due to CCD and the variables proportion of residents aged 65 or over and mortality rate per 1000 inhabitants due to diabetes. This means that when the proportion of older residents increases, the hospital admission rate due to CCD also increases. In turn, high mortality rates due to diabetes correspond to high hospital admission rates due to CCD. In addition, there seems to be a negative correlation between the hospital admission rate and the variables proportion of members in federations and population density, that is, as the proportion of players and population density increase, the hospital admission rate decreases. Furthermore, there appears to be a slight negative correlation between the admission rate and the proportion of male residents, that is, as this variable increases, the admission rate declines. Finally, the correlation between the hospital admission rate and the unemployment rate is

apparently null.

At municipality level, the maps are expressed in the graphs below:

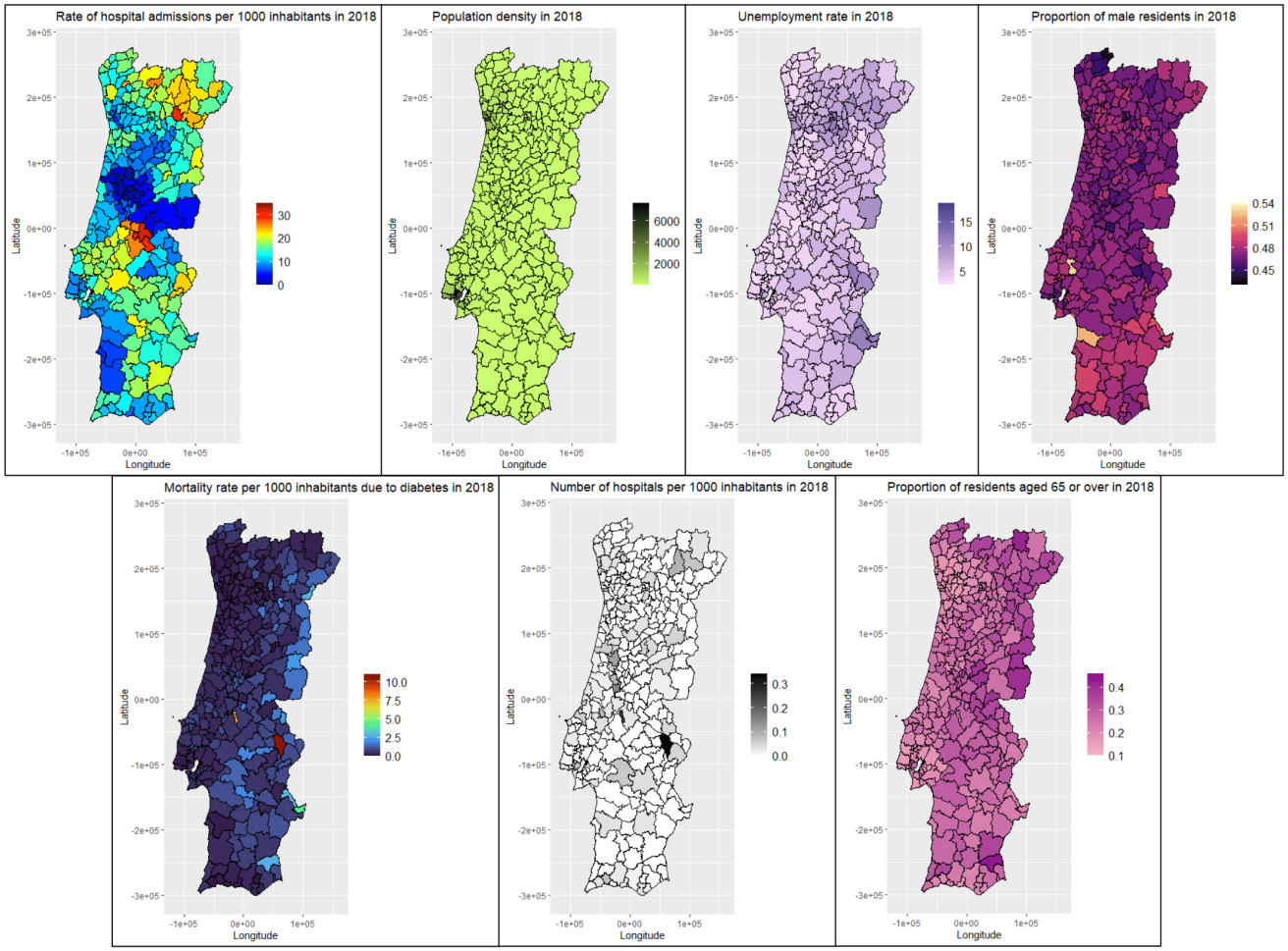

Figure 3.5: Maps at municipality level of the hospital admission rate and the variables from INE in 2018.

In 2018, the municipalities with the lowest hospital admission rate were Mortágua (0.226) and Miranda do Corvo (0.473), and the municipalities with the highest hospital admission rate were Mação (31.631) and Sardoal (31.292) (two municipalities that belong to Santarém). Concerning the INE variables, the municipalities with the highest population density were located in Lisboa and Porto in every year. The unemployment rate showed to be mostly higher in municipalities in the South and North of Portugal. However, Ourém, Batalha and Vila de Rei were the municipalities with the lowest unemployment rate, while Vila de Rei was the fifth municipality with the highest hospital admission rate. The proportion of male residents is practically constant throughout all municipalities (approximately half of the residents), and this happens in each year. The mortality rate due to diabetes was the highest in Monforte and Constância. The municipalities with the highest number of hospitals per 1000 inhabitants were Monforte (Portalegre) and Constância (Santarém), in all years. Regarding the proportion of residents aged 65 or over, it is the highest in Alcoutim, Vinhais and Idanha-a-Nova.

Finally, the scatter plots at municipality level for 2018 are represented below:

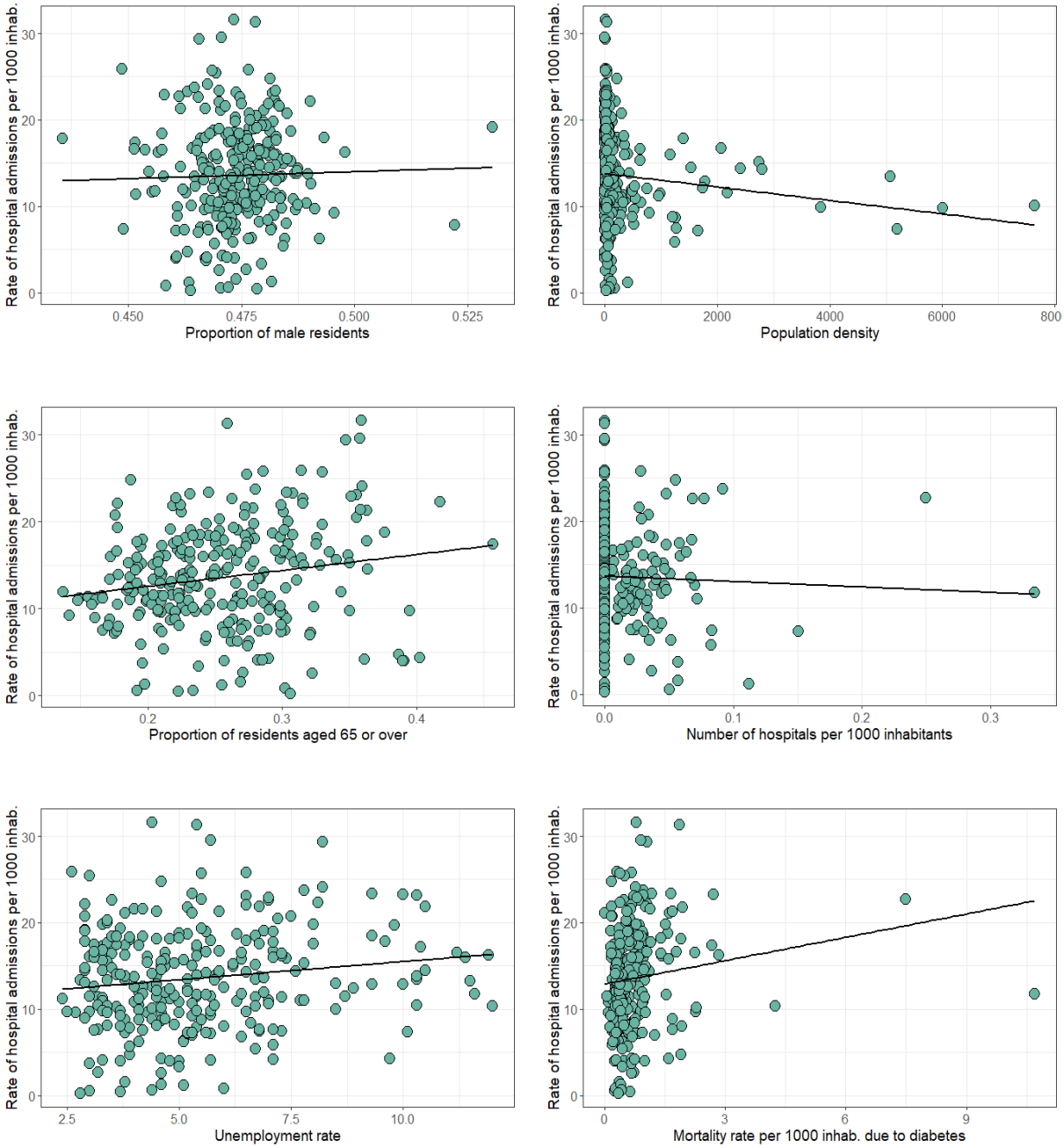

Figure 3.6: Scatter plots of the hospital admission rate vs all the variables at municipality level for 2018 and the respective regression line of the GLM model.

Analysing the scatter plots in Figure 3.6, there seems to be a positive correlation between the variable hospital admission rate due to CCD and the variable proportion of residents aged 65 or over, that is, when the proportion of older residents increase, the hospital hospital admission rate also increase. The regression line of the scatter plot of the mortality rate per 1000 inhabitants due to diabetes has a positive slope, however, it is difficult to analyse the data as there are many municipalities whose mortality rate is between 0 and 2.5 (which causes a large concentration of points at the beginning of the graph) and only three municipalities whose mortality rate is above 3 (these were considered to be influential values). There seems to be a slight positive correlation between the hospital admission rate and the proportion of male residents. The same is true for the unemployment rate (opposite of other years). Thus, when the

proportion of males and the unemployment rate increase, the hospital admission rate also goes up. The regression line of the scatter plot of the population density has a negative slope, but nevertheless, it is difficult to analyse the data, as there are many municipalities with very low population density (which causes a large concentration of points at the beginning of the graph). The same is true for the variable number of hospitals per 1000 inhabitants, since there are many municipalities without hospitals in all years. However, the correlation between the hospital admission rate and the number of hospitals appears to be slightly negative.

4. Application to the BDMH-ACSS data

The aim of this study is to analyse the regressors that best explain the number of hospital admissions due to CCD in Portugal over five years. To this end, we will perform two analyses, a spatial analysis and a spatio-temporal analysis, as explained below:

1. **Spatial analysis** - study the variation of admissions in different regions in a specific year (2018, in the case of this study), considering the spatial organisation of the areas and the fact that the areas are not independent of each other. This spatial structure is represented in the model by introducing structured random effects;

2. **Spatio-temporal analysis** - analyse the evolution of admissions in the regions over time. Thus, the spatial and temporal structure are considered and these two components are introduced in the model through the so-called structured random effects. Therefore, it is possible to have a space-time interaction.

For these analyses, generalised linear mixed models on the Bayesian hierarchical approach were used to model the number of hospital admissions, using the Poisson model. This procedure was described in Section 2. Consequently, the number of hospital admissions was modelled at area and time level using the package `INLA` of the software *R Studio* ®.

4.1 Spatial analysis

As previously mentioned, the first aim of this study is to analyse how the hospital admission rate is distributed in space and to identify which regressors best explain this spatial variation.

4.1.1 Data at district level

In the spatial analysis, the parameters associated with the regressors mentioned in Section 3.1 are as follows:

- β_1: Population density
- β_2: Unemployment rate
- β_3: Proportion of male residents
- β_4: Mortality rate per 1000 inhabitants due to diabetes
- β_5: Proportion of members in federations
- β_6: Proportion of residents aged 65 or over

4. APPLICATION TO THE BDMH-ACSS DATA

In statistics, to interpret variables expressed on different scales, it is necessary to standardise them. Considering the observed values related to variable X, the standardized data z is obtained taking into account the following formula:

$$z = \frac{x - \bar{x}}{s}. \tag{4.1}$$

Thus, the standardized values have mean 0 and standard deviation 1.

After the standardisation of the variables, the number of hospital admissions will be modelled under the latest available year (2018). As the response variable is a count, the distribution chosen was Poisson:

$$\begin{aligned} Y_i &\sim \text{Poisson}(\lambda_i),\ i = 1, ..., 18, \\ \log(\lambda_i) &= \eta_i. \end{aligned} \tag{4.2}$$

The initial model, without taking into account the spatial component (without random effects), is defined as follows:

$$\eta_i = b_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \text{offset}_i, i = 1, ..., 18, \tag{4.3}$$

As explained in Section 2.2.3.1, to include the spatial variation and the spatial dependence between areas, random effects must be incorporated into the model. Note that v_i is the unstructured effect and u_i is the structured effect.

Thus, the model with the unstructured effect (i.i.d model) is defined as follows:

$$\eta_i = b_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + v_i + \text{offset}_i, i = 1, ..., 18, \text{ with} \tag{4.4}$$

$$v_i \sim \text{Normal}(0, \sigma_v^2).$$

The model with the structured effect (Besag model) is defined in the following way:

$$\eta_i = b_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + u_i + \text{offset}_i, i = 1, ..., 18 \text{ with} \tag{4.5}$$

$$u_i | \tau_u \sim \text{iCAR}(\tau_u),$$

that is, u_i is modelled as a *intrinsic conditional autoregressive* (Equation 2.33).

Finally, the model with the unstructured and structured effects (BYM model) is represented as follows:

$$\eta_i = b_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + v_i + u_i + \text{offset}_i, i = 1, ..., 18, \tag{4.6}$$

where v_i and u_i were defined in the previous models.

Thus, by default, the priors for the hyperparameters are specified as follows:

$$
\begin{aligned}
b_0 &\sim \text{Normal}(0, 10^6), \\
\beta_j &\sim \text{Normal}(0, 10^6),\ j = 1, \ldots, 6, \\
\log(\tau_v) &\sim \text{logGamma}(1, 0.0005), \\
\log(\tau_u) &\sim \text{logGamma}(1, 0.0005).
\end{aligned}
$$

The best model was chosen using the variable selection process under the DIC criterion (note that the model with the lowest DIC is better). The DIC decreased considerably with the inclusion of all variables and, consequently, with the inclusion of random effects. The model with the unstructured random effect and the model with the structured and unstructured random effects have the same DIC (222.53). That said, the model chosen at district level is the simplest model, that is, the model with only the unstructured random effect (i.i.d model).

Table 4.1 presents the summary of the posterior statistics for the fixed and random effects of the spatial model at district level.

Table 4.1: Posterior mean, posterior standard deviation and posterior 95% credibility interval for the parameters and hyperparameters of the spatial model at district level.

Parameter	**Mean**	**Standard deviation**	**2.5% Quantile**	**97.5% Quantile**
(Intercept)	-4.255	0.050	-4.355	-4.156
Population density (β_1)	-0.036	0.086	-0.206	0.135
Unemployment rate (β_2)	-0.030	0.059	-0.148	0.088
Prop. male residents (β_3)	-0.049	0.078	-0.203	0.105
‰ of mort. due to diabetes (β_4)	0.063	0.092	-0.120	0.246
Prop. members federations (β_5)	-0.069	0.057	-0.181	0.044
Prop. residents +65 (β_6)	0.012	0.121	-0.229	0.253
τ_v	26.65	10.25	10.41	50.21

As can be seen from the table above, there are no significant variables for the model. However, the random effect is significant, meaning that this model takes into account that the districts are different from each other with respect to the study data.

Figure 4.1 represents the posterior mean of the unstructured spatial random effect v_i. The unstructured spatial random effect (there is no spatial correlation in the way the hospital admission rate is modelled) can be seen as the variation in the hospital admission rate explained by the spatial distribution of districts. Since the districts are coloured with different colours, the introduction of the random effect in the model is justified, so spatial differences in the hospital admission rate between districts can be noted.

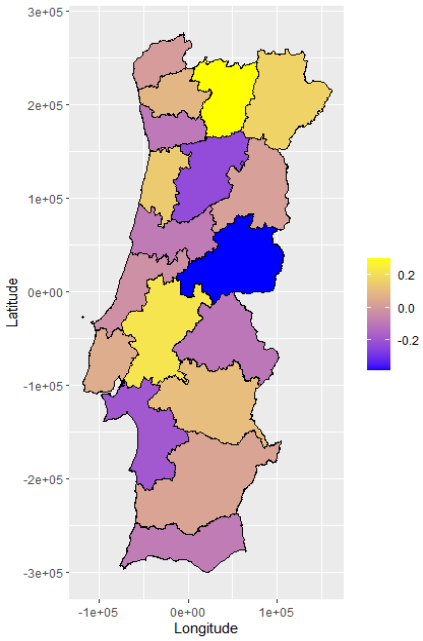

Figure 4.1: Posterior mean of the unstructured spatial random effect in the spatial model for data at district in 2018.

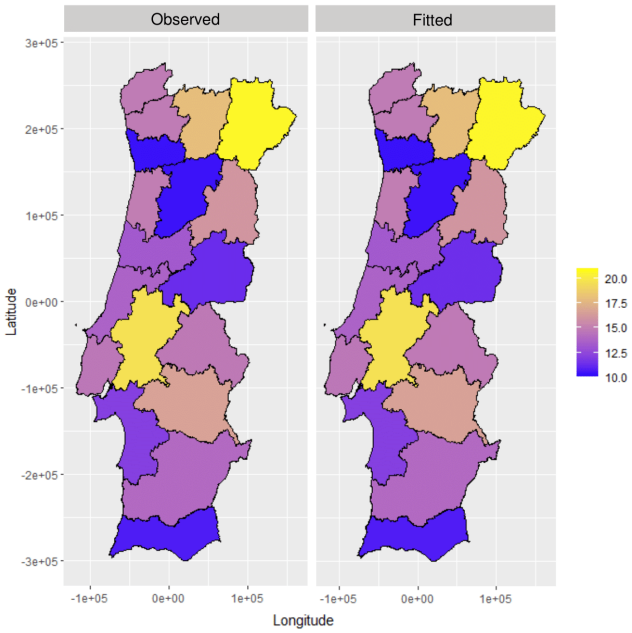

Figure 4.2: Maps of observed and fitted values of the hospital admission rate in 2018 at district level.

Figure 4.2 shows the observed values of the hospital admission rate (left) and the fitted values of the hospital admission rate (right) in 2018 at district level, given by η_i (Equation 4.4). These fitted values are the posterior mean of the hospital admission rate for each district, according to the model chosen. As the map on the left in the Figure 4.2 shows, in 2018, the districts with the highest hospital admission rate due to CCD were Bragança (20.647), Santarém (19.599) and Vila Real (19.973) and the districts with the lowest relative risk were Porto, Viseu and Faro (all around 10). Furthermore, the map on the right in Figure 4.2 reveals that the difference between the observed and fitted values is very small, so it is possible to conclude a good fit of the model to the data at district level for 2018.

4.1.2 Data at municipality level

At municipality level, the regressors of the model are defined as follows:

- β_1: Population density
- β_2: Unemployment rate
- β_3: Proportion of male residents
- β_4: Mortality rate per 1000 inhabitants due to diabetes
- β_5: Number of hospitals per 1000 inhabitants
- β_6: Proportion of residents aged 65 or over

Note that standardisation was also used in this set of variables, as it was in the variables at district level.

That said, the models at municipality level that include the unstructured, structured or both random effects are described in the same way as the models at district level, but i is defined from 1 to 278 (number

of municipalities). The best model was also chosen using the variable selection process under the DIC criterion. The DIC decreased, again, with the inclusion of all variables and, consequently, with the inclusion of random effects. The model with the structured random effect has the lowest DIC (2549.63), so it was the model chosen (Besag model).

Table 4.2 shows the summary of the posterior statistics for the fixed and random effects of the spatial model at municipality level.

Table 4.2: Posterior mean, posterior standard deviation and posterior 95% credible interval for the parameters and hyperparameters of the spatial model for the data at municipality level.

Parameter	**Mean**	**Standard deviation**	**2.5% Quantile**	**97.5% Quantile**
(Intercept)	-4.428	0.006	-4.439	-4.416
Population density (β_1)	-0.018	0.042	-0.102	0.065
Unemployment rate (β_2)	0.020	0.037	-0.052	0.092
Prop. male residents (β_3)	-0.001	0.028	-0.055	0.053
‰ of mortality due to diabetes (β_4)	0.036	0.031	-0.026	0.097
Hospitals per 1000 inhab. (β_5)	-0.032	0.030	-0.090	0.027
Prop. residents +65 (β_6)	0.097	0.035	0.028	0.165
τ_u	1.690	0.163	1.390	2.030

Through the previous table, it is concluded that the variable proportion of residents aged 65 or over is significant to explain the variation in the hospital admission rate due to CCD at municipality level, accounting for the effect of spatial dependence. As the coefficient is positive, it can be concluded that regions with a high proportion of residents aged 65 or over tend to have a high hospital admission rate, so the proportion of residents aged 65 or over is a risk factor at municipality level. In addition, the random spatial effect is also significant for the model, that is, the spatial correlation is significant in explaining the variation in the hospital admission rate, which is not explained by the covariates.

Figure 4.3 represents the posterior mean of the structured spatial random effect u_i, that is, the spatial neighbouring structure is taken into account. The structured spatial random effect can be seen as the variation in the hospital admission rate explained by the correlation spatial. As the districts are coloured with different colours, it is justified to introduce the random effect in the model, that is, there is a spatial correlation between the municipalities in relation to the hospital admission rate due to CCD.

Figure 4.4 displays the observed values of the hospital admission rate (left) and the fitted values of the hospital admission rate (right) in 2018 at municipality level. These fitted values are the posterior mean of the hospital admission rate for each municipality, according to the model chosen. According to the map on the left in Figure 4.4, in 2018, the municipalities with the highest hospital admission rate due to CCD were located in the Médio Tejo region and in Bragança. The municipalities with the lowest hospital admission rate were located in Viseu, Coimbra, Aveiro and Castelo Branco. The map on the right in Figure 4.4 shows that the difference between the observed and fitted values is very small (the colors of

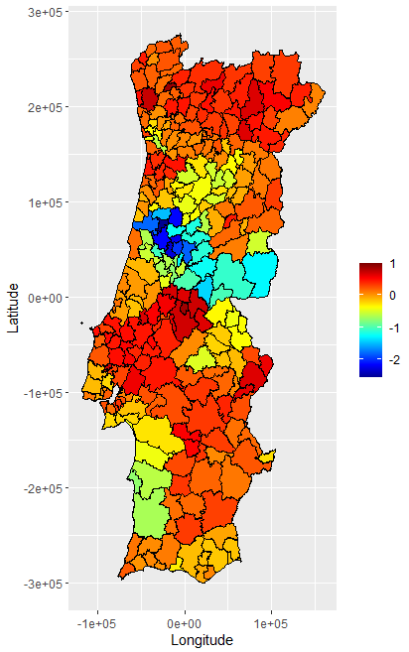

Figure 4.3: Posterior mean of the spatial random effect in the spatial model for data in 2018 at municipality level.

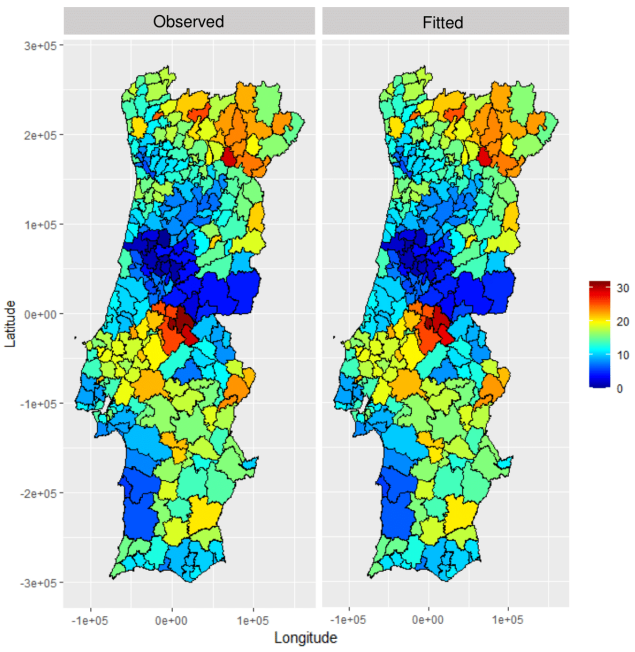

Figure 4.4: Maps of observed and fitted values of the hospital admission rate in 2018 at municipality level.

the map on the right are slightly lighter, that is, the fitted values are smaller), so it is possible to conclude a good fit of the model to the data at municipality level for 2018.

4.1.3 Diagnosis

As explained in Section 2.2.4 there are some methods to check and select the best model. The Deviance Information Criteria (DIC) (a method based on the deviance) for each model was specified throughout the spatial and spatio-temporal analysis without/with interaction. Note that models with smaller DIC are better supported by the data. However, in this chapter, the posterior predictive check will be used in order to analyse the fit of the models to the data. In accordance with this method, there are two quantities of interest: the *posterior predictive distribution* and the *posterior predictive p-value*. The posterior predictive distribution is represented by a scatter plot of the posterior mean for the predictive distributions against the observed values. So that the respective model fits the data reasonably well, the distribution of values in the scatter plot should be similar to a straight line. On the other hand, the representation of the predictive p-values is done by means of a histogram of the values. According to Blangiardo and Cameletti (2015), values of $p(y_i^* \leq y_i|\boldsymbol{y})$ near to 0 or 1 indicate that the model fails to fit the data, that is, the closer they are to 0.5 the better.

The Figures 4.5 and 4.6 show the scatter plots of the posterior mean for the predictive distributions against the observed values (left) and the histogram of the posterior predictive p-value (right) for the spatial model at district level and at municipality level, respectively.

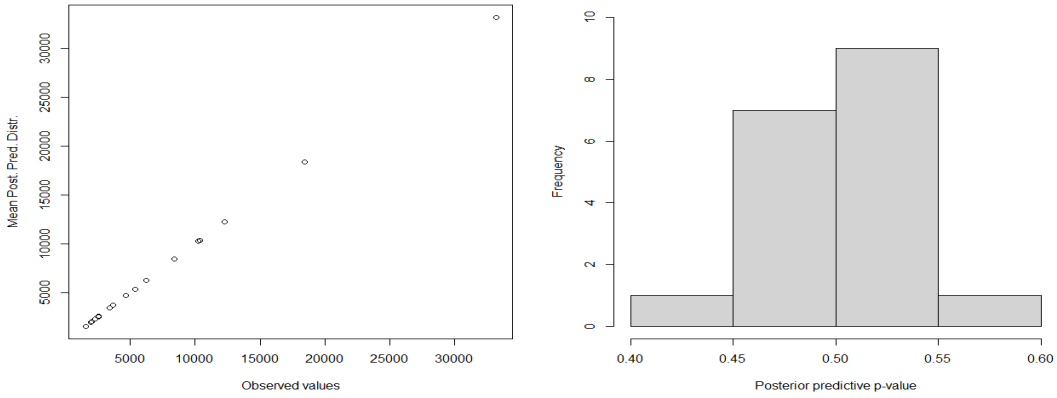

Figure 4.5: Scatter plot of the posterior mean for the predictive distributions against the observed values (left) and the histogram of the posterior predictive p-value (right) - spatial model at district level

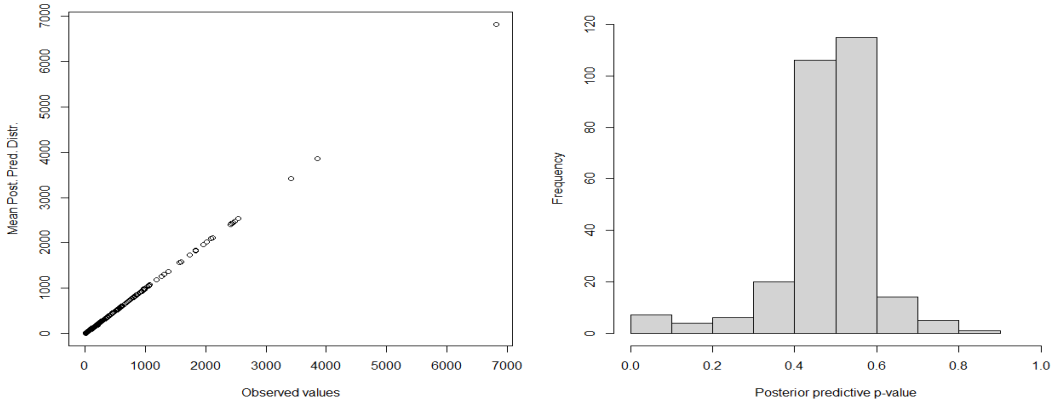

Figure 4.6: Scatter plot of the posterior mean for the predictive distributions against the observed values (left) and the histogram of the posterior predictive p-value (right) - spatial model at municipality level

As the distribution of the points in the scatter plots is similar to a straight line, it is possible to conclude that, on average, the prediction is very close to the observed values. On the other hand, there is a high number of areas with p-values close to 0.5 (in the middle of the histograms) and few areas whose p-value is very low or high. Thus, these graphs in Figures 4.5 and 4.6 suggest that the spatial model at district level and the spatial model at municipality level fit the data well.

4.2 Spatio-temporal analysis

As mentioned before, the second aim of this study is to analyse how the hospital admission rate is distributed in space and time and to identify which regressors best explain this spatial and temporal variation. As explained in Section 2.2.3.2, to investigate a spatial pattern over time, spatio-temporal models are used. In this type of models, the spatial and temporal variation and the spatial and temporal dependence between areas and years, respectively, must be included in the model. For this purpose, random effects will be incorporated in the model: v_i and ϕ_t are the unstructured effects for area and time, respectively, and u_i and γ_t are the structured effects for area and time, respectively.

4.2.1 Data at district level

Applying it to the case study, the number of hospital admissions in area i and time t is modelled according to the Poisson distribution:

$$\begin{aligned} Y_{it} &\sim \text{Poisson}(\lambda_{it}),\ i = 1,\ldots,18 \text{ and } t = 1,\ldots,5, \\ \log(\lambda_{it}) &= \eta_{it}. \end{aligned} \tag{4.7}$$

Considering that the best spatial model is the i.i.d model (see the model in Equation 4.4), the model with the unstructured temporal random effect is defined as follows:

$$\begin{aligned} &\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + v_i + \phi_t + \text{offset}_{it}, \\ &i = 1,...,18 \text{ and } t = 1,...,5, \text{ with} \end{aligned} \tag{4.8}$$

$$\begin{aligned} v_i &\sim \text{Normal}(0, \sigma_v^2), \\ \phi_t &\sim \text{Normal}(0, \sigma_\phi^2). \end{aligned}$$

For the temporal structured random effect, the model is as follows:

$$\begin{aligned} &\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + v_i + \gamma_t + \text{offset}_{it}, \\ &i = 1,...,18 \text{ and } t = 1,...,5, \text{ with} \end{aligned} \tag{4.9}$$

$$\begin{aligned} \gamma_t|\gamma_{t-1} &\sim \text{Normal}(\gamma_{t+1}, \sigma^2), \text{ if RW of order 1,} \\ \gamma_t|\gamma_{t-1}, \gamma_{t-2} &\sim \text{Normal}(2\gamma_{t+1} + \gamma_{t-2}, \sigma^2), \text{ if RW of order 2.} \end{aligned}$$

Finally, the model with the two temporal random effects is the following:

$$\begin{aligned} &\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + v_i + \phi_t + \gamma_t + \text{offset}_{it}, \\ &i = 1,...,18 \text{ and } t = 1,...,5, \end{aligned} \tag{4.10}$$

where ϕ_t and γ_t were defined in the previous models.

Thus, the priors for the hyperparameters are specified as follows:

$$\begin{aligned} b_0 &\sim \text{Normal}(0, 10^6), \\ \beta_j &\sim \text{Normal}(0, 10^6),\ j = 1,\ldots,6, \\ \log(\tau_v) &\sim \text{logGamma}(1, 0.0005), \\ \log(\tau_\phi) &\sim \text{logGamma}(1, 0.0005), \\ \log(\tau_\gamma) &\sim \text{logGamma}(1, 0.0005). \end{aligned}$$

The following table shows the process of including temporal random effects in the model chosen in the spatial analysis at district level.

Table 4.3: Process of inclusion of temporal random effects in the spatio-temporal model for the data at district level and the respective DIC.

Type of model for random effects	Model	DIC
I.I.D	$\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + v_i + \phi_t + \text{offset}_{it}$	2951.64
RW1	$\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + v_i + \gamma_t + \text{offset}_{it}$	2942.30
RW2	$\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + v_i + \gamma_t + \text{offset}_{it}$	**2934.61**
I.I.D and RW1	$\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + v_i + \phi_t + \gamma_t + \text{offset}_{it}$	2943.82
I.I.D and RW2	$\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + v_i + \phi_t + \gamma_t + \text{offset}_{it}$	2935.64

As can be seen in Table 4.3, the model with the lowest DIC is the model that takes into account that the hospital admission rate is different across districts (unstructured spatial random effect) and that there are temporal correlation between years (structured temporal random effect defined by RW2 - Equation 4.9).

Table 4.4 presents the summary of the posterior statistics for the fixed and random effects of the spatio-temporal model without interaction at district level.

Table 4.4: Posterior mean, posterior standard deviation and posterior 95% credible interval for the parameters and hyperparameters of the spatio-temporal model without interaction for the data at district level.

Parameter	Mean	Standard deviation	2.5% Quantile	97.5% Quantile
(Intercept)	-4.182	0.456	-5.087	-3.278
Population density (β_1)	1.787	0.122	1.548	2.026
Unemployment rate (β_2)	0.001	0.006	-0.010	0.012
Prop. male residents (β_3)	-0.076	0.011	-0.099	-0.054
‰ of mortality due to diabetes (β_4)	-0.050	0.006	-0.063	-0.038
Prop. members federations (β_5)	-0.039	0.004	-0.047	-0.031
Prop. residents +65 (β_6)	0.045	0.021	0.004	0.087
τ_v	0.302	0.106	0.140	0.552
τ_γ	292.383	192.181	59.70	784.231

Through the previous table, it is concluded that the variables population density, proportion of male residents, mortality rate per 1000 inhabitants due to diabetes, proportion of members in federations and proportion of residents aged 65 or over are significant to explain the variation in the hospital admission rate due to CCD at district level, accounting for the effect of spatial variation and temporal dependence. As the coefficients of population density and the proportion of residents aged 65 or over are positive, it can be concluded that regions with a high population density and a high proportion of residents aged 65 or

over tend to have a high hospital admission rate. Thus, these two variables are risk factors for the hospital admission rate due to CCD at district level over the years. In turn, since the coefficients of the proportion of male residents, the mortality rate per 1000 population due to diabetes and the proportion of members in federations are negative, it can be concluded that regions with high values of these variables tend to have low hospital admission rates. Furthermore, the spatial random effect is also significant, that is, the variation in the hospital admission rate due to CCD is different across districts, but these differences are maintained over time. In addition, the temporal random effect is also significant in explaining the admission rate, that is, there is a temporal correlation between years that is not related to the distribution of areas.

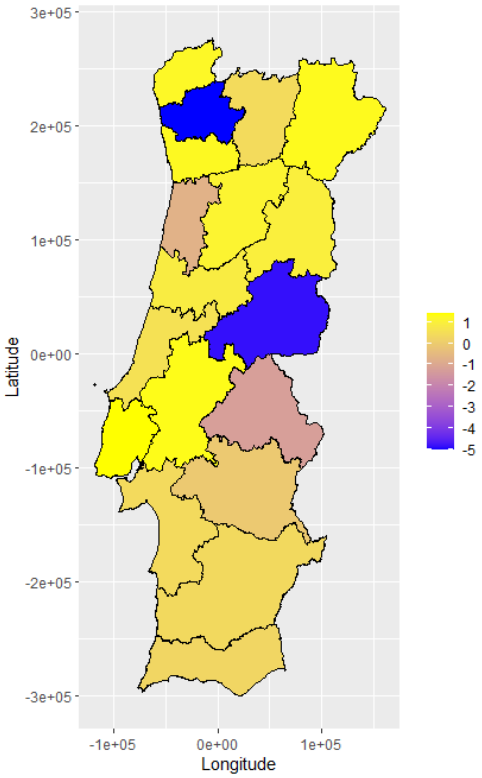

Figure 4.7: Posterior mean of the spatial random effect in the spatio-temporal model at district level.

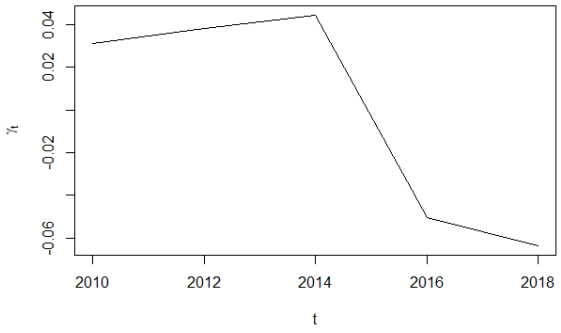

Figure 4.8: Posterior mean of the temporal random effect in the spatio-temporal model at district level.

Figure 4.7 represents the posterior mean of the unstructured spatial random effect v_i but now taking into account the insertion of the temporal structured component in the model. Through of this map it is possible to observe that there is not so much diversity of tones as in the map in Figure 4.1. These changes are justified by the insertion of the temporal structured component, which dissolves part of the variability previously explained only by the spatial component. Thus, the variability of the admission rate in some districts, which was previously explained by their location, is now explained by temporal correlation.

In turn, Figure 4.8 exposes the posterior mean of the structured temporal random effect γ_t. As all years are different from each other, it is justified to introduce the temporal random effect in the model, that is, the hospital admission rate due to CCD varies over time.

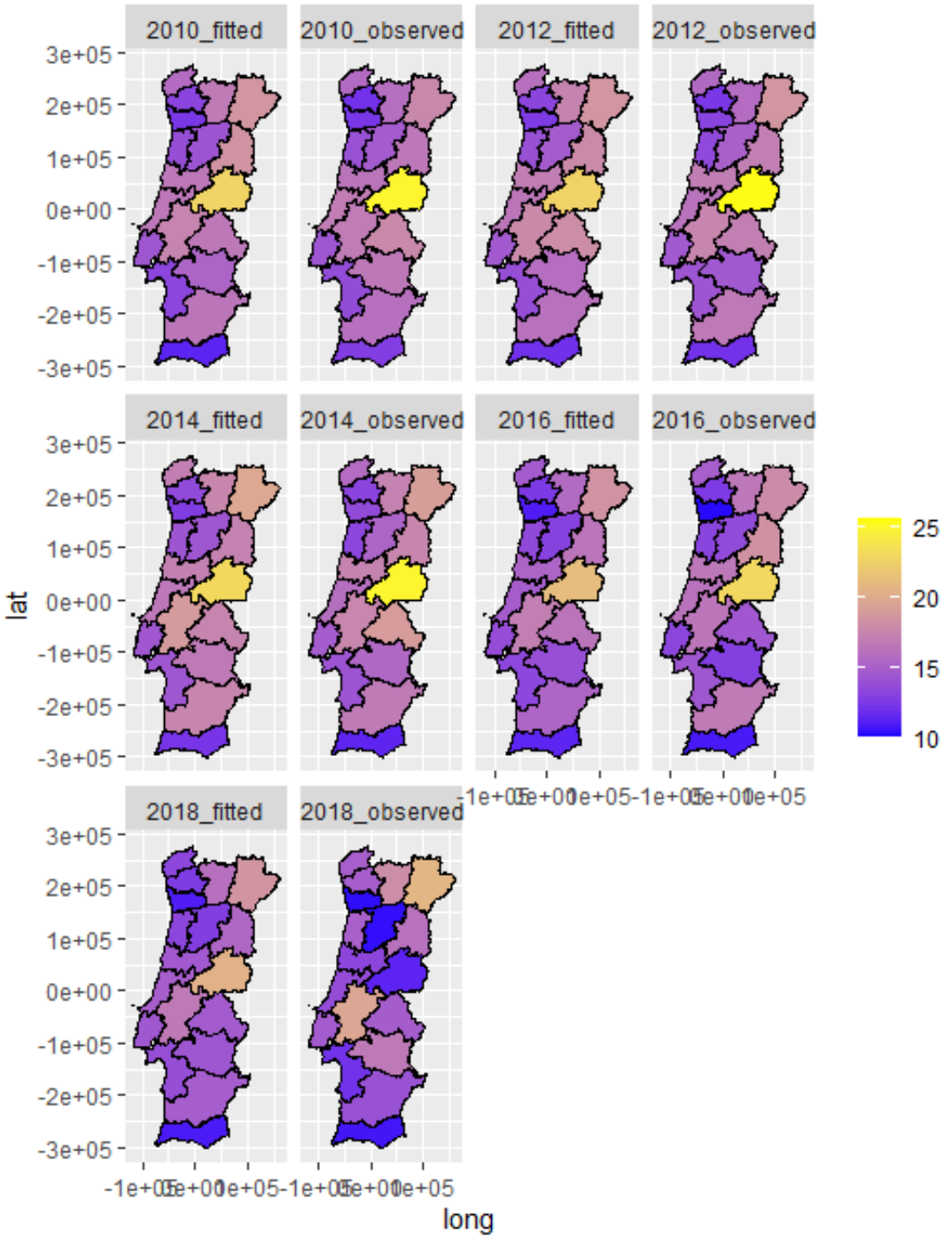

Figure 4.9: Spatial and temporal distribution of the hospital admission rate at district level - without interaction.

Figure 4.9 demonstrates the fitted values and the observed values of the hospital admission rate for all years at district level. These fitted values are the posterior mean of the hospital admission rate for each district and for each year, according to the model chosen. According to the maps of observed values, Castelo Branco was the district with the highest hospital admission rate due to CCD from 2010 to 2016. In 2016, it is possible to see a slight decrease in the hospital admission rates in the districts overall, comparing with previously years. In 2018, an increase in the hospital admission rate could again be observed in Bragança (the second district with the highest rate in 2012 and 2014 and the third district with the highest rate in 2010 and 2016), Santarém and Évora and a decrease in the rate in other districts in general. This decreased is specifically highlighted in Castelo Branco, where the hospital admission

rate decreased from 22.953 to 11.266, suffering a decline by half. In addition, Braga was the district with the lowest admission rate in 2010, Faro was the district with the lowest rate in 2012 and 2014 and Porto in 2016 and 2018. Regarding the fitted values, the maps are very similar to the maps of observed values from 2010 to 2016, so it is possible to conclude a good fit of the model to the data at district level from 2010 to 2016. However, the map of fitted values for 2018 is very different from the map of observed values. This difference is quite visible in Castelo Branco, which became the district with the lowest admission rate and in previous years was the district with the highest rate. Therefore, there is a suspicion of a poor fit of the model (with spatial and temporal components) at district level for 2018.

4.2.1.1 Space-time interactions

The models explained above can be extended through the interaction between space and time. At district level, the best spatial model is the i.i.d. model. Consequently, the best spatio-temporal model is the model whose spatial variation is defined by the unstructured random effect and whose time dependence is defined by the structured random effect, modelled as RW2. Therefore, the natural interaction between space and time is the type II interaction (explained in the Equation 2.39). According to the study data, the model with the type II interaction is described as follows:

$$\begin{aligned} &\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + v_i + \gamma_t + \delta_{it} + \text{offset}_{it}, \\ &i = 1, ..., 18 \text{ and } t = 1, ..., 5, \end{aligned} \tag{4.11}$$

where δ_{it} is defined concerning the equation mentioned in the before paragraph.

In the following table is represented the DIC for the defined model:

Table 4.5: DIC of the spatio-temporal model with interaction term for data at district level.

Interaction	Parameter interaction	DIC
II	v_i and γ_t	1121.19

The posterior mean of the parameters and hyperparameters of the model chosen, the posterior standard deviation and the posterior 2.5% and 97.5% quantiles are presented in Table 4.6.

Table 4.6: Posterior mean, posterior standard deviation and posterior 95% credible interval for the parameters and hyperparameters of the spatio-temporal model with interaction for the data at district level.

Parameter	Mean	Standard deviation	2.5% Quantile	97.5% Quantile
(Intercept)	0.148	5792.620	-11372.717	11363.523
Population density (β_1)	0.722	2.467	-4.129	5.566
Unemployment rate (β_2)	0.007	0.047	-0.086	0.100
Prop. male residents (β_3)	-0.279	0.118	-0.511	-0.048
‰ of mortality due to diabetes (β_4)	0.017	0.026	-0.034	0.067
Prop. members federations (β_5)	-0.019	0.040	-0.099	0.060
Prop. residents +65 (β_6)	-1.192	0.572	-2.315	-0.069

τ_v	67.31	0.142	67.00	67.62
τ_γ	106.61	0.292	105.90	107.19
τ_δ	178.10	0.488	176.89	179.08

From the table above, it can be concluded that the variables proportion of male residents and proportion of residents aged 65 or over are significant to explain the variation in the hospital admission rate due to CCD at district level over time, accounting for the effect of spatial variation, temporal dependence and interaction effect. Since the coefficients are negative, it can be concluded that regions with a high proportion of male residents and a high proportion of residents aged 65 or over tend to have a low hospital admission rate. In addition, the random effects are also significant for the model, that is, the spatial variation, the temporal correlation and the interaction space-time are significant to explain the variation in the hospital admission rate due to CCD at district level over time.

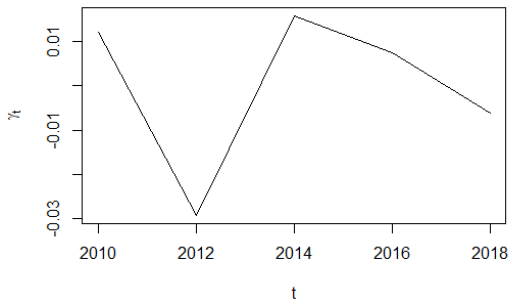

Figure 4.10: Posterior mean of the temporal random effect in the spatio-temporal model with interaction space-time at district level.

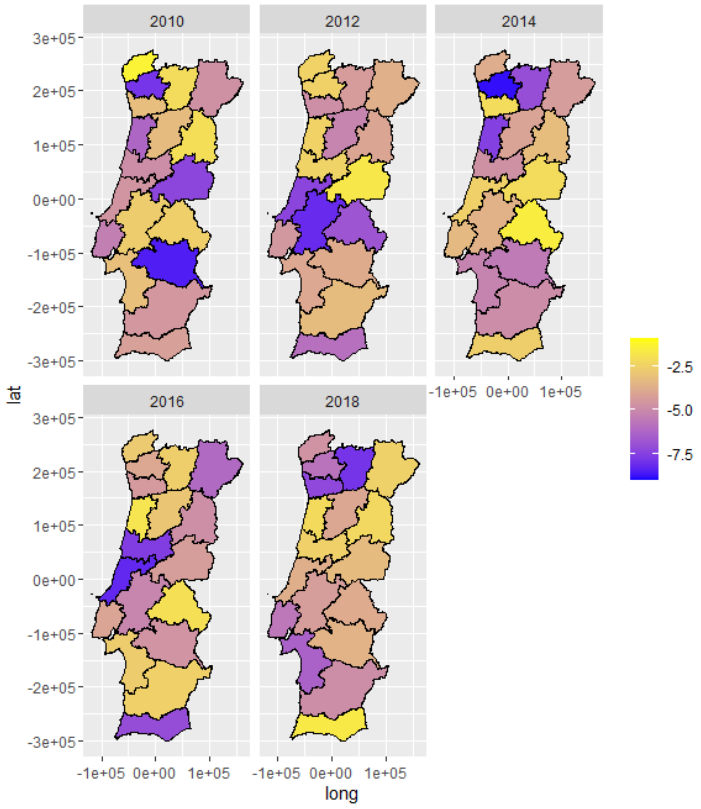

Figure 4.11: Posterior mean of the interaction random effect in the spatio-temporal model with interaction space-time at district level.

Despite being significant for the model, the posterior mean of the spatial and temporal random effects of the spatio-temporal model with interaction are almost zero. However the temporal random effect exhibits a small heterogeneity. Thus, there is a small variation in the admission rate explained only by the temporal effect. In turn, the posterior mean of the interaction random effect is represented in Figure 4.11. Since the districts are coloured with different colours in all years, the relationship between the districts varies over time, so it is justified to introduce the space-time interaction in the model. Thus, it is possible to conclude that the variation in the hospital admission rate, which is not explain by the covariates, is

mostly explain by the combined space-time effect.

The difference between this model and the spatio-temporal model without interaction is that when the interaction term is introduced in the model, the variability that was previously accommodated in the spatial and temporal effects, is now accommodated mainly in the space-time interaction. Thus, in most cases, the hospital admission rate is different across districts, and this difference varies over time.

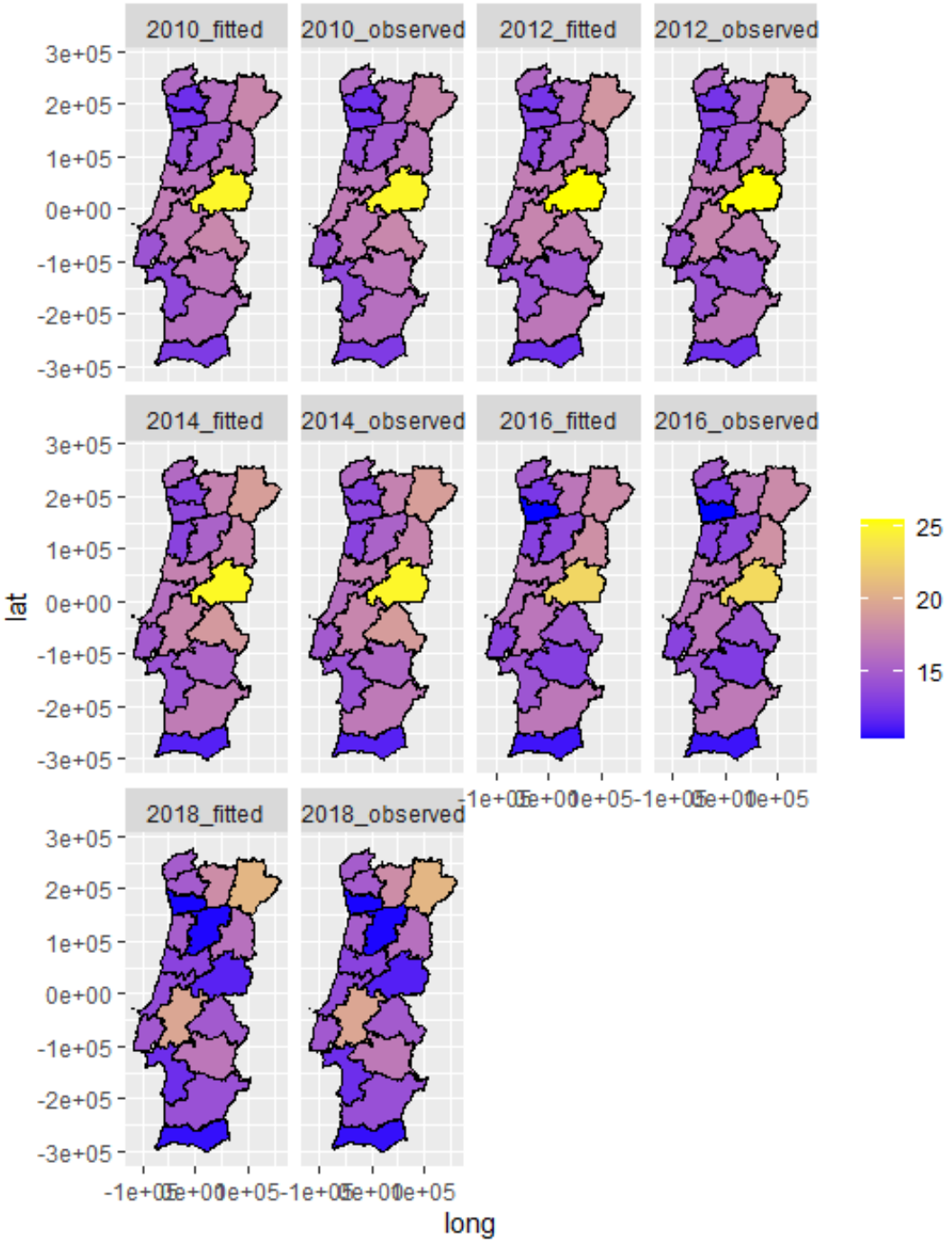

Figure 4.12: Spatial and temporal distribution of the hospital admission rate at district level - with interaction.

The Figure 4.12 shows the fitted values and the observed values of the hospital admission rate for all years at district level. These fitted values are the posterior mean of the hospital admission rate for each district and year, according to the model for the data at district level with the unstructured random effect

for area, the structured random effect for time (defined by RW2) and the interaction term between the latter two random effects. Concerning the maps of observed values, the conclusions are the same as in Figure 4.9. The maps of the fitted values are very similar with the maps of the observed values in all years, contrary to what happened in 4.9. Thus, it is possible to conclude a good fit of the model to the data at district level from 2010 to 2018.

4.2.1.2 Diagnosis

The Figures 4.13 and 4.14 present the scatter plots of the posterior mean for the predictive distributions against the observed values (left) and the histogram of the posterior predictive p-value (right) for the spatio-temporal model without interaction and for the spatio-temporal model with interaction at district level, respectively.

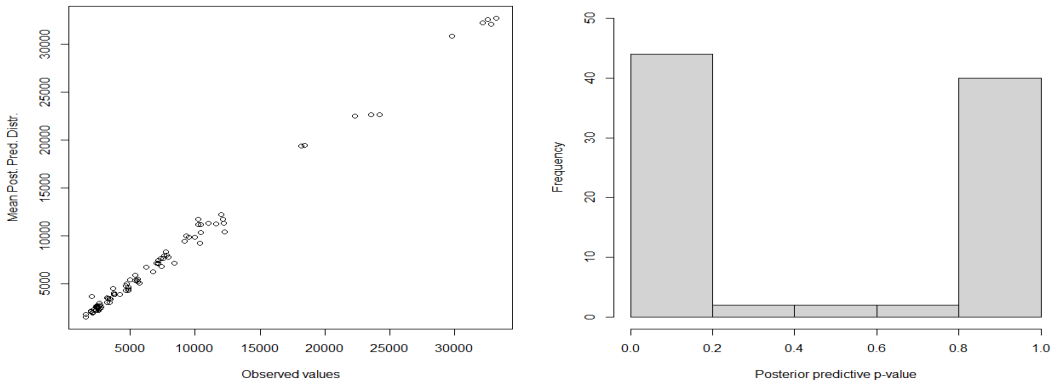

Figure 4.13: Scatter plot of the posterior mean for the predictive distributions against the observed values (left) and the histogram of the posterior predictive p-value (right) - spatio-temporal model at district level

Although the distribution of the points in Figure 4.13 looks like a straight line, this distribution is not as linear as in Figure 4.5. However, it is possible to conclude that, on average, the prediction is close to the observed values. On the other hand, by observing the histogram it is possible to see that there is a high number of areas with low and high p-values. Thus, the histogram in Figure 4.13 suggests that the spatio-temporal model without interaction at district level does not fit the data well.

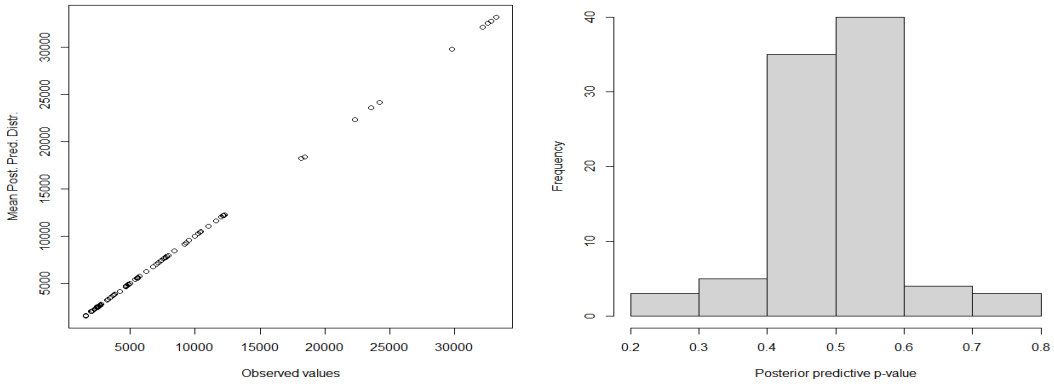

Figure 4.14: Scatter plot of the posterior mean for the predictive distributions against the observed values (left) and the histogram of the posterior predictive p-value (right) - spatio-temporal model with interaction at district level

As the distribution of the points in the scatter plot in Figure 4.14 is similar to a straight line, it is possible to conclude that, on average, the prediction is very close to the observed values. On the other hand,

there is a high number of areas with p-values close to 0.5 (in the middle of the histogram) and few areas whose p-value is very low or high. Thus, these graphs in Figure 4.14 suggest that the spatio-temporal model with interaction at district level fits the data well.

Consequently, the best spatio-temporal model, for the data at district level, is the model with interaction, defined by the Equation 4.11.

4.2.2 Data at municipality level

The number of hospital admissions in area i and year t is modelled according to the Poisson distribution:

$$\begin{aligned} Y_{it} &\sim \text{Poisson}(\lambda_{it}),\ i = 1,\ldots,278 \text{ and } t = 1,\ldots,5, \\ \log(\lambda_{it}) &= \eta_{it}. \end{aligned} \tag{4.12}$$

Since the best spatial model is the Besag model, the model with the unstructured temporal random effect is defined as follows:

$$\begin{aligned} &\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + u_i + \phi_t + \text{offset}_{it}, \\ &i = 1,...,278 \text{ and } t = 1,...,5, \text{ with} \\ &\qquad u_i|\tau_u \sim \text{iCAR}(\tau_u), \\ &\qquad \phi_t \sim \text{Normal}(0, \sigma^2_\phi). \end{aligned} \tag{4.13}$$

For the structured random effect, the model is the following:

$$\begin{aligned} &\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + u_i + \gamma_t + \text{offset}_{it}, \\ &i = 1,...,278 \text{ and } t = 1,...,5, \text{ with} \\ &\qquad \gamma_t|\gamma_{t-1} \sim \text{Normal}(\gamma_{t+1}, \sigma^2), \text{ if RW of order 1,} \\ &\qquad \gamma_t|\gamma_{t-1}, \gamma_{t-2} \sim \text{Normal}(2\gamma_{t+1} + \gamma_{t-2}, \sigma^2), \text{ if RW of order 2.} \end{aligned} \tag{4.14}$$

Finally, the model with the two temporal random effects is as follows:

$$\begin{aligned} &\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + u_i + \phi_t + \gamma_t + \text{offset}_{it}, \\ &i = 1,...,278 \text{ and } t = 1,...,5, \end{aligned} \tag{4.15}$$

where ϕ_t is the unstructured random effect and γ_t is the structured random effect, defined previously. The priors for the hyperparameters were specified in Section 4.2.1.

In the following table is represented the process of inclusion of the temporal random effects in the model chosen in the spatial analysis, with data at municipality level.

Table 4.7: Process of inclusion of random effects in the spatio-temporal model for the data at municipality level and the respective DIC.

Type of model for random effects	Model	DIC
I.I.D	$\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + u_i + \phi_t + \text{offset}_{it}$	**25148.59**
RW1	$\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + u_i + \gamma_t + \text{offset}_{it}$	25168.60
RW2	$\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + u_i + \gamma_t + \text{offset}_{it}$	25150.00
I.I.D and RW1	$\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + u_i + \phi_t + \gamma_t + \text{offset}_{it}$	25148.87
I.I.D and RW2	$\eta_{it} = b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + u_i + \phi_t + \gamma_t + \text{offset}_{i}$	25150.28

As can be seen in Table 4.7, the model with the lowest DIC is the model that takes into account the spatial correlation between municipalities (structured spatial random effect) and that the hospital admission rate is different over time (unstructured temporal random effect).

Table 4.8: Posterior mean, posterior standard deviation and posterior 95% credible interval for the parameters and hyperparameters of the spatio-temporal model without interaction for the data at municipality level.

Parameter	Mean	Standard deviation	2.5% Quantile	97.5% Quantile
(Intercept)	-4.201	0.051	-4.304	-4.099
Population density (β_5)	-0.009	0.007	-0.022	0.004
Unemployment rate (β_3)	-0.029	0.004	-0.037	-0.021
Prop. male residents (β_6)	-0.093	0.008	-0.108	-0.077
‰ of mortality due to diabetes (β_4)	0.002	0.004	-0.006	0.010
Hospitals per 1000 inhab. (β_2)	-0.012	0.006	-0.024	0.000
Prop. residents +65 (β_1)	-0.013	0.013	-0.038	0.011
τ_u	8.63	0.919	6.95	10.56
τ_ϕ	113.94	66.436	28.22	280.90

Through the Table 4.8, it is concluded that the variables unemployment rate and proportion of male residents are significant to explain the variation in the hospital admission rate due to CCD at municipality level, accounting for the effect of spatial dependence and temporal variation. Since the coefficients are negative, it can be concluded that regions with high unemployment rates and high proportions of male residents tend to have low hospital admission rates at municipality level over the years. Moreover, the spatial random effect is significant for the model, that is, there is a spatial correlation between districts that is not related to the years. In addition, the temporal random effect is also significant in explaining the admission rate, that is, the variation in the hospital admission rate due to CCD is different across years, but these differences are kept over districts.

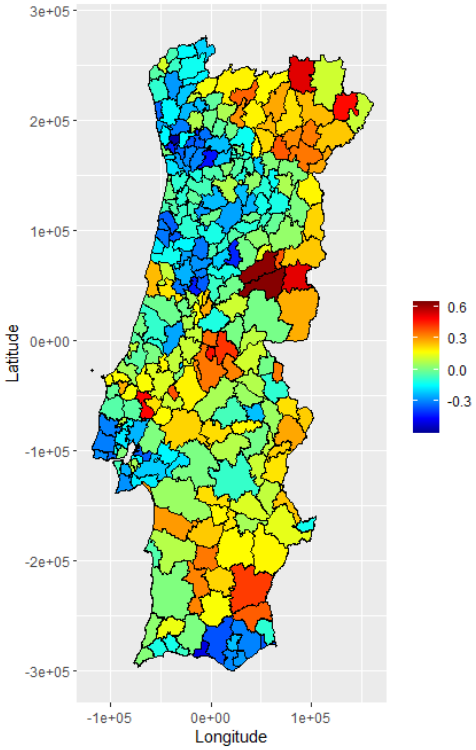

Figure 4.15: Posterior mean of the spatial random effect in the spatio-temporal model at municipality level.

Figure 4.16: Posterior mean of the temporal random effect in the spatio-temporal model at district level.

Figure 4.15 represents the posterior mean of the structured spatial random effect u_i but now taking into account the insertion of the temporal unstructured component in the model. This map shows a great diversity of colours, in contrast to the map in Figure 4.3. Thus, it is possible to conclude that a high part of the variability of the admission rate, which is not explained by the covariates, is explained by the spatial random effect.

In turn, Figure 4.16 exhibits the posterior mean of the unstructured temporal random effect ϕ_t. As all years are different from each other, it is justified to introduce the temporal random effect in the model, that is, the hospital admission rate due to CCD varies with time.

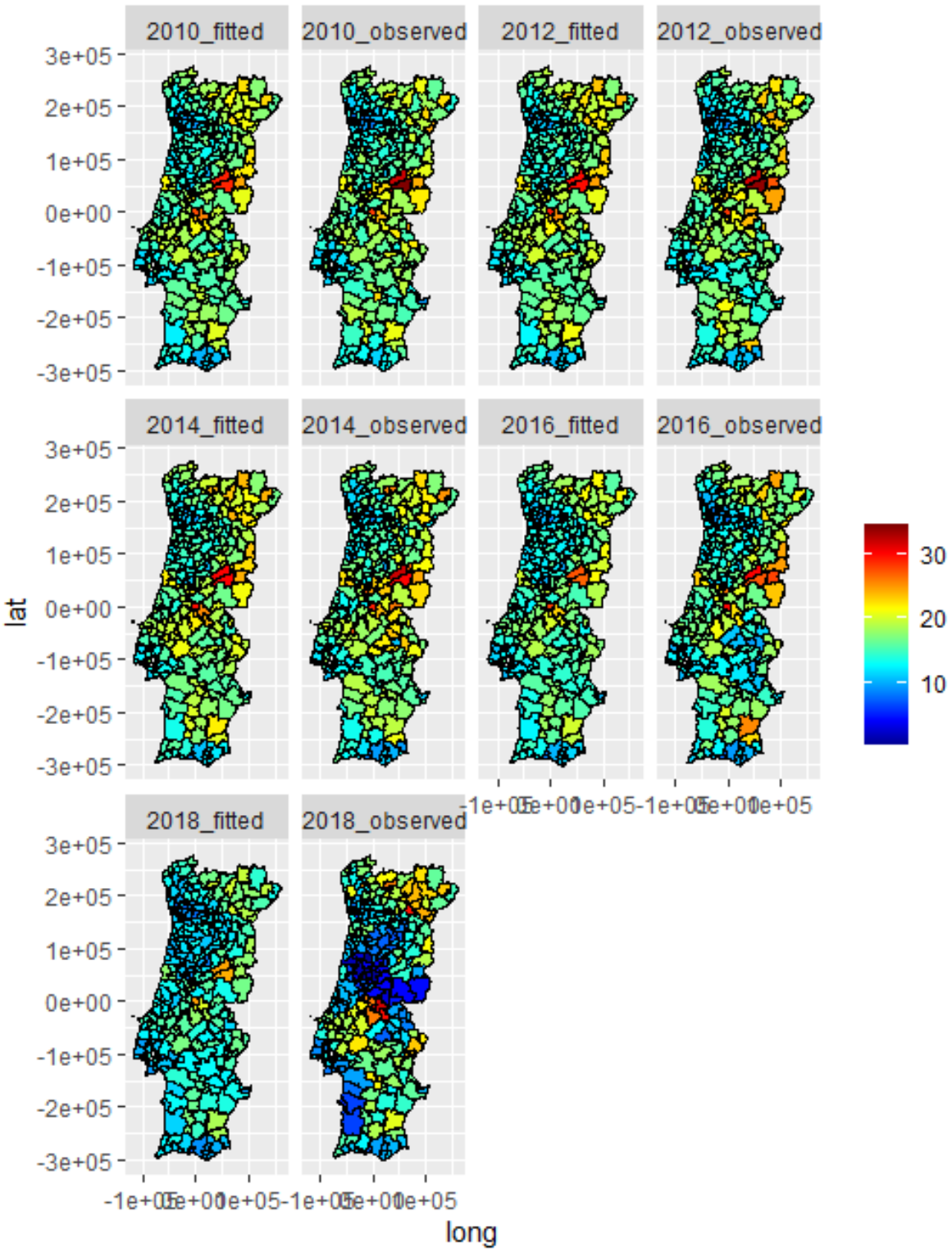

Figure 4.17: Spatial and temporal distribution of the hospital admission rate at municipality level - without interaction.

Figure 4.17 shows the fitted and observed values of the hospital admission rate for all years at municipality level. These fitted values are the posterior mean of the hospital admission rate for each municipality and for each year, according to the model chosen. In relation to the maps of observed values, Covilhã and Fundão, located in Castelo Branco, were the municipalities with the highest hospital admission rate due to CCD from 2010 to 2014. In 2016, a slight decrease in hospital admission rates was observed in most municipalities compared to previous years. However, Covilhã and Vila Real, located

in Castelo Branco, were the municipalities with the highest admission rates in 2016. In 2018, there was again an increase in hospital admission rate in some municipalities located in Bragança and Santarém, more specifically Mação, Sardoal and Gavião. Nevertheless, in general, a decrease in the rate is observed in the remaining municipalities. This decrease is very accentuated in the municipalities located in Castelo Branco and Coimbra. Regarding the fitted values, the maps are very similar to the maps of observed values from 2010 to 2016, with the fitted values being slightly lower than the observed values. Thus, it is possible to conclude a good fit of the model to the data at municipality level from 2010 to 2016. Meanwhile, the map of fitted values for 2018 is very different to the map of observed values, so there is suspicion of a poor model fit (with spatial and temporal components) at municipality level for 2018.

4.2.2.1 Space-time interactions

As seen in previous sections, the best spatial model, at municipality level, is the Besag model (model with structured random effect). Consequently, the best spatio-temporal model is the model whose spatial dependence is defined by the structured random effect and whose time variation is defined by the unstructured random effect. Therefore, the natural interaction between space and time is the type III interaction (explained in Equation 2.40). Adapting to the study data, the model with type III interaction is described as follows:

$$\begin{aligned}\eta_{it} &= b_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + u_i + \phi_t + \delta_{it} \\ &+ \text{offset}_{it}, i = 1, ..., 278 \text{ and } t = 1, ..., 5,\end{aligned} \tag{4.16}$$

where δ_{it} is defined concerning the equation mentioned in the above paragraph.

In the following table is represented the DIC for the explained model:

Table 4.9: DIC of the spatio-temporal model with interaction term for data at municipality level.

Interaction	**Parameter interaction**	**DIC**
III	u_i and ϕ_t	12914.28

The posterior mean of the parameters and hyperparameters of the model chosen, the posterior standard deviation and the posterior 2.5% and 97.5% quantiles are presented in Table 4.10.

Table 4.10: Posterior mean, posterior standard deviation and posterior 95% credible interval for the parameters and hyperparameters of the spatio-temporal model with interaction for the data at municipality level.

Parameter	**Mean**	**Standard deviation**	**2.5% Quantile**	**97.5% Quantile**
(Intercept)	0.029	2300.560	-4516.844	4512.727
Population density (β_5)	-0.001	0.0165	-0.032	0.031
Unemployment rate (β_3)	0.025	0.012	0.002	0.048
Prop. male residents (β_6)	-0.012	0.011	-0.034	0.009
‰ of mortality due to diabetes (β_4)	0.010	0.007	-0.004	0.025

Hospitals per 1000 inhab. (β_2)	-0.004	0.009	-0.021	0.013
Prop. residents +65 (β_1)	0.117	0.014	0.090	0.144
τ_u	17.458	2.454	13.542	23.110
τ_ϕ	0.394	0.477	0.001	1.660
τ_δ	15.579	0.810	14.062	17.250

As shown in Table 4.10, it is concluded that the variables unemployment rate and proportion of residents aged 65 or over are significant to explain the variation in the hospital admission rate due to CCD at municipality level, accounting for the effect of spatial dependence, temporal variation and the effect of space-time interaction. As the coefficients are positive, it can be concluded that regions with high unemployment rates and high proportions of residents aged 65 or over tend to have high hospital admission rates. Thus, these two variables are risk factors for the hospital admission rate due to CCD at municipality level over the years. Moreover, the random effects are also significant to explain the variation in the hospital admission rate due to CCD at municipality level over time.

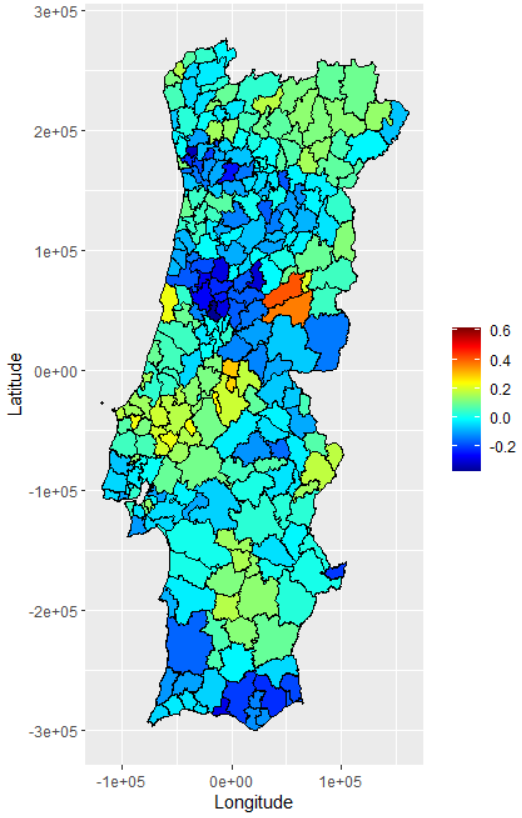

Figure 4.18: Posterior mean of the spatial random effect in the spatio-temporal model with interaction at municipality level.

Figure 4.18 represents the posterior mean of the structured spatial random effect u_i but now taking into account the insertion of the temporal unstructured component and the interaction in the model. Since the municipalities are coloured with different colours, there are a spatial correlation between the areas. Despite of the temporal effect being significant for the model, the posterior mean is almost zero, that is, there is a very small variation in the admission rate explained only by the temporal effect.

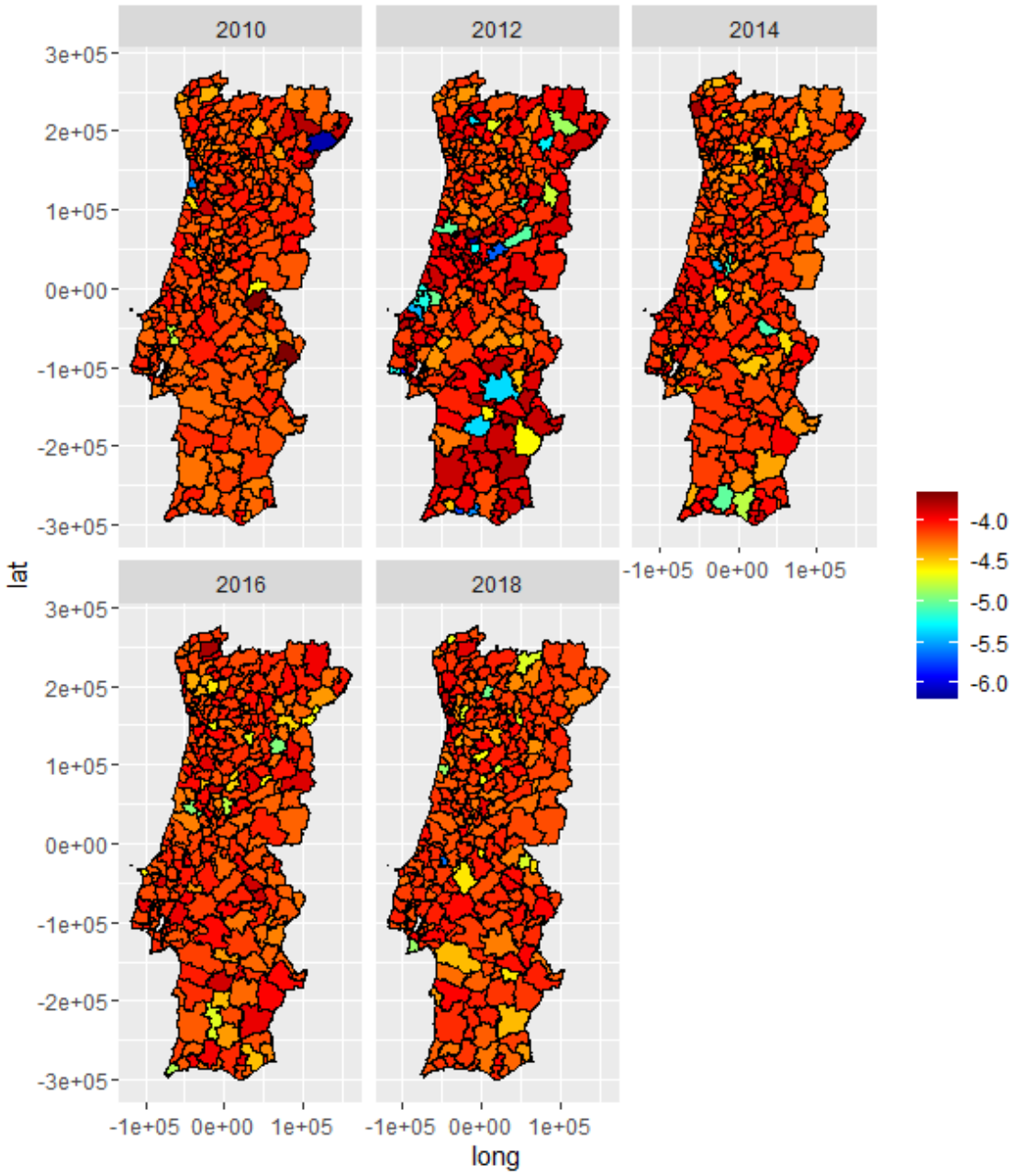

Figure 4.19: Posterior mean of the interaction random effect in the spatio-temporal model with interaction space-time at municipality level.

Figure 4.19 shows the posterior mean of the interaction random effect. Since the color of some districts varies over time, it is possible to conclude that the relationship between the districts varies with time. Therefore, it is justified to introduce the space-time interaction in the model. Thus, it is can to conclude that the variation in the hospital admission rate, which is not explain by the covariates, is explain by the spatial correlation and by space-time combined effect.

The difference between this model and the spatio-temporal model without interaction is that when the interaction term is introduced in the model, the variability that was previously accommodated only in the spatial effect and the temporal effect is now mainly accommodated in the spatial random effect (part of the variation in the admission rate is explained only by space, that is, it remains unchanged over time) and in the interaction random effect (part of the variation in the hospital admission rate is explained by

the space-time combined effect, that is, the rate is different across districts, and this difference varies over time).

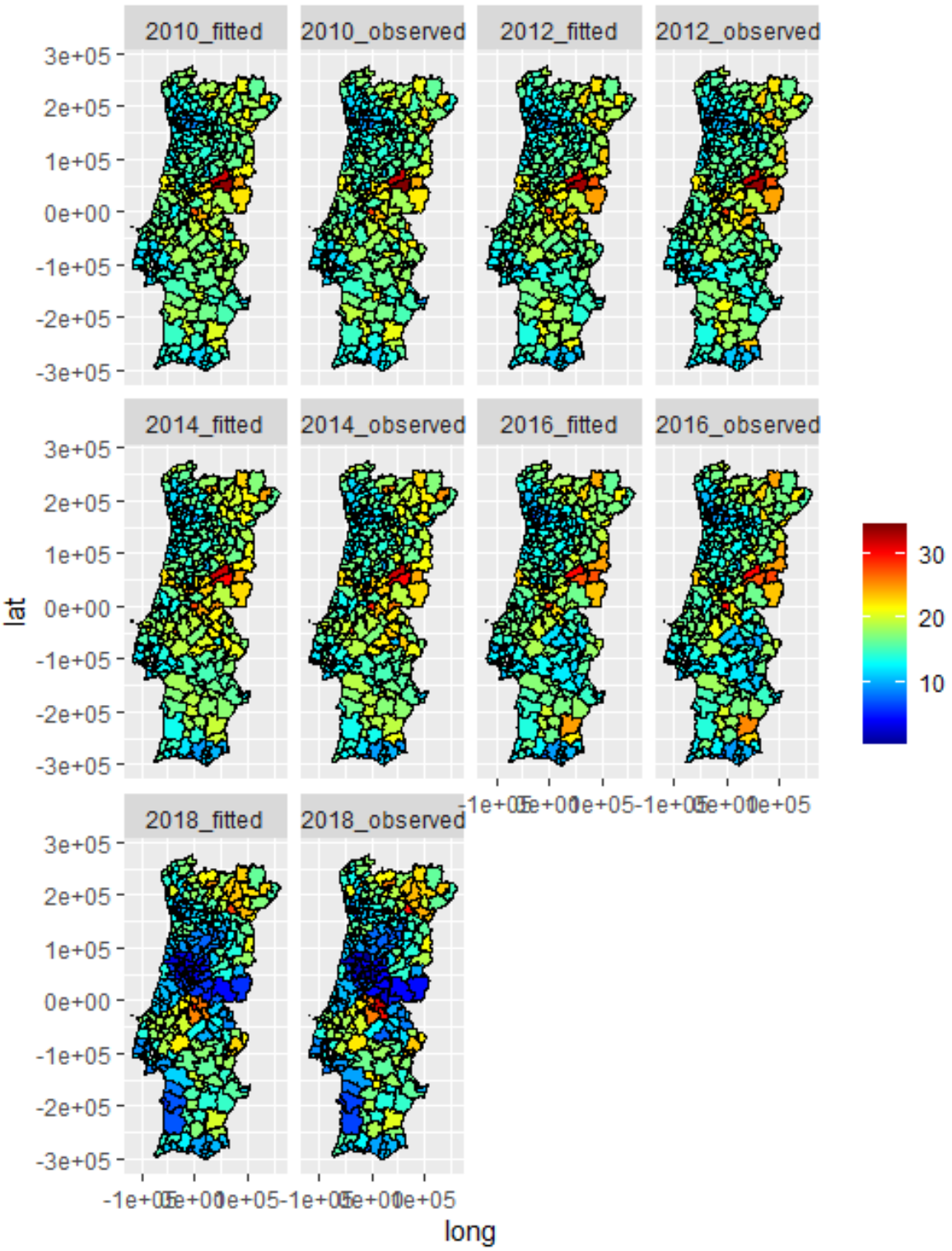

Figure 4.20: Spatial and temporal distribution of the hospital admission rate at municipality level - with interaction.

Figure 4.20 shows the observed values and the fitted values obtained from the model, for the data at municipality level, with the structured random effect for area, the unstructured random effect for time and the interaction term between the latter two random effects. According to the maps of observed values,

the conclusions are the same as the Figure 4.17. Regarding the fitted values, all maps are very similar to the observed maps, contrary to what happened in Figure 4.17. Thus, it is possible to conclude a good fit of the model to the data at municipality level from 2010 to 2018.

4.2.2.2 Diagnosis

The Figures 4.21 and 4.22 show the scatter plots of the posterior mean for the predictive distributions against the observed values (left) and the histogram of the posterior predictive p-value (right) for the spatio-temporal model without interaction and for the spatio-temporal model with interaction at municipality level, respectively.

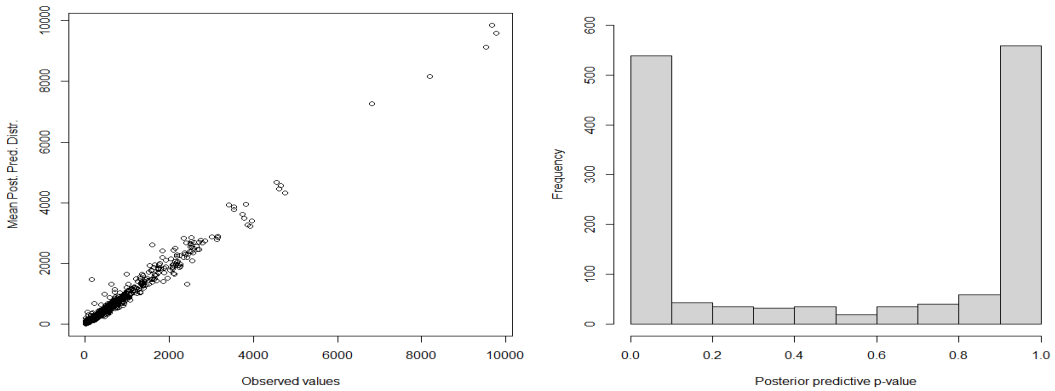

Figure 4.21: Scatter plot of the posterior mean for the predictive distributions against the observed values (left) and the histogram of the posterior predictive p-value (right) - spatio-temporal model at municipality level

Although the distribution of the points in Figure 4.21 looks like a straight line, this distribution is not as linear as in Figure 4.6. However, it is possible to conclude that, on average, the prediction is close to the observed values. On the other hand, by observing the histogram it is possible to see that there is a high number of areas with low and high p-values. Thus, the histogram in Figure 4.21 suggests that the spatio-temporal model without interaction, at municipality level, does not fit the data well.

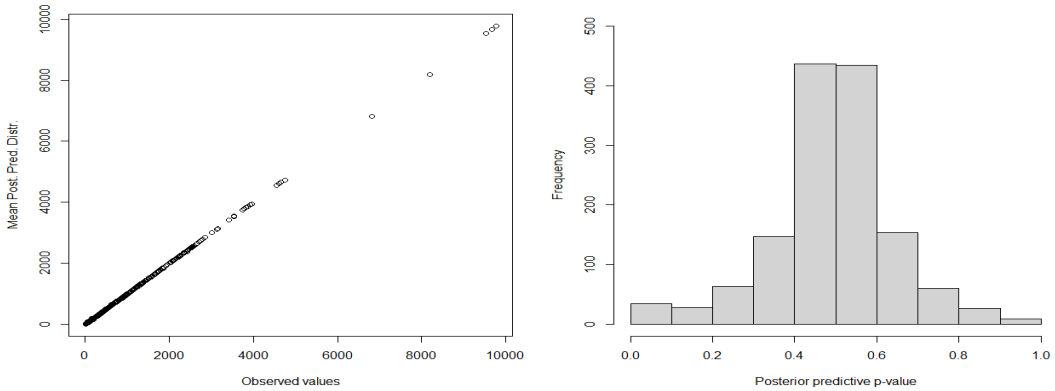

Figure 4.22: Scatter plot of the posterior mean for the predictive distributions against the observed values (left) and the histogram of the posterior predictive p-value (right) - spatio-temporal model with interaction at municipality level

As the distribution of the points in the scatter plot in Figure 4.22 is similar to a straight line, it is possible to conclude that, on average, the prediction is very close to the observed values. On the other hand, there is a high number of areas with p-values close to 0.5 (in the middle of the histogram) and few areas

whose p-value is very low or high. Thus, these graphs in Figure 4.22 suggest that the spatio-temporal model with interaction, at municipality level, fits the data well.

Consequently, the best spatio-temporal model, for the data at municipality level, is the model with interaction, defined by the Equation 4.16.

5. Discussion and main conclusions

All previous studies carried out in Portugal only analyse the mortality rate due to some cerebro-cardiovascular diseases, in particular stroke and myocardial infarction, and the associated risk factors. The DGS reports present the mortality rate caused by cerebrovascular diseases and cardiovascular diseases, separately. In addition, these reports present the spatial distribution of the percentage of admissions that arrived at the hospital through the "vias verdes" of stroke units. The Centre and Alentejo regions are the most problematic. Although these results only include strokes, it is also in these regions that the highest rate of hospital admission due to CCD is recorded (taking into account the results of this analysis).

Since cerebro-cardiovascular diseases are the main cause of death in Portugal, it is imperative to study all diseases in this group, and not only the main ones. Thus, this study aims to characterise the spatial and temporal distribution of the admissions due to CCD and their risk factors and to detect and assess the spatial and/or temporal patterns of consumption of hospital resources and identify which regressors best explain the spatial and temporal variation. For this purpose, the consumption of hospital resources was studied in the form of hospital admission rate at spatial and temporal level. Therefore, this study differs from the others because it includes all the diseases in the CCD group and analyses the spatial distribution of admissions in districts and municipalities over time. The second focus of this study was to analyse the risk factors for the hospital admission rate due to CCD.

Given the real-life scenario, the interest lies in the space-time analysis. Thus, the goal of the study is to analyse the spatio-temporal model with space-time interaction, which is the best spatio-temporal model, both at district and municipality level.

That said, taking into account the random effects plots at the district level, the hospital admission rate varies mainly across districts over time. However, a small variation in the admission rate is explained only by the temporal dependence, that is, there is still a slight dependence between the years that is maintained across districts. In terms of hospital admission rate values, Castelo Branco was the district with the highest rate from 2010 to 2016. In 2018, Bragança stood out, which is also the second district with the highest hospital admission rate in 2012 and 2014 and the third district with the highest rate in 2010 and 2016. On the other hand, Braga, Porto and Faro were the districts with the lowest hospital admission rate in 2010, 2012 and 2016. In 2014, Faro, Braga and Aveiro were the three districts with the lowest rate and in 2018 they were Porto, Viseu and Faro. At district level, the proportion of male residents and the proportion of residents aged 65 or over are the significant risk factors in explaining the variation in the hospital admission rate across space and time, accounting for the effect of spatial variation, temporal dependence and interaction effect. Since the coefficients are negative, it can be concluded that regions with a high proportion of male residents and a high proportion of older residents tend to have a low hospital admission rate. Consequently, it is suspected that women may go to hospital more frequently,

and therefore represent a higher percentage of hospital admissions. Thus, we suspect that the patient's gender may be a proxy for a certain behaviour regarding health. Regarding the proportion of residents aged 65 years or over, this result was not as expected, but this finding should be interpreted with caution because the number of observations in the model at district level is small. Furthermore, although being male and being older are risk factors for circulatory system diseases (given the literature), these results cannot be directly related, as being a risk factor for CCD is different from being a risk factor for being admitted to hospital.

However, when spatial and temporal effects and the combined effect of both are not taken into account, all variables are significant for the hospital admission rate due to CCD at district level, except the unemployment rate. According to these results, and to the results presented in the last paragraph, we conclude that it is not the values of population density, the proportion of members in federations and the mortality rate due to diabetes that directly affect the evolution of the hospital admission rate due to CCD, but rather the location of the areas characterised by these values over time. An illustrative example of the behaviour shown by these variables can be given by analysing the mortality rate due to diabetes. As this variable is no longer significant when random effects are considered, it is concluded that it is not the fact of having diabetes in a more advanced state (which could lead to death) that becomes a risk factor for the hospital admission rate due to CCD, but rather the fact that this type of diabetics live in risk areas for the hospital admission rate, for this type of disease. This situation is portrayed when considering the fact that access to healthcare is more limited in rural areas, and consequently, when patients living in these regions go to hospital, they present a more degraded state of health, which can lead to death. Following this line of thought, one is not dismissing the association between diabetes and the propensity for circulatory system diseases, but rather considering that it is not the severity of diabetes (here represented by the mortality rate) that is associated with the hospital admission rate due to CCD, but rather the patients' area of residence.

At municipality level, taking into account the random effects plots, part of the evolution of the hospital admission rate is explained by the space-time combined effect, that is, the hospital admission rate varies from municipality to municipality over time. Meanwhile, the other part of the variation in the hospital admission rate, which is not explained by the covariates, is only explained by the spatial dependence, that is, there is a dependence between municipalities is maintained over time. Regarding the rate values, Covilhã and Fundão, located in Castelo Branco, were the municipalities with the highest hospital admission rate from 2010 to 2014. In 2016, it was possible to observe a slight decrease in the hospital admission rate in most municipalities, compared to previous years. However Covilhã and Vila Real, located in Castelo Branco, were the municipalities with the highest admission rates. In 2018, there was again an increase in the hospital admission rate in some municipalities located in Bragança and Santarém. Nevertheless, a decrease in the rate was observed in other municipalities in general. This decrease was very accentuated in the municipalities located in Castelo Branco and Coimbra. In addition, the unemployment rate and the proportion of residents aged 65 or over are the significant variables to explain the evolution of the admission rate in space and time, accounting for the effect of spatial dependence, temporal variation and interaction effect. Since the coefficients are positive, it can be stated that regions with high unemployment rates and high proportions of residents aged 65 or over tend to have high hospital admission rates. This finding is in accordance with the facts found in the literature, as being older is a risk factor for circulatory system diseases. On the other hand, being unemployed can cause stress in an individual, which becomes a risk factor for CCD, as suspected in the literature. Thus, it leads us to believe that these variables are

considered risk factors for both hospital admission rate and propensity for CCD, however, these results cannot be fully compared since the response variables are different, as explained for the districts.

Moreover, without considering spatial and temporal effects and the interaction between the two, all variables are significant for the hospital admission rate due to CCD. Hence, according to these results, it is not the values of the proportion of male residents, the population density, the number of hospitals and the mortality rate due to diabetes in each area that directly affect the evolution of the hospital admission rate, but rather the geographical location of the areas characterised by these values over time.

In summary, it is not possible to directly compare this study with the other studies presented in the literature because this one analyses the risk factors for the hospital admission rate due to CCD and the others analyse the propensity for CCD. Additionally, for variables that are not significant when random effects are taken into account, the models indicate that differences in hospital admission rates due to CCD stem from the location of the regions over time, rather than from population characteristics.

Finally, the models were built given the limitations of the data, essentially present in odd years. The sample size of odd years is half the sample size of even years, which makes us doubt the credibility of the data. Although to the odd years have been excluded to the study data, these results should not be taken as absolute certainty because of all the limitations of the data. In addition, there are important risk factors that were not included in the models (e.g. smoking and alcoholism) as they are not available at the intended area level (district and municipality).

In the future, I would like to replicate this study, but using a database with patient health-level characteristics to profile the population admitted to hospitals due to CCD, in order to identify the risk factors associated with these patients, so as to control this disease and avoid hospital admissions. Furthermore, I would like to divide this group of diseases into two subgroups: one group composed of the diseases with high hospital admission (e.g. stroke) and another group with diseases with low probability of taking an individual to hospital, that is, more controlled diseases. Thus, I intended to see if the results would be different, taking into account that these two groups are very different in relation to the health status of individuals and the associated risk.

Bibliography

Banerjee, Sudipto et al. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall.

Bermudez, Patricia (2021). "Apontamentos da Unidade Curricular Estatística Bayesiana".

Besag, Julian et al. (1991). "Bayesian image restoration, with two applications in spatial statistics". In: *Annals of the Institute of Statistical Mathematics* 43, pp. 1–59.

Blangiardo, Marta and Michela Cameletti (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*.

Blangiardo, Marta, Michela Cameletti, et al. (n.d.). *A tutorial in spatial and spatio-temporal models with R-INLA*. Tech. rep.

Bourbon, Mafalda et al. (n.d.). *Doenças Cardiovasculares*.

Buuren, Stef van (2007). "Multiple imputation of discrete and continuous data by fully conditional specification". In: *Statistical Methods in Medical Research* 16.3, pp. 219–242.

Buuren, Stef van and Karin Groothuis-Oudshoorn (2011). "mice: Multivariate imputation by chained equations in R". In: *Journal of Statistical Software* 45.3.

Cadima, Jorge (2015). "Apontamentos da Unidade Curricular Modelos Matemáticos e Aplicações".

Cressie, Noel (1993). *Statistics for Spatial Data*. Wiley.

Dobson, Annette J. (2002). *An introduction to generalized linear models*.

Ferreira, Rui Cruz et al. (2016). *Doenças Cérebro-Cardiovasculares em Números – 2015*. Tech. rep.

Fox, John and Sanford Weisberg (2015). "Mixed Effects Models in R". In: *An R Companion to Applied Regression*.

Gelfand, Alan et al. (2010). *Handbook of Spatial Statistics*.

Gulbenkian Descobrir and Maratona da Saúde (2016). *Dossiê Doenças Cardiovasculares*.

Juárez, M. A. (2018). "Appointments of the Curriculum Unit Bayesian Statistics".

Katitas, Aycan (2019). *Getting Started with Multiple Imputation in R*.

Kleinman, Ken et al. (2004). "A Generalized Linear Mixed Models Approach for Detecting Incident Clusters of Disease in Small Areas, with an Application to Biological Terrorism". In: *American Journal of Epidemiology* 159.3, pp. 217–224.

Kleinschmidt, I. et al. (2001). *Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in KwaZulu Natal, South Africa*. Tech. rep. 12, pp. 1213–1221.

Lawson, Andrew B. (2008). *Bayesian Disease Mapping : Hierarchical Modeling in Spatial Epidemiology*.

Lawson, Andrew B. and Fiona L. R. Williams (2001). *An Introductory Guide to Disease Mapping*. Vol. 10. 6.

Lindgren, Finn and Håvard Rue (2015). "Bayesian spatial modelling with R-INLA". In: *Journal of Statistical Software* 63.19.

Médis (2018). *Como prevenir Doenças Cardiovasculares?*

Oliveira, Antônio Neco de (2018). "Modelação por regressão incorporando dependência espacial e temporal". PhD thesis. Escola de Ciências da Universidade do Minho.

BIBLIOGRAPHY

Pires, Vera Lúcia Alonso (2018). "A Codificação Clínica e os Problemas Associados à Qualidade dos Dados: Perspetiva dos Codificadores". PhD thesis. Faculdade de Medicina da Universidade do Porto.

Rubin, Donald B. (1976). *Inference and missing data*.

Stuart, Elizabeth A. et al. (2009). "Multiple imputation with large data sets: A case study of the children's mental health initiative". In: *American Journal of Epidemiology* 169.9, pp. 1133–1139.

Turkman, M. Antónia Amaral and Carlos Daniel Paulino (2015). *Estatística Bayesiana Computacional - uma introdução*. Sociedade Portuguesa de Estatística.

Turkman, M. Antônia Amaral and Giovani Loiola Silva (2000). "Modelos Lineares Generalizados: da teoria à prática." In: *VIII Congresso Anual da Sociedade Portuguesa de Estatística*.

Zuur, Alain F. et al. (2009). *Mixed Effects Models and Extensions in Ecology with R*.

Appendices

A Tables of exploratory analysis

Table 1: Descriptive values of the variables per year.

Years	‰ of hospital admissions	Population density	% of unemployment	Proportion of male residents	‰ of mortality due to diabetes	Proportion of members in federations	Proportion of residents aged 65 or over	Hospitals per 1000 inhab.
2010	14.370	112.900	7.991	0.478	0.451	0.048	0.190	0.021
2012	14.596	112.300	9.602	0.476	0.472	0.049	0.197	0.021
2014	14.818	111.000	9.343	0.474	0.433	0.051	0.206	0.021
2016	13.521	110.300	7.738	0.473	0.423	0.056	0.214	0.021
2018	13.574	109.800	5.307	0.472	0.429	0.063	0.222	0.022

Table 2: Descriptive values of the study variables per district in 2010.

Districts	‰ of hospital admissions	Population density	% of unemployment	Proportion of male residents	‰ of mortality due to diabetes	Proportion of members in federations	Proportion of residents aged 65 or over
Aveiro	13.941	255.213	8.051	0.481	0.443	0.056	0.173
Beja	15.991	14.916	8.159	0.492	0.705	0.045	0.248
Braga	12.061	313.630	9.161	0.482	0.257	0.040	0.140
Bragança	17.571	20.700	8.330	0.482	0.681	0.030	0.282
Castelo Branco	24.823	29.694	8.373	0.477	0.595	0.031	0.269
Coimbra	16.305	108.441	6.402	0.473	0.388	0.049	0.221
Évora	16.539	22.587	7.054	0.482	0.976	0.047	0.240
Faro	12.739	90.319	8.555	0.488	0.348	0.059	0.191
Guarda	16.500	29.157	7.302	0.474	0.558	0.051	0.275
Leiria	16.880	134.521	6.384	0.481	0.543	0.048	0.199
Lisboa	14.290	803.183	6.705	0.472	0.422	0.044	0.189
Portalegre	17.643	19.535	8.844	0.479	1.001	0.053	0.265
Porto	12.265	780.004	10.404	0.478	0.346	0.056	0.152
Santarém	16.777	67.642	6.630	0.478	0.744	0.044	0.227
Setúbal	13.610	163.384	7.343	0.481	0.433	0.048	0.179
Viana do Castelo	15.621	110.557	6.843	0.467	0.465	0.050	0.225
Vila Real	15.929	48.081	8.956	0.478	0.604	0.041	0.236
Viseu	14.581	75.499	7.812	0.477	0.545	0.047	0.221

Table 3: Descriptive values of the study variables per district in 2012.

Districts	‰ of hospital admissions	Population density	% of unemployment	Proportion of male residents	‰ of mortality due to diabetes	Proportion of members in federations	Proportion of residents aged 65 or over
Aveiro	13.423	253.467	8.748	0.479	0.490	0.059	0.181
Beja	16.615	14.666	10.107	0.491	0.857	0.044	0.246
Braga	12.351	312.797	10.627	0.481	0.326	0.043	0.147
Bragança	18.562	20.207	9.714	0.481	0.765	0.031	0.285
Castelo Branco	25.309	28.927	9.603	0.475	0.699	0.032	0.272
Coimbra	16.930	106.352	8.375	0.471	0.412	0.047	0.228
Évora	14.471	22.179	9.100	0.480	0.890	0.048	0.244
Faro	12.064	88.935	10.707	0.484	0.358	0.059	0.198
Guarda	17.034	28.306	8.548	0.472	0.715	0.059	0.279
Leiria	16.211	133.306	7.835	0.479	0.663	0.053	0.206
Lisboa	14.503	800.872	8.321	0.471	0.420	0.046	0.198
Portalegre	17.050	19.010	9.990	0.478	0.856	0.042	0.265
Porto	13.039	775.434	12.298	0.476	0.350	0.054	0.161
Santarém	17.471	66.880	8.553	0.476	0.774	0.045	0.232
Setúbal	14.225	163.809	8.965	0.479	0.464	0.047	0.188
Viana do Castelo	15.614	109.137	8.304	0.466	0.500	0.046	0.229
Vila Real	15.741	47.180	10.080	0.477	0.566	0.039	0.240
Viseu	15.009	74.327	9.434	0.475	0.454	0.051	0.225

Table 4: Descriptive values of the study variables per district in 2014.

Districts	‰ of hospital admissions	Population density	% of unemployment	Proportion of male residents	‰ of mortality due to diabetes	Proportion of members in federations	Proportion of residents aged 65 or over
Aveiro	13.227	250.965	8.138	0.476	0.437	0.065	0.191
Beja	16.702	14.351	9.410	0.488	0.726	0.043	0.249
Braga	13.139	310.276	9.514	0.479	0.295	0.047	0.157
Bragança	19.097	19.672	9.735	0.478	0.686	0.031	0.292
Castelo Branco	24.819	28.190	9.408	0.475	0.674	0.033	0.277
Coimbra	17.281	104.341	8.440	0.470	0.379	0.052	0.239
Évora	15.451	21.622	9.036	0.478	0.732	0.058	0.249
Faro	11.346	88.351	9.164	0.482	0.392	0.061	0.206
Guarda	17.463	27.508	7.846	0.470	0.782	0.076	0.286
Leiria	15.916	131.799	7.040	0.477	0.630	0.061	0.214
Lisboa	14.685	793.769	8.302	0.470	0.368	0.046	0.207
Portalegre	18.968	18.422	9.638	0.476	1.142	0.041	0.268
Porto	13.536	767.211	12.599	0.474	0.282	0.053	0.173
Santarém	17.533	65.759	7.468	0.475	0.634	0.046	0.239
Setúbal	14.130	163.500	8.792	0.477	0.482	0.046	0.199
Viana do Castelo	15.748	107.261	7.959	0.464	0.387	0.045	0.237
Vila Real	17.233	46.113	10.578	0.475	0.604	0.047	0.248
Viseu	15.207	73.009	10.104	0.474	0.506	0.066	0.233

Table 5: Descriptive values of the study variables per district in 2016.

Districts	‰ of hospital admissions	Population density	% of unemployment	Proportion of male residents	‰ of mortality due to diabetes	Proportion of members in federations	Proportion of residents aged 65 or over
Aveiro	13.113	249.527	6.487	0.476	0.298	0.067	0.201
Beja	16.708	14.025	9.025	0.487	0.827	0.052	0.251
Braga	12.543	307.675	7.167	0.478	0.258	0.045	0.168
Bragança	17.831	19.166	9.460	0.477	0.704	0.032	0.300
Castelo Branco	22.953	27.630	7.968	0.474	0.612	0.038	0.283
Coimbra	16.513	103.263	7.109	0.470	0.405	0.060	0.247
Évora	12.836	21.128	7.808	0.478	1.043	0.066	0.255
Faro	10.787	88.351	6.936	0.479	0.340	0.076	0.211
Guarda	18.273	26.832	6.788	0.470	0.660	0.072	0.290
Leiria	16.203	131.045	5.369	0.477	0.629	0.071	0.220
Lisboa	13.247	798.242	6.949	0.469	0.372	0.047	0.215
Portalegre	14.357	17.847	9.030	0.476	0.884	0.053	0.270
Porto	10.247	761.715	10.483	0.473	0.303	0.062	0.185
Santarém	16.399	64.843	6.099	0.474	0.601	0.052	0.244
Setúbal	14.348	163.295	7.532	0.475	0.504	0.048	0.207
Viana do Castelo	14.858	105.376	6.292	0.464	0.475	0.056	0.245
Vila Real	16.507	45.130	9.892	0.474	0.504	0.066	0.258
Viseu	13.532	71.968	8.576	0.473	0.485	0.069	0.240

Table 6: Descriptive values of the study variables per district in 2018.

Districts	‰ of hospital admissions	Population density	% of unemployment	Proportion of male residents	‰ of mortality due to diabetes	Proportion of members in federations	Proportion of residents aged 65 or over
Aveiro	14.929	248.381	4.368	0.474	0.335	0.071	0.211
Beja	14.011	13.756	6.492	0.486	0.857	0.064	0.254
Braga	14.804	306.211	4.880	0.476	0.273	0.047	0.178
Bragança	20.647	18.879	6.782	0.473	0.779	0.035	0.306
Castelo Branco	11.266	27.014	5.571	0.473	0.681	0.044	0.290
Coimbra	13.233	101.987	4.814	0.470	0.442	0.068	0.255
Évora	16.649	20.676	4.894	0.477	0.864	0.072	0.261
Faro	10.664	87.830	4.733	0.478	0.335	0.089	0.216
Guarda	16.072	26.079	5.138	0.469	0.845	0.077	0.294
Leiria	13.698	129.670	3.558	0.476	0.576	0.072	0.228
Lisboa	14.612	806.706	4.635	0.468	0.386	0.055	0.220
Portalegre	14.733	17.336	6.521	0.475	1.195	0.068	0.274
Porto	10.356	762.609	7.183	0.471	0.312	0.071	0.195
Santarém	19.599	63.962	4.195	0.473	0.624	0.057	0.249
Setúbal	12.031	163.467	5.306	0.473	0.424	0.051	0.214
Viana do Castelo	14.838	104.087	3.546	0.462	0.476	0.089	0.251
Vila Real	17.973	44.549	7.836	0.471	0.479	0.073	0.266
Viseu	10.382	70.752	6.440	0.472	0.454	0.077	0.248

B Explanatory analysis - scatter plots and maps

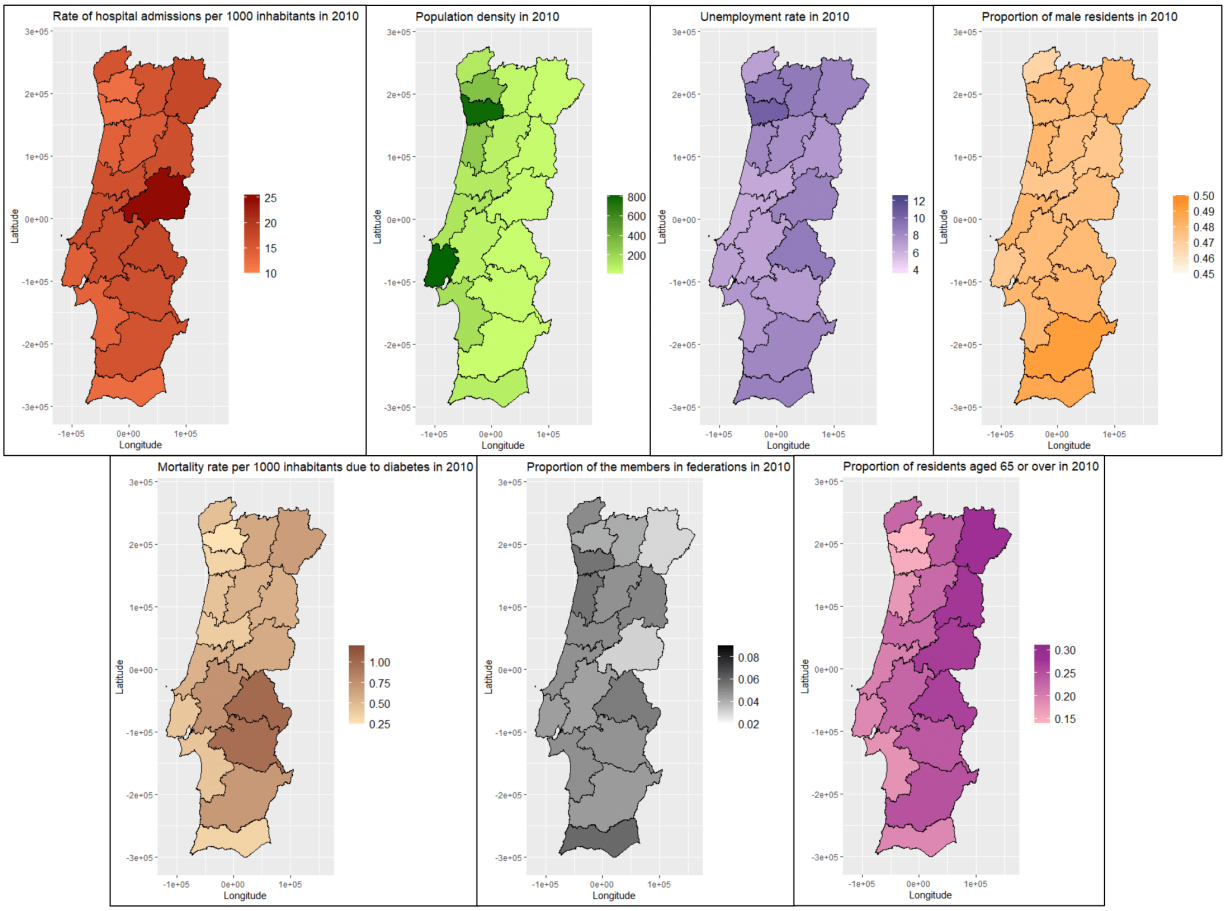

Figure 1: Maps at district level of the hospital admission rate and the variables from INE in 2010.

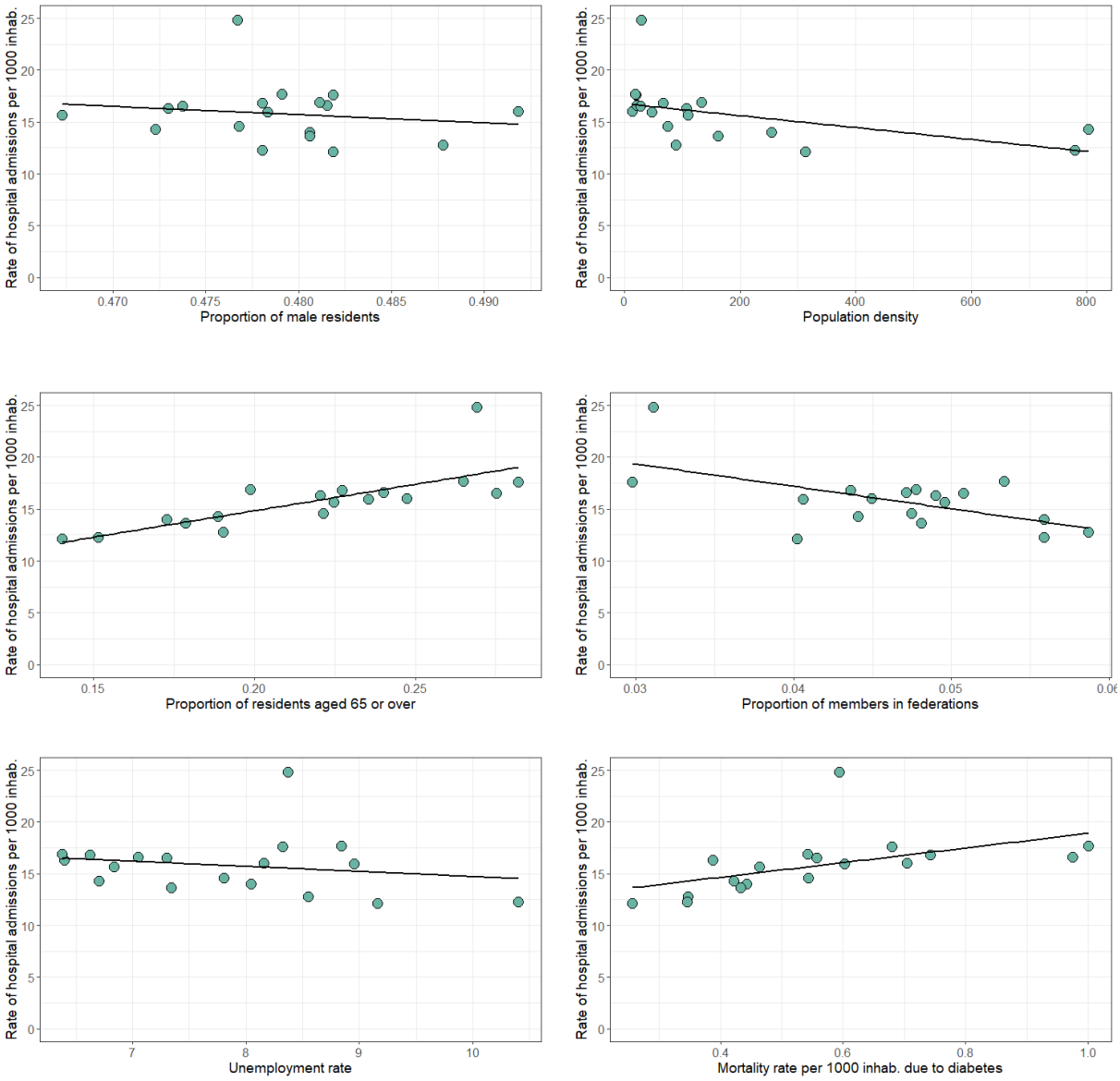

Figure 2: Scatter plot of the hospital admission rate vs all the variables at district level for 2010 and the respective regression line of the GLM model.

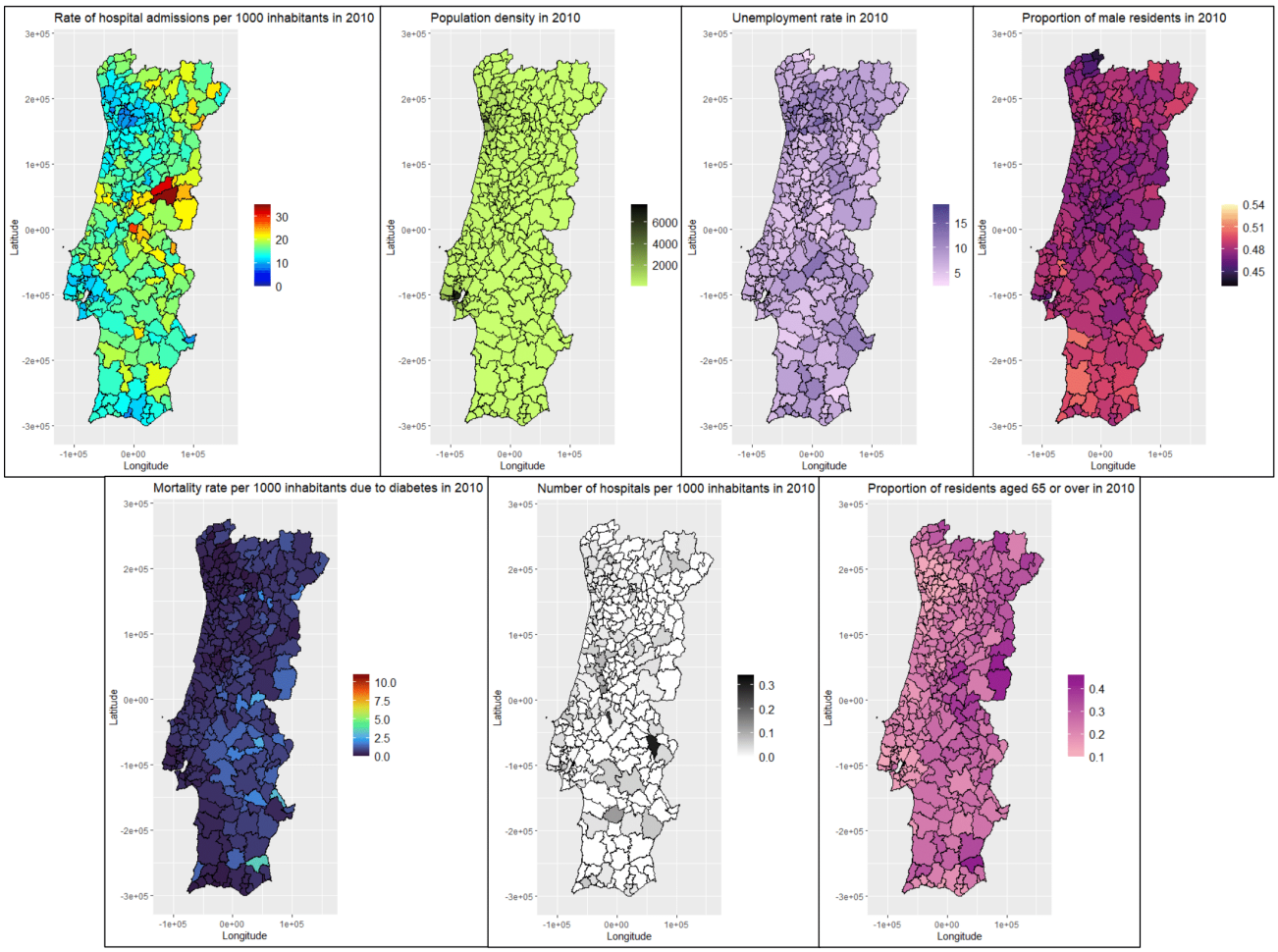

Figure 3: Maps at municipality level of the hospital admission rate and the variables from INE in 2010.

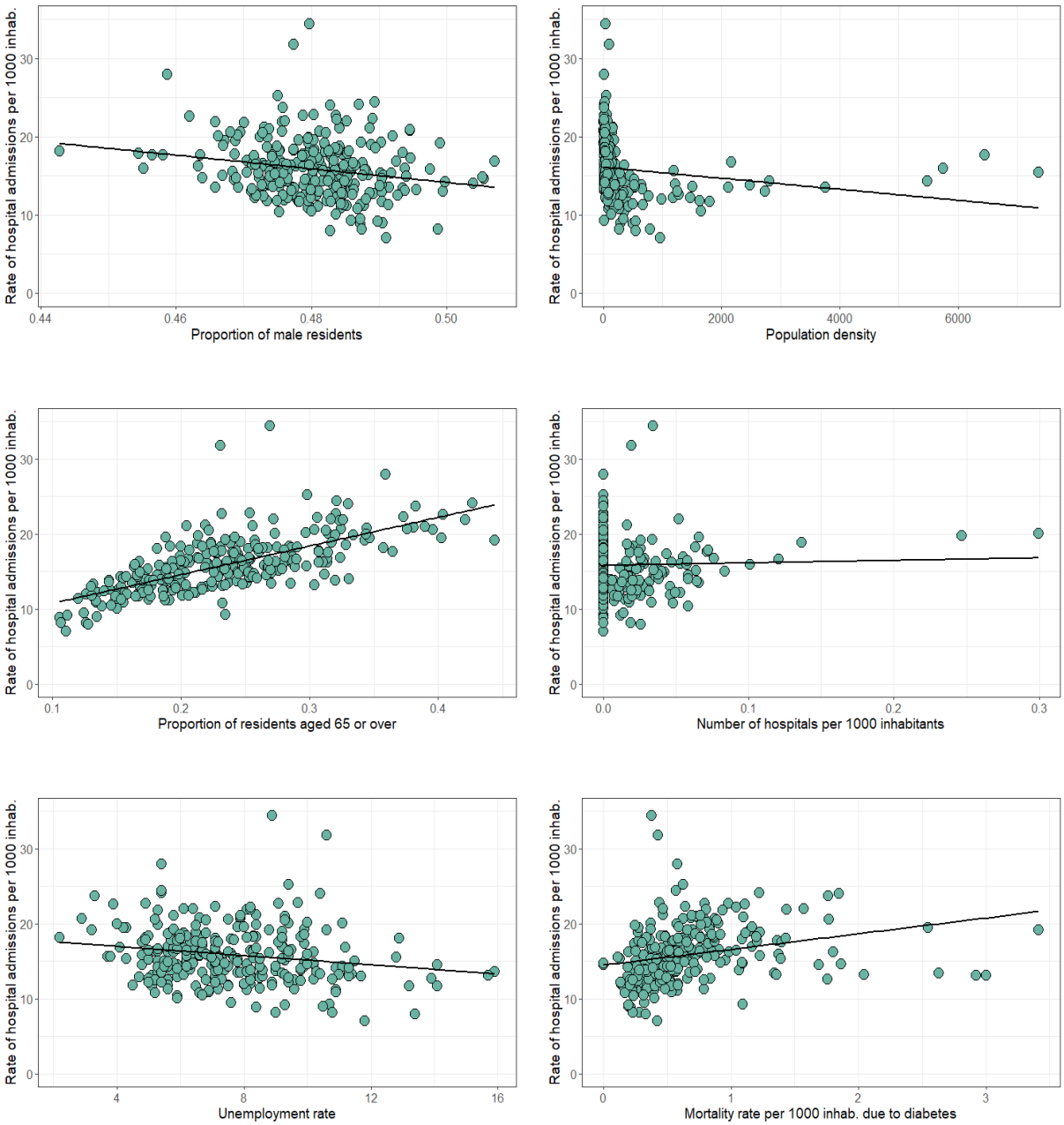

Figure 4: Scatter plot of the hospital admission rate vs all the variables at municipality level for 2010 and the respective regression line of the GLM model.

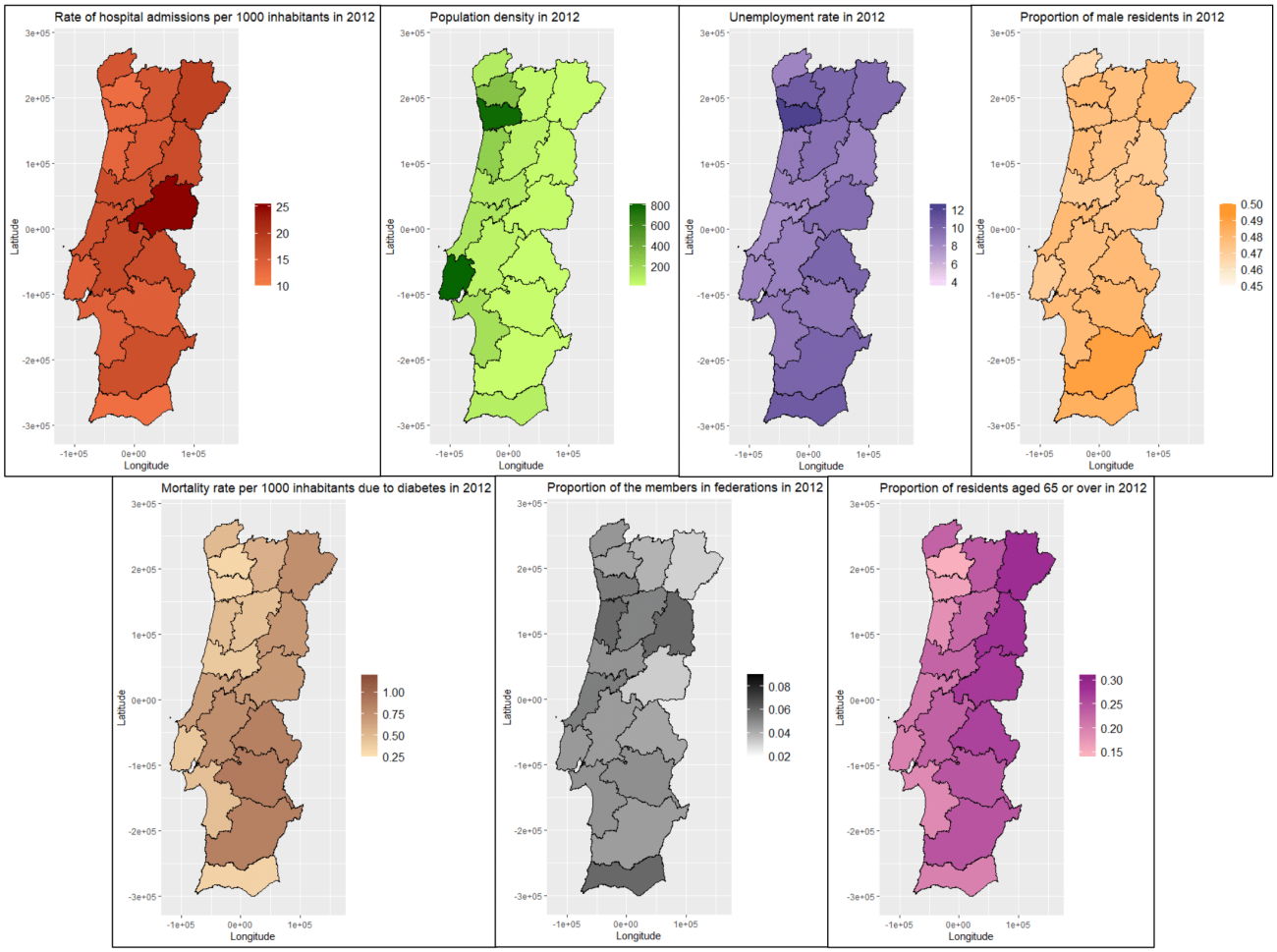

Figure 5: Maps at district level of the hospital admission rate and the variables from INE in 2012.

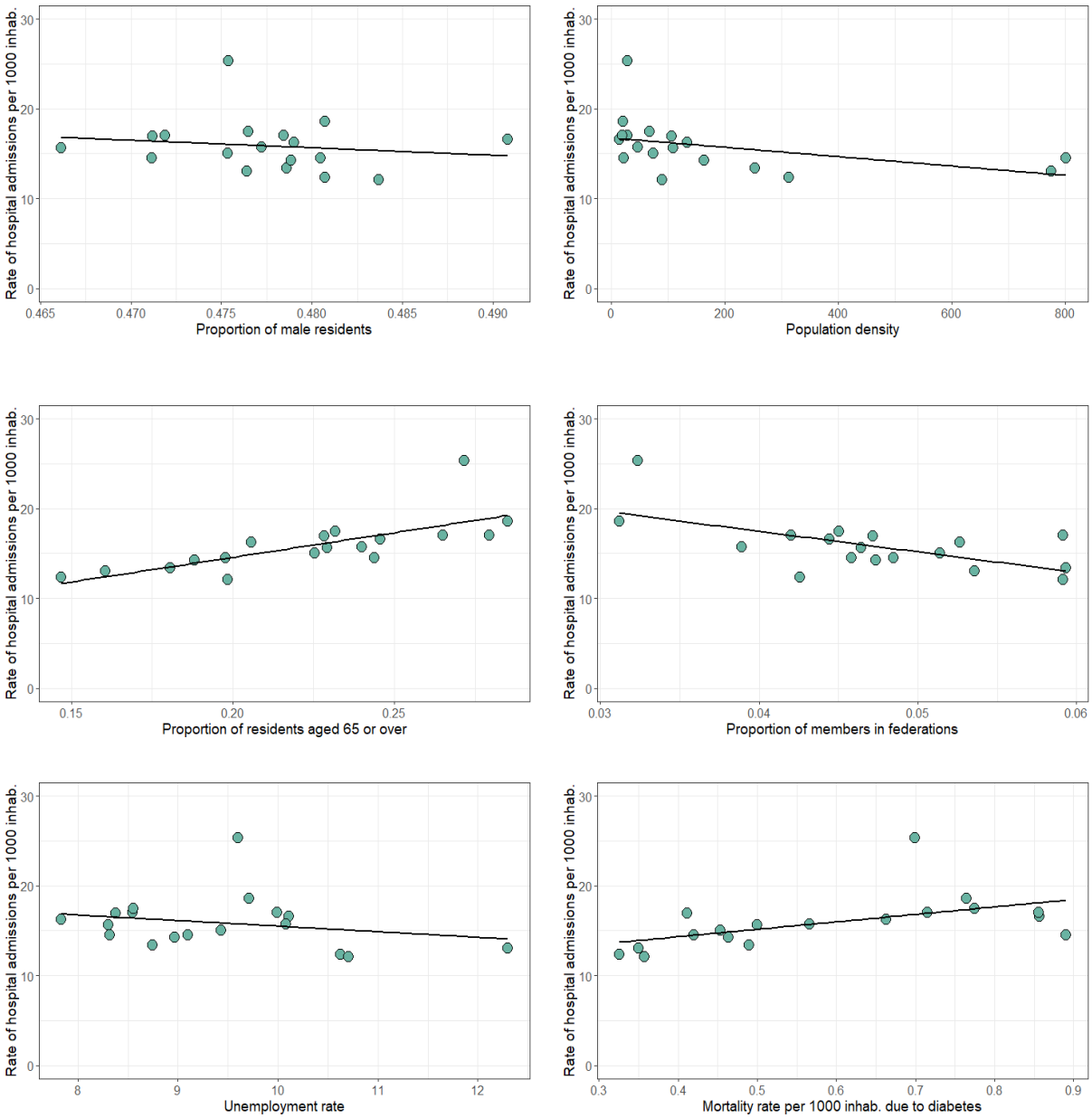

Figure 6: Scatter plot of the hospital admission rate vs all the variables at district level for 2012 and the respective regression line of the GLM model.

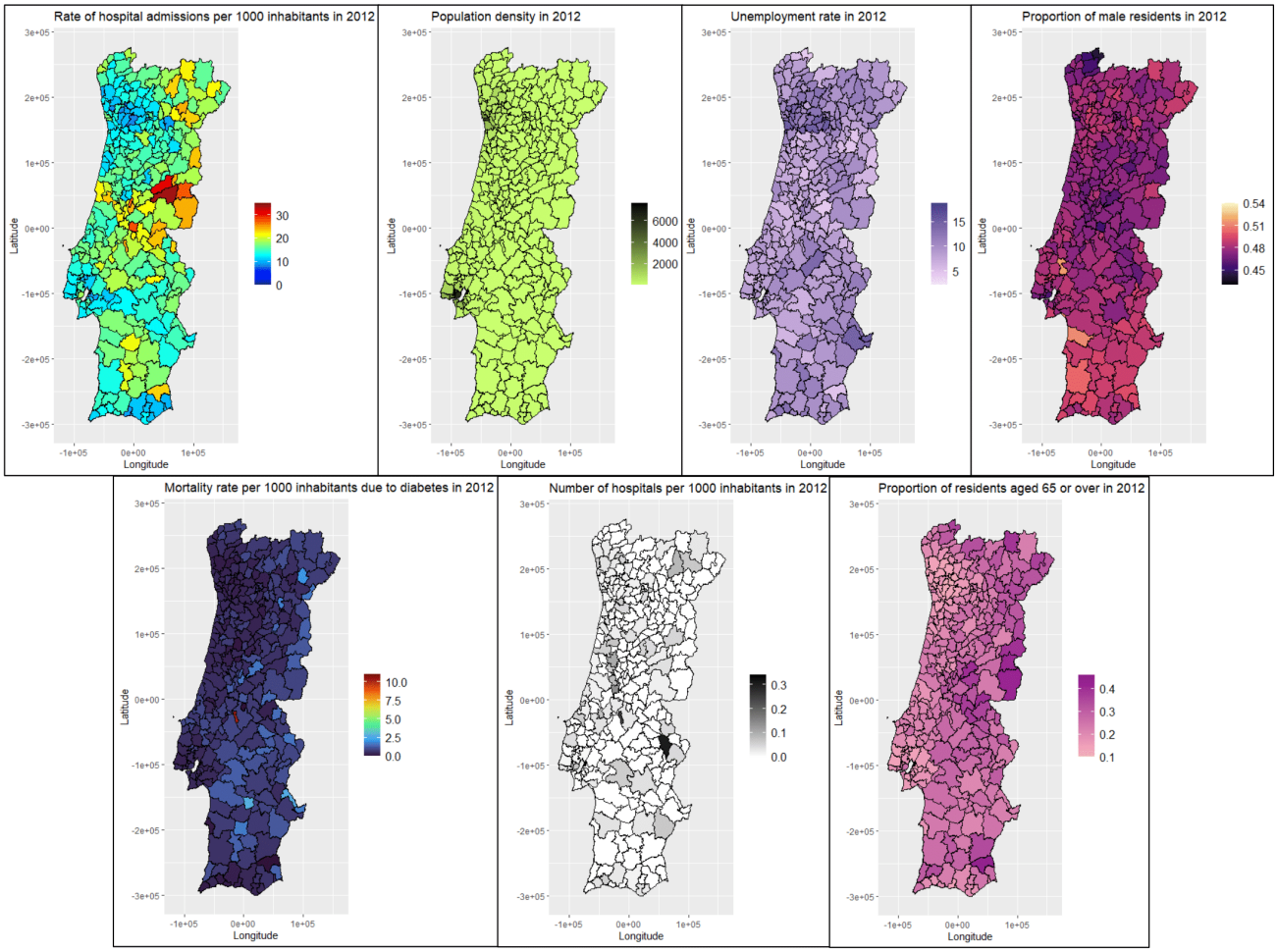

Figure 7: Maps at municipality level of the hospital admission rate and the variables from INE in 2012.

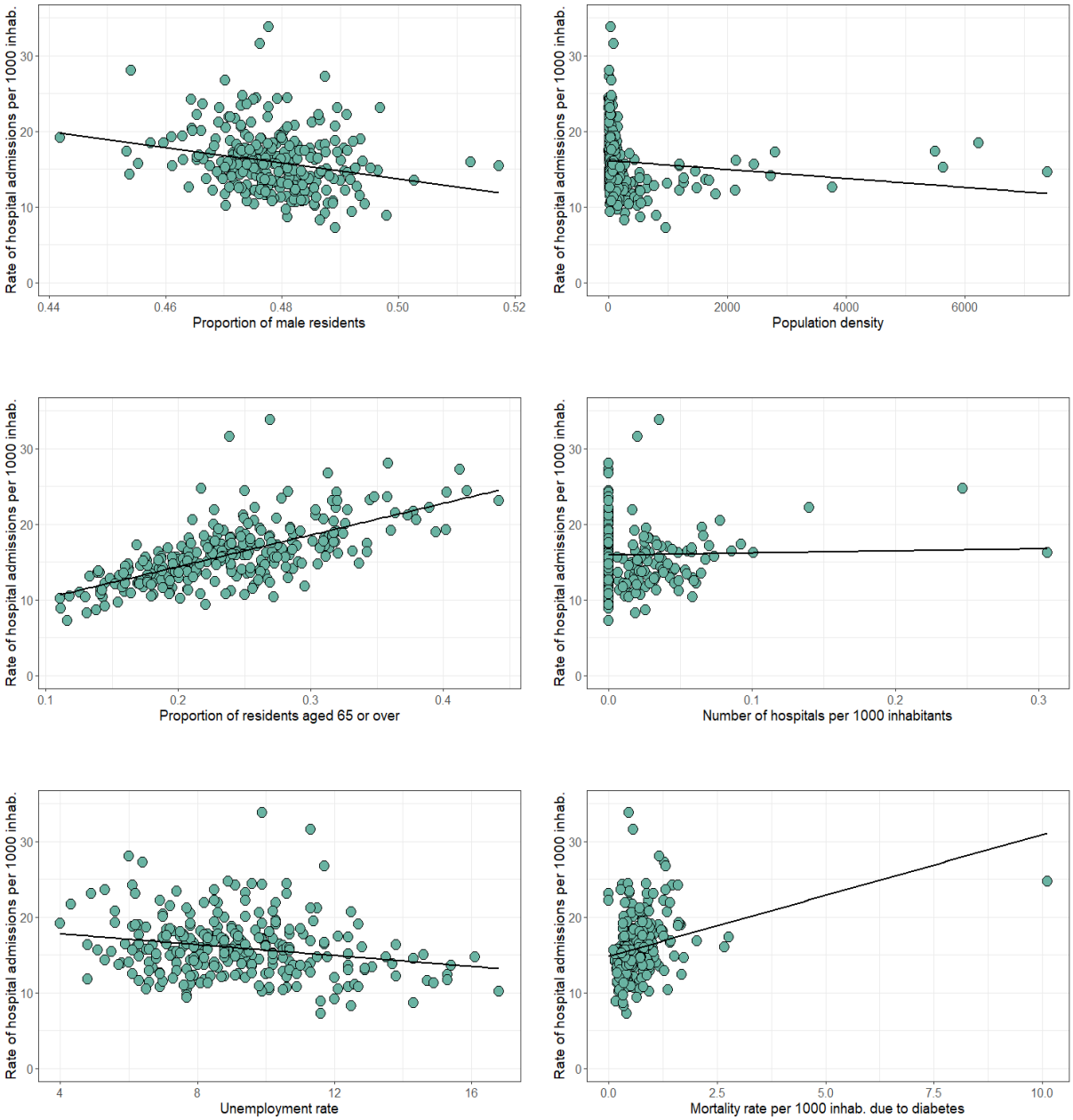

Figure 8: Scatter plot of the hospital admission rate vs all the variables at municipality level for 2012 and the respective regression line of the GLM model.

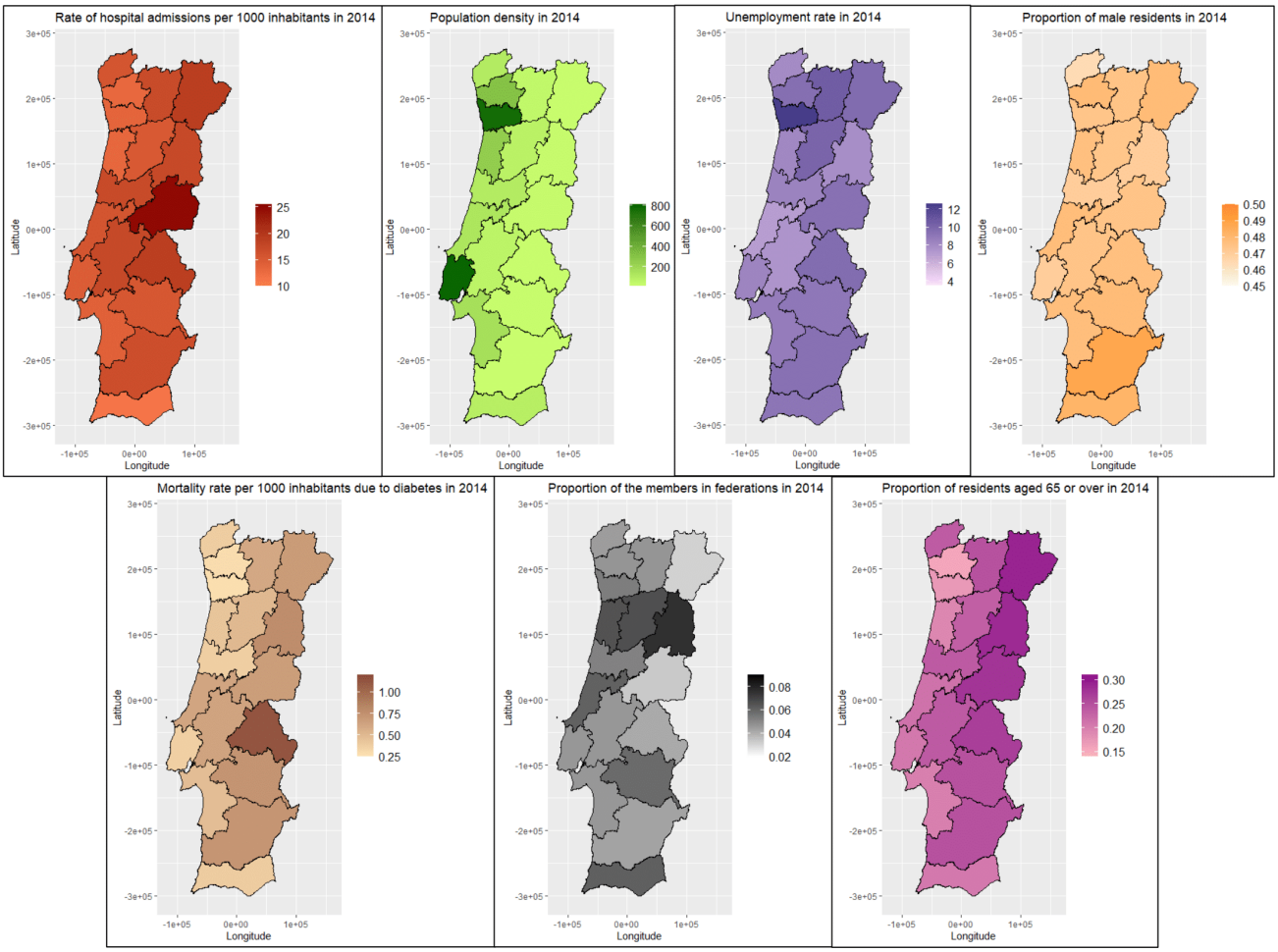

Figure 9: Maps at district level of the hospital admission rate and the variables from INE in 2014.

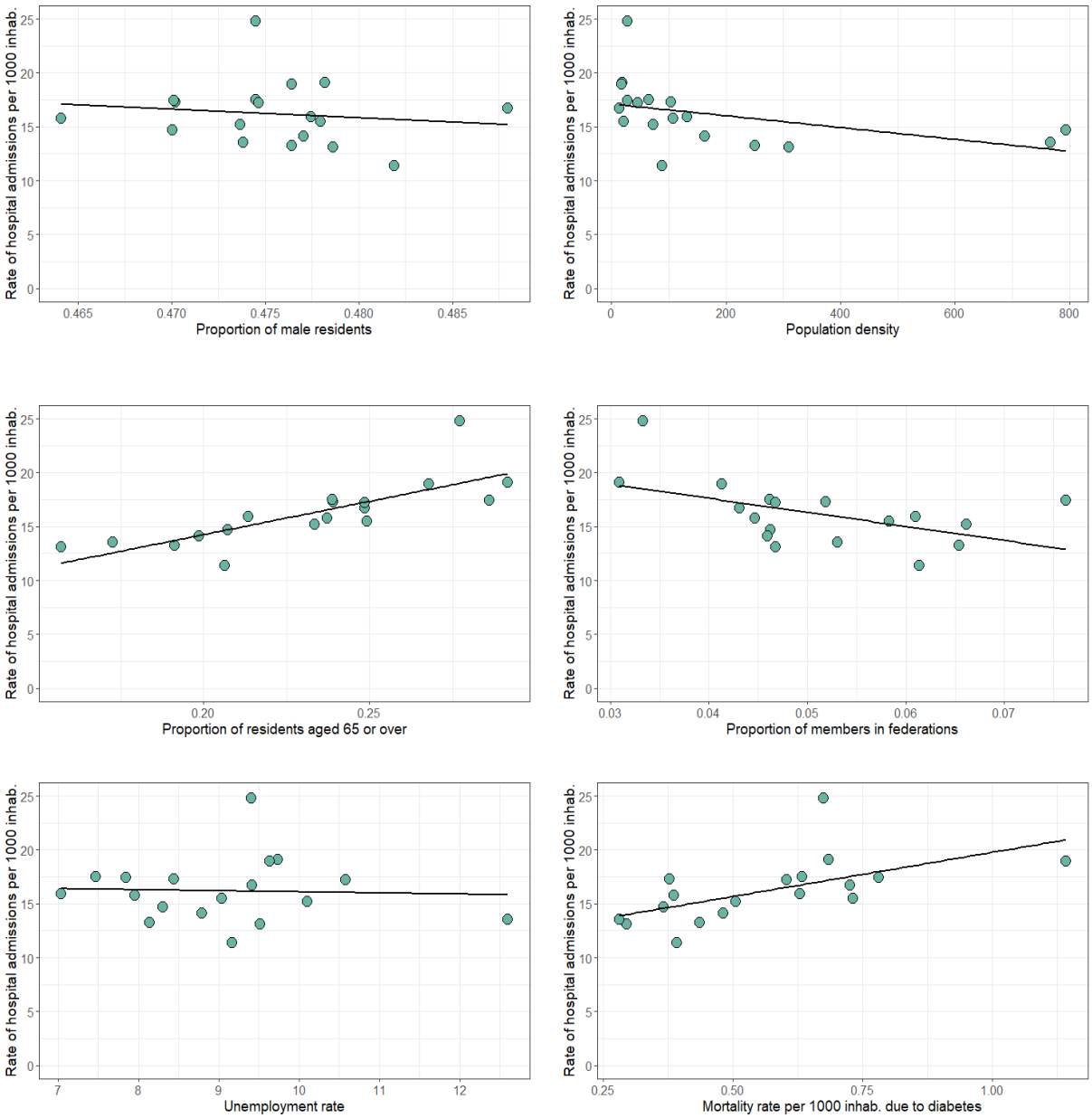

Figure 10: Scatter plot of the hospital admission rate vs all the variables at district level for 2014 and the respective regression line of the GLM model.

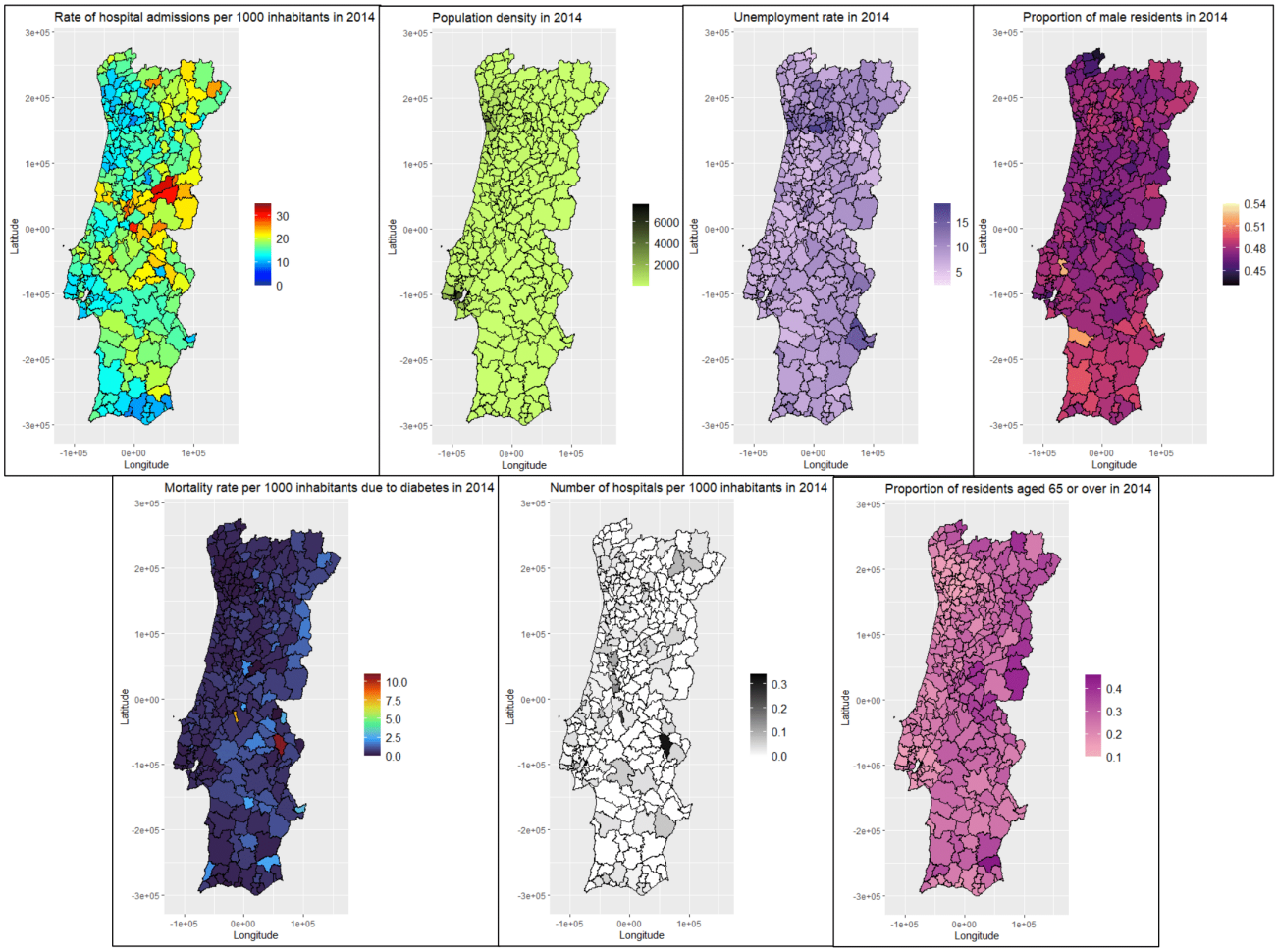

Figure 11: Maps at municipality level of the hospital admission rate and the variables from INE in 2014.

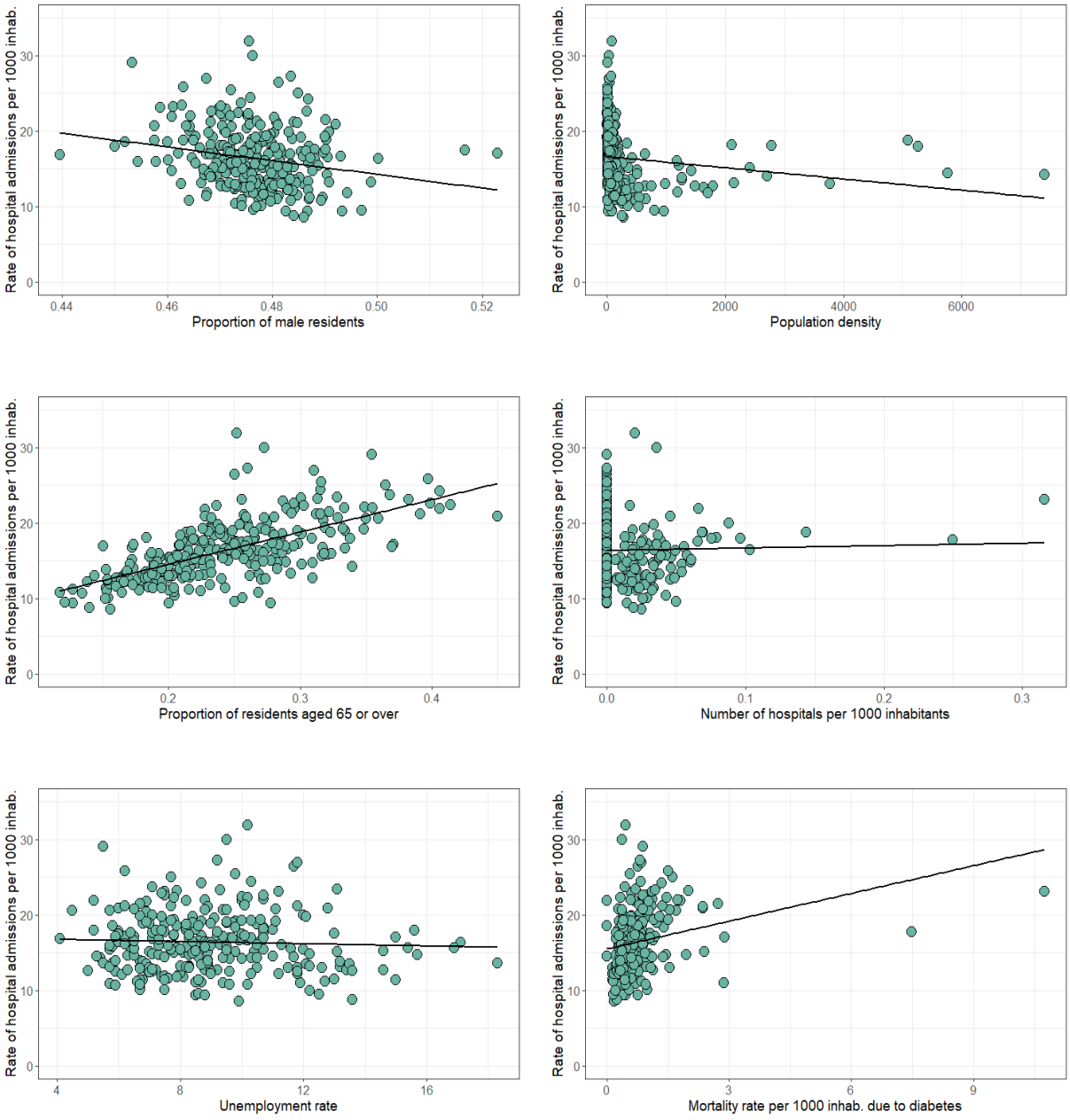

Figure 12: Scatter plot of the hospital admission rate vs all the variables at municipality level for 2014 and the respective regression line of the GLM model.

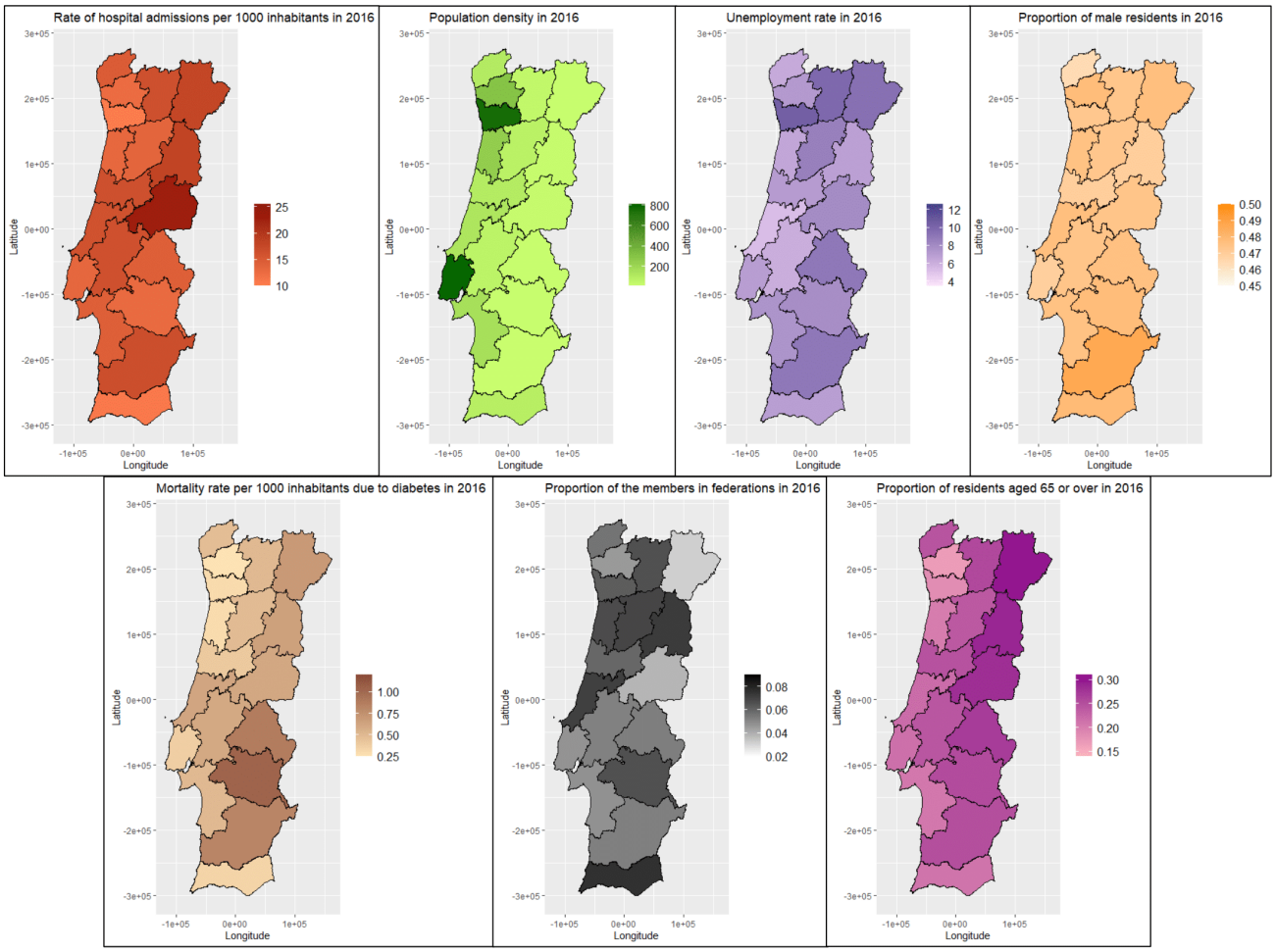

Figure 13: Maps at district level of the hospital admission rate and the variables from INE in 2016.

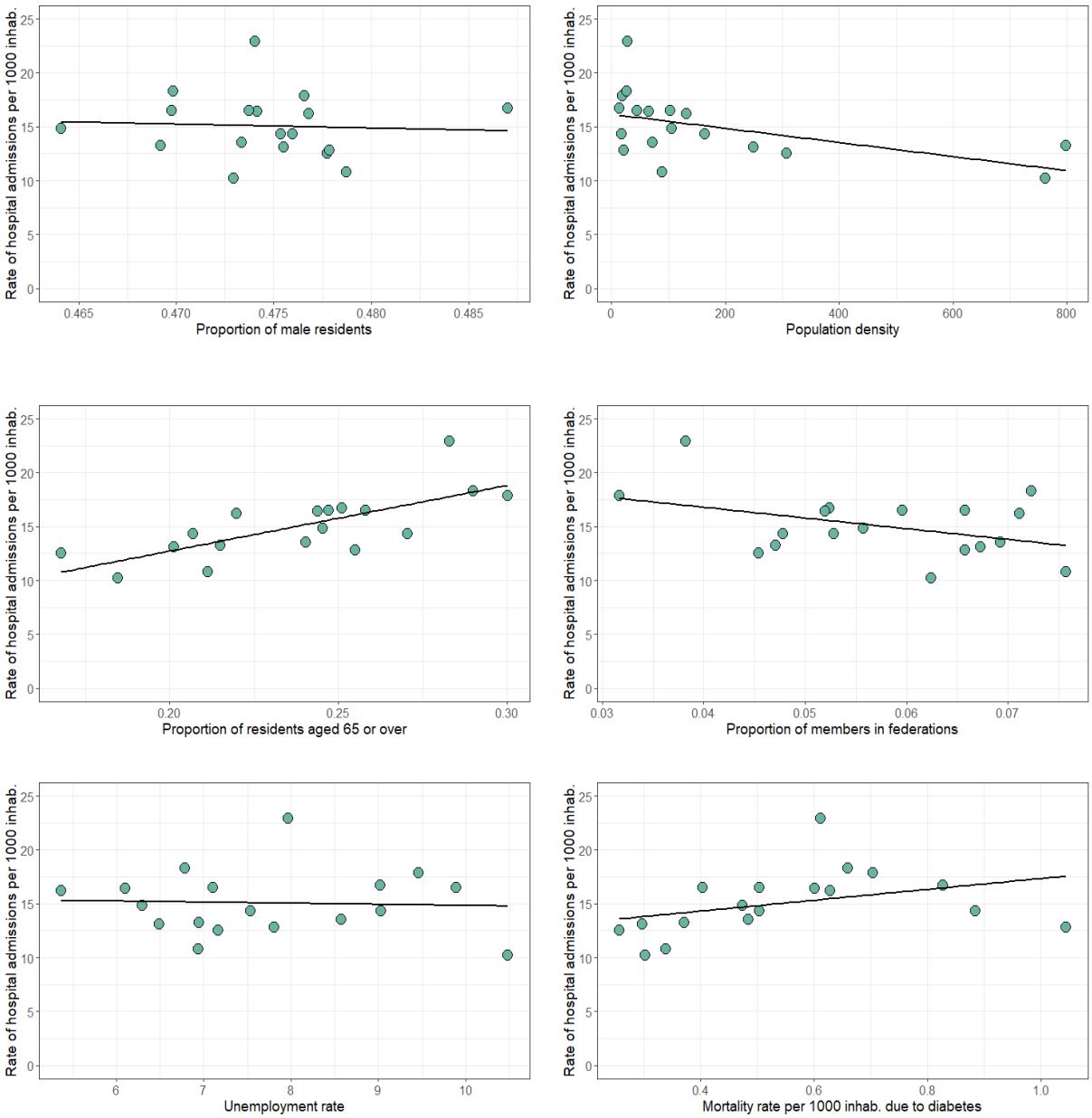

Figure 14: Scatter plot of the hospital admission rate vs all the variables at district level for 2016 and the respective regression line of the GLM model.

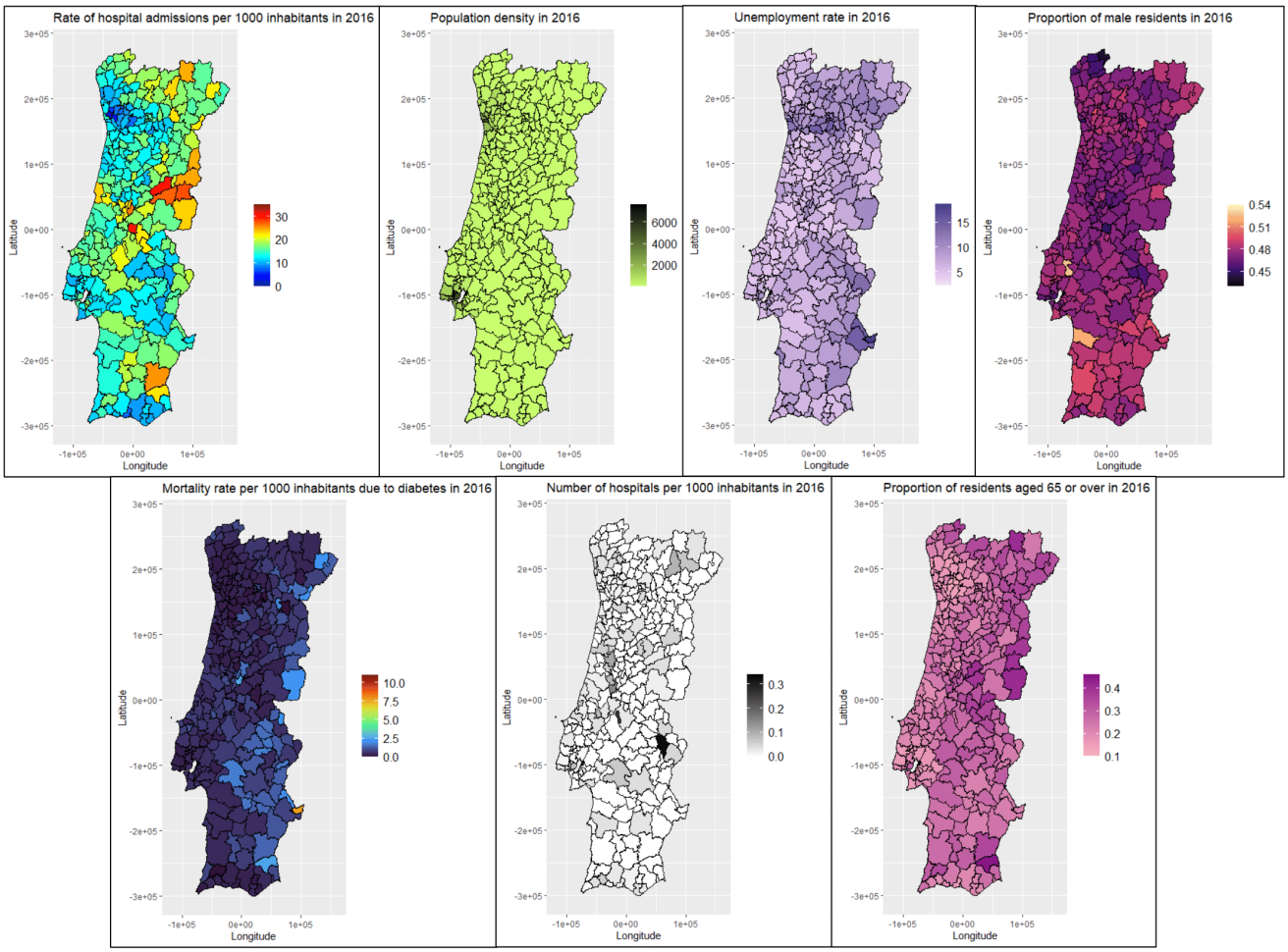

Figure 15: Maps at municipality level of the hospital admission rate and the variables from INE in 2016.

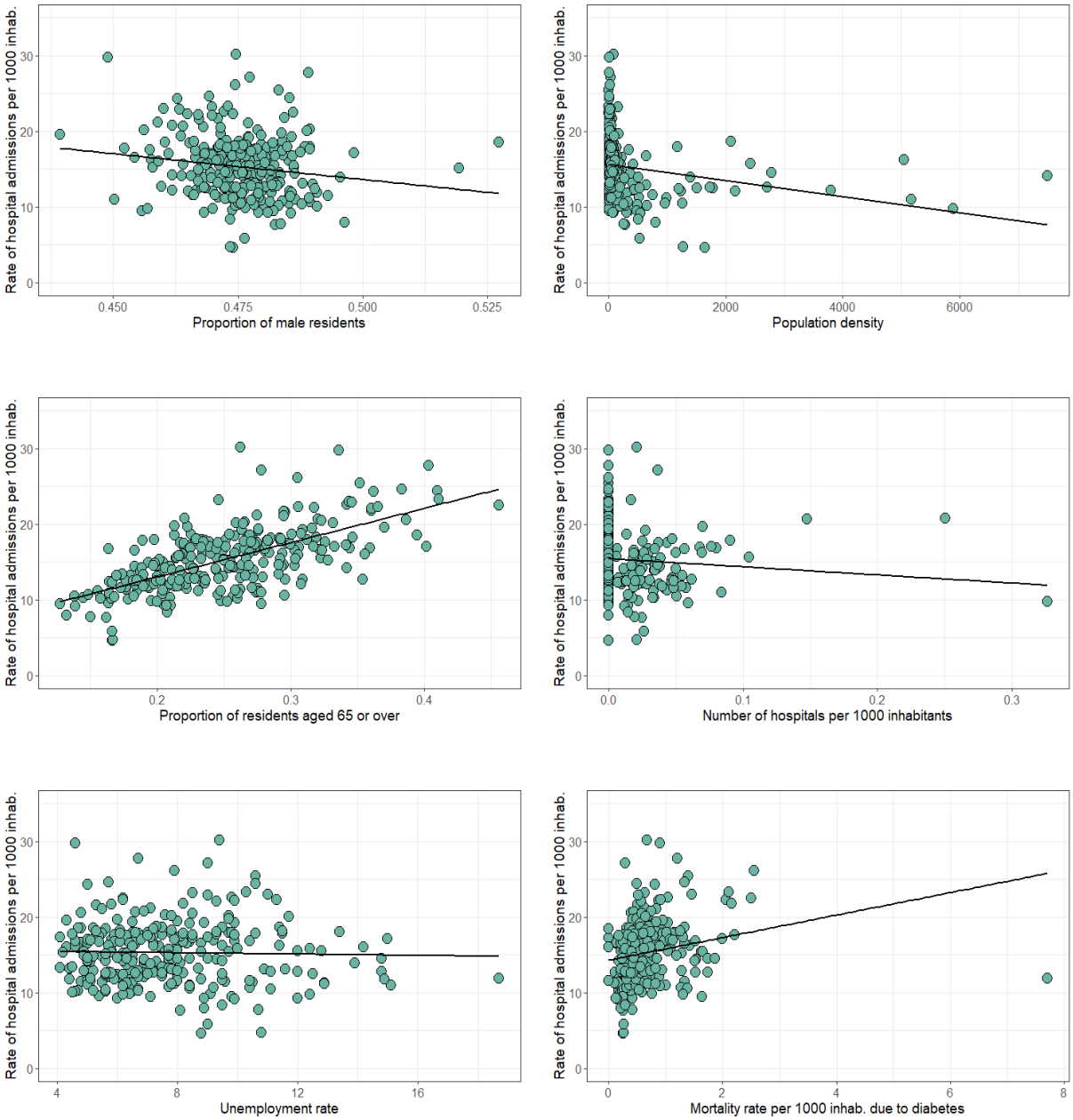

Figure 16: Scatter plot of the hospital admission rate vs all the variables at municipality level for 2016 and the respective regression line of the GLM model.

C R-code for Multiple Imputation

```
library(finalfit)
library(mice)
library(readxl)
library(data.table)
library(stringr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(lattice)
```

```
data<-cbind(BD_concelhos_without_diabetes[,-c(1:6)],N_mortes_diabetes)

md.pattern(data)
p <- md.pairs(data);p
p_missing <- unlist(lapply(data, function(x) sum(is.na(x))))/nrow(data)
sort(p_missing[p_missing > 0], decreasing = TRUE)

init = mice(data,print=FALSE)
as.character(init$loggedEvents[, "out"])

predM = init$predictorMatrix
predM[, c("Prop_males")]<-0

imp_pmm_5 <- mice(data,predictorMatrix=predM, m = 5, maxit = 20,
                  seed = 30031, print=FALSE)

stripplot(imp_pmm_5,pch = 20, cex = 1.2)
#the imputed values are plausible

plot(imp_pmm_5, c("N_mortes_diabetes"))
#healthy convergence

imp1_pmm<-complete(imp_pmm_5,1)
imp2_pmm<-complete(imp_pmm_5,2)
imp3_pmm<-complete(imp_pmm_5,3)
imp4_pmm<-complete(imp_pmm_5,4)
imp5_pmm<-complete(imp_pmm_5,5)
```

```
#Table with final data

BD_concelhos_with_diabetes<-cbind(BD_concelhos_without_diabetes,
                            c(apply(cbind(imp1_pmm$N_mortes_diabetes,
                            imp2_pmm$N_mortes_diabetes,
                            imp3_pmm$N_mortes_diabetes,
                            imp4_pmm$N_mortes_diabetes,
                            imp5_pmm$N_mortes_diabetes),1,mean)))
names(BD_concelhos_with_diabetes)[12]<-"N_mortes_diabetes"
BD_concelhos_with_diabetes$N_mortes_diabetes<-
          round(BD_concelhos_with_diabetes$N_mortes_diabetes,0)


grafico_imput = function(real,imputado,x){
  data.frame("Antes de imputar" = real,
             "Depois de imputar" = imputado) %>%
    pivot_longer(everything()) %>%
    na.omit() %>%
    ggplot(mapping = aes(value,col= name))+
    geom_density()+
    theme_minimal()+
    xlab(x)+
    ylab("Density")+
    theme(panel.border =
        element_rect(colour = "black", fill = NA, size = 0.2))+
    scale_color_manual(values = c("red", "blue"),
    labels = c("Before imputation","After imputation"))+
    guides(color=guide_legend(title=""))
  }

grafico_imput(data$N_mortes_diabetes,
              BD_concelhos_with_diabetes$N_mortes_diabetes,
              "Number of deaths due to diabetes")
#As the densities are quite similar, it follows that the
imputation was plausible.
```

D R-code for Bayesian approach

D.1 Packages

```
library(tidyverse)
library(fields)
```

```
library(maps)
library(rgdal)
library(broom)
library(rgeos)
library(ggpubr)
library(spdep)
library(INLA)
library(maptools)
library(ggregplot)
```

D.2 Spatial analysis at district level

D.2.1 Data preparation

```
Dist_mapas<-readOGR("Dist.shp")
Dist_mapas<-Dist_mapas[1:18,] #remove the islands

Dist_mapas_s<-gSimplify(Dist_mapas,tol=50)
Dist_mapas_dt<-tidy(Dist_mapas_s)

temp<-poly2nb(Dist_mapas)
nb2INLA("Dist.graph", temp)
Dist.adj <- paste(getwd(),"/Dist.graph",sep="")

#NEIGHBOURHOOD PLOT
H<-inla.read.graph(filename="Dist.graph")
image(inla.graph2matrix(H),xlab="",ylab="")

BD_final_distrito_MI<-readRDS("BD_final_distrito_MI.rds")
BD_final_distrito_MI<-BD_final_distrito_MI[BD_final_distrito_MI$Year
                                    %in% c("2010","2012","2014","2016",
                                    "2018"),]

names(BD_final_distrito_MI)[1]<-"ID.area"

ano_2018_MI<-BD_final_distrito_MI[BD_final_distrito_MI$Year==2018,]
ano_2018_MI$ID.area<-as.numeric(ano_2018_MI$ID.area)
```

D.2.2 Models

```
#IID MODEL
formula_iid<-N_admissions~scale(Density_pop)+scale(Unemployment_rate)+
  scale(Prop_males)+scale(Mort_rate_diabetes_1000)+
```

```
  scale(Prop_pract_federations)+scale(Prop_residents_65)+
  f(ID.area,model="iid",graph = Dist.adj)
model_iid <- inla(formula_iid,family="poisson",data=ano_2018_MI,
                  offset=log(Total_inhabitants),
                  control.predictor=list(compute=TRUE),
                  control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_iid)

#BESAG MODEL
formula_besag<-N_admissions~scale(Density_pop)+scale(Unemployment_rate)+
  scale(Prop_males)+scale(Mort_rate_diabetes_1000)+
  scale(Prop_pract_federations)+scale(Prop_residents_65)+
  f(ID.area,model="besag",graph = Dist.adj)
model_besag <- inla(formula_besag,family="poisson",data=ano_2018_MI,
                    offset=log(Total_inhabitants),
                    control.predictor=list(compute=TRUE),
                    control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_besag)

#BYM MODEL
formula_bym<-N_admissions~scale(Density_pop)+scale(Unemployment_rate)+
  scale(Prop_males)+scale(Mort_rate_diabetes_1000)+
  scale(Prop_pract_federations)+scale(Prop_residents_65)+
  f(ID.area,model="bym",graph = Dist.adj)
model_bym <- inla(formula_bym,family="poisson",data=ano_2018_MI,
                  offset=log(Total_inhabitants),
                  control.predictor=list(compute=TRUE),
                  control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_bym)
```

D.2.3 Graphs of the random effect models

```
id_maps <-rep(0:17)
posterior_spatial_random_2018<-model_iid$summary.random$ID.area$mean
BD_spatial_random_effect_2018<-data.frame(id_maps=as.factor(id_maps),
    posterior_spatial_random_2018)
BD_spatial_random_effect_2018_maps<-left_join(Dist_mapas_dt,
    BD_spatial_random_effect_2018,by=c("id"="id_maps"))

ggplot(data=BD_spatial_random_effect_2018_maps,aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=posterior_spatial_random_2018),
               colour="black")+
  xlab("Longitude") + ylab("Latitude") +
```

```
  labs(fill = "")+
  scale_fill_gradient(low="blue", high="yellow")+
  coord_fixed(1.1)
```

D.2.4 Graphs of the observed and fitted values

```
id_maps <-rep(0:17)
rate_admission_observed<-(ano_2018_MI$N_admissions/
                          ano_2018_MI$Total_inhabitants)*1000
ano_2018_MI<-cbind(id_maps=as.factor(id_maps),ano_2018_MI,
                   rate_admission_observed)
observed_data_2018<-left_join(Dist_mapas_dt,ano_2018_MI,by=c("id"="id_maps"))

ggplot(data=observed_data_2018,aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=rate_admission_observed),
               colour="black")+
  xlab("Longitude") + ylab("Latitude") +
  scale_fill_gradient(low="blue", high="yellow",limits=c(10,21))+
  coord_fixed(1.1)

rate_admission_fitted<-(model_iid$summary.fitted.values$mean/
                            ano_2018_MI$Total_inhabitants)*1000
BD_fitted_rate<-data.frame(id_maps=as.factor(id_maps),ano_2018_MI$District,
                             rate_admission_fitted)
BD_fitted_data_2018<-left_join(Dist_mapas_dt,
                               BD_fitted_rate,by=c("id"="id_maps"))

ggplot(data=BD_fitted_data_2018,aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=rate_admission_fitted),
               colour="black")+
  xlab("Longitude") + ylab("Latitude") +
  scale_fill_gradient(low="blue", high="yellow",limits=c(10,21))+
  coord_fixed(1.1)
```

D.3 Spatial analysis at municipality level

D.3.1 Data preparation

```
Mun_mapas<-readOGR("Mun.shp")
Mun_mapas<-Mun_mapas[1:278,] #remove the islands

Mun_mapas_s<-gSimplify(Mun_mapas,tol=50)
Mun_mapas_dt<-tidy(Mun_mapas_s)
```

```
temp<-poly2nb(Mun_mapas)
nb2INLA("Conc.graph", temp)
Conc.adj <- paste(getwd(),"/Conc.graph",sep="")

#NEIGHBOURHOOD PLOT
H<-inla.read.graph(filename="Conc.graph")
image(inla.graph2matrix(H),xlab="",ylab="")

BD_final_concelhos_MI<-readRDS("BD_final_concelhos_MI.rds")
BD_final_concelhos_MI<-BD_final_concelhos_MI[BD_final_concelhos_MI$Year %in%
                                               c("2010","2012","2014","2016",
                                               "2018"),]

names(BD_final_concelhos_MI)[1]<-"ID.area"

ano_2018_conc_MI<-BD_final_concelhos_MI[BD_final_concelhos_MI$Year==2018,]
ano_2018_conc_MI$ID.area<-as.integer(ano_2018_conc_MI$ID.area)
```

D.3.2 Models

```
#IID MODEL
formula_conc_iid<-N_admissions~scale(Density_pop)+scale(Unemployment_rate)+
  scale(Prop_males)+scale(Mort_rate_diabetes_1000)+scale(Hospitals_1000)+
  scale(Prop_residents_65)+
  f(ID.area,model="iid",graph = Conc.adj)
model_conc_iid <- inla(formula_conc_iid,family="poisson",
                       data=ano_2018_conc_MI,
                       offset=log(Total_inhabitants),
                       control.predictor=list(compute=TRUE),
                       control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_conc_iid)

#BESAG MODEL
formula_conc_besag<-N_admissions~scale(Density_pop)+scale(Unemployment_rate)+
  scale(Prop_males)+scale(Mort_rate_diabetes_1000)+scale(Hospitals_1000)+
  scale(Prop_residents_65)+
  f(ID.area,model="besag",graph = Conc.adj)
model_conc_besag <- inla(formula_conc_besag,family="poisson",
                         data=ano_2018_conc_MI,
                         offset=log(Total_inhabitants),
                         control.predictor=list(compute=TRUE),
                         control.fixed=list(mean=list(0),prec=list(0.0001),
                                            mean.intercept=0,
```

```
                                              prec.intercept=0.0001),
                          control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_conc_besag)

#BYM MODEL
formula_conc_bym<-N_admissions~scale(Density_pop)+scale(Unemployment_rate)+
  scale(Prop_males)+scale(Mort_rate_diabetes_1000)+scale(Hospitals_1000)+
  scale(Prop_residents_65)+
  f(ID.area,model="bym",graph = Conc.adj)
model_conc_bym <- inla(formula_conc_bym,family="poisson",
                       data=ano_2018_conc_MI,
                       offset=log(Total_inhabitants),
                       control.predictor=list(compute=TRUE),
                       control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_conc_bym)
```

D.3.3 Graphs of the random effect models

```
id_maps <-rep(0:277)
posterior_spatial_random_2018_conc<-
           model_conc_besag$summary.random$ID.area$mean
BD_spatial_random_effect_2018_conc<-data.frame(id_maps=as.factor(id_maps),
                                       posterior_spatial_random_2018_conc)
BD_spatial_random_effect_2018_conc_maps<-left_join(Mun_mapas_dt,
                                    BD_spatial_random_effect_2018_conc,
                                    by=c("id"="id_maps"))

ggplot(data=BD_spatial_random_effect_2018_conc_maps,aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=posterior_spatial_random_2018_conc),
               colour="black")+
  xlab("Longitude") + ylab("Latitude") +
  labs(fill = "")+
  scale_fill_gradientn(colours=tim.colors(99), limits=c(-2.56,1))+
  coord_fixed(1.1)
```

D.3.4 Graphs of the observed and fitted values

```
id_maps <-rep(0:277)
rate_admission_observed_2018_conc<-(ano_2018_conc_MI$N_admissions/
                             ano_2018_conc_MI$Total_inhabitants)*1000
ano_2018_conc_MI<-cbind(id_maps=as.factor(id_maps),ano_2018_conc_MI,
                        rate_admission_observed_2018_conc)
BD_observed_rate_2018_conc_maps<-left_join(Mun_mapas_dt,ano_2018_conc_MI,
```

```
                                          by=c("id"="id_maps"))

ggplot(data=BD_observed_rate_2018_conc_maps,aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=rate_admission_observed_2018_conc),
               colour="black")+
  xlab("Longitude") + ylab("Latitude") +
  labs(fill = "")+
  scale_fill_gradientn(colours=tim.colors(99),limits=c(0,32)) +
  coord_fixed(1.1)

rate_admission_fitted_conc<-(model_conc_besag$summary.fitted.values$mean/
                             ano_2018_conc_MI$Total_inhabitants)*1000
BD_fitted_rate_conc<-data.frame(id_maps=as.factor(id_maps),
                                ano_2018_conc_MI$Municipalities,
                                rate_admission_fitted_conc)
names(BD_fitted_rate_conc)<-c("id_maps","Municipalities",
                              "Rate_admission_fitted")

BD_fitted_rate_2018_conc_maps<-left_join(Mun_mapas_dt,BD_fitted_rate_conc,
                                         by=c("id"="id_maps"))

ggplot(data=BD_fitted_rate_2018_conc_maps,aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=Rate_admission_fitted),
               colour="black")+
  xlab("Longitude") + ylab("Latitude") +
  labs(fill = "")+
  scale_fill_gradientn(colours=tim.colors(99),limits=c(0,32))+
  coord_fixed(1.1)
```

D.4 Diagnosis of the spatial models chosen

```
#DISTRICT LEVEL
model_iid <- inla(formula_iid,family="poisson",data=ano_2018_MI,
                  offset=log(Total_inhabitants),
                  control.predictor=list(link=1,compute=TRUE),
                  control.fixed=list(mean=0,prec=0.00001),
                  control.compute=list(return.marginals.predictor=TRUE))

par(mfrow=c(1,2))
plot(ano_2018_MI$N_admissions,model_iid$summary.fitted.values$mean,
     xlab="Observed values",ylab="Mean Post. Pred. Distr.")
```

```
predicted.p.value<-c()
for(i in (1:18)) {
  predicted.p.value[i] <- inla.pmarginal(q=ano_2018_MI$N_admissions[i],
                       marginal=model_iid$marginals.fitted.values[[i]])
}
hist(predicted.p.value,main="",
     xlab="Posterior predictive p-value",
     breaks=5,xlim=c(0.4,0.6))

#MUNICIPALITY LEVEL
model_conc_besag <- inla(formula_conc_besag,family="poisson",
                         data=ano_2018_conc_MI,
                         offset=log(Total_inhabitants),
                         control.predictor=list(link=1,compute=TRUE),
                         control.fixed=list(mean=0,prec=0.00001),
                       control.compute=list(return.marginals.predictor=TRUE))

par(mfrow=c(1,2))
plot(ano_2018_conc_MI$N_admissions,
     model_conc_besag$summary.fitted.values$mean,
     xlab="Observed values",ylab="Mean Post. Pred. Distr.")


predicted.p.value<-c()
for(i in (1:278)) {
  predicted.p.value[i] <- inla.pmarginal(q=ano_2018_conc_MI$N_admissions[i],
                      marginal=model_conc_besag$marginals.fitted.values[[i]])
}
hist(predicted.p.value,main="",
     xlab="Posterior predictive p-value",
     breaks=9,xlim=c(0,1.0),ylim=c(0,100))
```

D.5 Spatio-temporal analysis without interaction at district level

D.5.1 Models

```
ID.area<-rep(1:18,each=5)
ID.year<-rep(1:5,18)
ID.year1<-ID.year

formula_year_iid<-N_admissions~scale(Density_pop)+scale(Unemployment_rate)+
  scale(Prop_males)+scale(Mort_rate_diabetes_1000)+
  scale(Prop_pract_federations)+scale(Prop_residents_65)+
  f(ID.area,model="iid",graph = Dist.adj)+f(ID.year,model="iid")
```

```
model_year_iid <- inla(formula_year_iid,family="poisson",
                       data=BD_final_distrito_MI,
                       offset=log(Total_inhabitants),
                       control.predictor=list(compute=TRUE),
                       control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_year_iid)

formula_year_rw1<-N_admissions~scale(Density_pop)+scale(Unemployment_rate)+
  scale(Prop_males)+scale(Mort_rate_diabetes_1000)+
  scale(Prop_pract_federations)+scale(Prop_residents_65)+
  f(ID.area,model="iid",graph = Dist.adj)+f(ID.year,model="rw1")
model_year_rw1 <- inla(formula_year_rw1,family="poisson",
                       data=BD_final_distrito_MI,
                       offset=log(Total_inhabitants),
                       control.predictor=list(compute=TRUE),
                       control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_year_rw1)

formula_year_rw2<-N_admissions~scale(Density_pop)+scale(Unemployment_rate)+
  scale(Prop_males)+scale(Mort_rate_diabetes_1000)+
  scale(Prop_pract_federations)+scale(Prop_residents_65)+
  f(ID.area,model="iid",graph = Dist.adj)+f(ID.year,model="rw2")
model_year_rw2 <- inla(formula_year_rw2,family="poisson",
                       data=BD_final_distrito_MI,
                       offset=log(Total_inhabitants),
                       control.predictor=list(compute=TRUE),
                       control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_year_rw2)

formula_year_iid_rw1<-N_admissions~scale(Density_pop)+
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Prop_pract_federations)+
  scale(Prop_residents_65)+
  f(ID.area,model="iid",graph = Dist.adj)+
  f(ID.year,model="rw1")+f(ID.year1,model="iid")
model_year_iid_rw1 <- inla(formula_year_iid_rw1,family="poisson",
                           data=BD_final_distrito_MI,
                           offset=log(Total_inhabitants),
                           control.predictor=list(compute=TRUE),
                           control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_year_iid_rw1)

formula_year_iid_rw2<-N_admissions~scale(Density_pop)+
```

```
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Prop_pract_federations)+
  scale(Prop_residents_65)+
  f(ID.area,model="iid",graph = Dist.adj)+
  f(ID.year,model="rw2")+f(ID.year1,model="iid")
model_year_iid_rw2 <- inla(formula_year_iid_rw2,family="poisson",
                           data=BD_final_distrito_MI,
                           offset=log(Total_inhabitants),
                           control.predictor=list(compute=TRUE),
                           control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_year_iid_rw2)
```

D.5.2 Graphs of the random effect models

```
#Map of spatial random effect
id_maps <-rep(0:17)
posterior_spatial_random_years<-model_year_rw2$summary.random$ID.area$mean
BD_spatial_random_effect_years<-data.frame(id_maps=as.factor(id_maps),
                                           posterior_spatial_random_years)
BD_spatial_random_effect_years_maps<-left_join(Dist_mapas_dt,
                                    BD_spatial_random_effect_years,
                                    by=c("id"="id_maps"))

ggplot(data=BD_spatial_random_effect_years_maps,aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=posterior_spatial_random_years),
               colour="black")+
  xlab("Longitude") + ylab("Latitude") +
  labs(fill = "")+
  scale_fill_gradient(low="blue", high="yellow")+
  coord_fixed(1.1)

#Map of temporal random effect
years<-c(2010,2012,2014,2016,2018)

plot(years,model_year_rw2$summary.random$ID.year$mean,
      xlab="t",ylab=expression(paste(gamma[t])),cex=0)
lines(years,model_year_rw2$summary.random$ID.year$mean,
      xlab="t",ylab=expression(paste(gamma[t])),cex=0)
```

D.5.3 Graphs of the observed and fitted values

```
rate_admission_observed_temp<-(BD_final_distrito_MI$N_admissions/
                               BD_final_distrito_MI$Total_inhabitants)*1000
```

```
rate_admission_fitted_temp<-(model_year_rw2$summary.fitted.values$mean/
                               BD_final_distrito_MI$Total_inhabitants)*1000

rate_observed_fitted_without_int<-c(rate_admission_observed_temp,
                                    rate_admission_fitted_temp)
BD_observed_fitted_rate_temp<-data.frame(rep(rep(0:17,each=5),2),
                                 rep(BD_final_distrito_MI$District,2),
                                 c(rep(c("2010_observed","2012_observed",
                                         "2014_observed","2016_observed",
                                         "2018_observed"),18),
                                   rep(c("2010_fitted","2012_fitted",
                                         "2014_fitted","2016_fitted",
                                         "2018_fitted"),18)),
                                 rate_observed_posterior_without_int)

names(BD_observed_fitted_rate_temp)<-c("id_maps","District","Year",
                                    "Rate_admission")

BD_observed_fitted_rate_temp_maps<-left_join(Dist_mapas_dt,
                                             BD_observed_fitted_rate_temp,
                                             by=c("id"="id_maps"))

ggplot(data=BD_observed_fitted_rate_temp_maps, aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=Rate_admission),
               colour="black")+
  coord_equal() +
  labs(fill = "")+
  scale_fill_gradient(low="blue", high="yellow", limits=c(10,25.5))+
  facet_wrap(~BD_observed_fitted_rate_temp_maps$Year)
```

D.6 Spatio-temporal analysis with interaction at district level

D.6.1 Models

```
ID.area.year<-1:(18*5) #18 districts x 5 years

formula_int_type_I<-N_admissions~scale(Density_pop)+
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Prop_pract_federations)+
  scale(Prop_residents_65)+
  f(ID.area,model="iid",graph = Dist.adj)+f(ID.year,model="rw2")+
  f(ID.area.year,model="iid")
model_int_type_I <- inla(formula_int_type_I,family="poisson",
```

```
                          data=BD_final_distrito_MI,
                          offset=log(Total_inhabitants),
                          control.predictor=list(compute=TRUE),
                          control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_int_type_I)

ID.area.int<-ID.area
ID.year.int<-ID.year
formula_int_type_II<-N_admissions~scale(Density_pop)+
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Prop_pract_federations)+
  scale(Prop_residents_65)+
  f(ID.area,model="iid",graph = Dist.adj)+f(ID.year,model="rw2")+
  f(ID.area.int,model="iid",group=ID.year.int,
  control.group = list(model="rw2"))
model_int_type_II <- inla(formula_int_type_II,family="poisson",
                          data=BD_final_distrito_MI,
                          offset=log(Total_inhabitants),
                          control.predictor=list(compute=TRUE),
                          control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_int_type_II)

formula_int_type_III<-N_admissions~scale(Density_pop)+
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Prop_pract_federations)+
  scale(Prop_residents_65)+
  f(ID.area,model="iid",graph = Dist.adj)+f(ID.year,model="rw2")+
  f(ID.year.int,model="iid",group=ID.area.int,
  control.group = list(model="iid"))
model_int_type_III <- inla(formula_int_type_III,family="poisson",
                          data=BD_final_distrito_MI,
                          offset=log(Total_inhabitants),
                          control.predictor=list(compute=TRUE),
                          control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_int_type_III)

formula_int_type_IV<-N_admissions~scale(Density_pop)+
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Prop_pract_federations)+
  scale(Prop_residents_65)+
  f(ID.area,model="iid",graph = Dist.adj)+f(ID.year,model="rw2")+
  f(ID.area.int,model="iid",group=ID.year.int,
  control.group = list(model="rw2"))
```

```
model_int_type_IV <- inla(formula_int_type_IV,family="poisson",
                          data=BD_final_distrito_MI,
                          offset=log(Total_inhabitants),
                          control.predictor=list(compute=TRUE),
                          control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_int_type_IV)
```

D.6.2 Graphs of the random effect models

```
#Map of spatial random effect
id_maps <-rep(0:17)
posterior_spatial_random_int<-model_int_type_II$summary.random$ID.area$mean
BD_spatial_random_effect_int<-data.frame(id_maps=as.factor(id_maps),
                                         posterior_spatial_random_int)
BD_spatial_random_effect_int_maps<-left_join(Dist_mapas_dt,
                                             BD_spatial_random_effect_int,
                                             by=c("id"="id_maps"))

ggplot(data=BD_spatial_random_effect_int_maps,aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=posterior_spatial_random_int),
               colour="black")+
  xlab("Longitude") + ylab("Latitude") +
  labs(fill = "")+
  scale_fill_gradient(low="coral", high="darkred")+
  coord_fixed(1.1)

#Map of temporal random effect
years<-c(2010,2012,2014,2016,2018)

plot(years,model_int_type_II$summary.random$ID.year$mean,
     xlab="t",ylab=expression(paste(gamma[t])),cex=0)
lines(years,model_int_type_II$summary.random$ID.year$mean,
      xlab="t",ylab=expression(paste(gamma[t])),cex=0)

#Map for interaction random effect
posterior_interaction_random<-
          model_int_type_II$summary.random$ID.area.int$mean

BD_posterior_interaction<-data.frame(rep(0:17,each=5),
                                     BD_final_distrito_MI$District,
                                     BD_final_distrito_MI$Year,
                                     posterior_interaction_random)
names(BD_posterior_interaction)<-c("id_maps","District","Year",
```

```
                                    "Posterior_interaction")
BD_posterior_interaction_maps<-left_join(Dist_mapas_dt,
                                         BD_posterior_interaction,
                                         by=c("id"="id_maps"))

ggplot(data=BD_posterior_interaction_maps, aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=Posterior_interaction),
               colour="black")+
  coord_equal() +
  labs(fill = "")+
  scale_fill_gradient(low="blue", high="yellow")+
  facet_wrap(BD_posterior_interaction_maps$Year)
```

D.6.3 Graphs of the observed and fitted values

```
rate_admission_observed_with_int<-(BD_final_distrito_MI$N_admissions/
                                   BD_final_distrito_MI$Total_inhabitants)*1000

rate_admission_fitted_with_int<-
              (model_int_type_II$summary.fitted.values$mean/
              BD_final_distrito_MI$Total_inhabitants)*1000

rate_observed_fitted_with_int<-c(rate_admission_observed_with_int,
                                 rate_admission_fitted_with_int)
BD_observed_fitted_rate_with_int<-data.frame(rep(rep(0:17,each=5),2),
                                 rep(BD_final_distrito_MI$District,2),
                                 c(rep(c("2010_observed","2012_observed",
                                         "2014_observed","2016_observed",
                                         "2018_observed"),18),
                                   rep(c("2010_fitted","2012_fitted",
                                         "2014_fitted","2016_fitted",
                                         "2018_fitted"),18)),
                                 rate_observed_fitted_with_int)
names(BD_observed_fitted_rate_with_int)<-c("id_maps","District","Year",
                                           "Rate_admission")

BD_observed_fitted_rate_with_int_maps<-left_join(Dist_mapas_dt,
                                           BD_observed_fitted_rate_with_int,
                                           by=c("id"="id_maps"))

ggplot(data=BD_observed_fitted_rate_with_int_maps, aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=Rate_admission),
               colour="black")+
```

```
  coord_equal() +
  labs(fill = "")+
  scale_fill_gradient(low="blue", high="yellow")+
  facet_wrap(~BD_observed_fitted_rate_with_int_maps$Year)
```

D.7 Diagnosis of the spatio-temporal models chosen at district level

```
model_year_rw2 <- inla(formula_year_rw2,family="poisson",
                       data=BD_final_distrito_MI,
                       offset=log(Total_inhabitants),
                       control.predictor=list(link=1,compute=TRUE),
                       control.fixed=list(mean=0,prec=0.00001),
                       control.compute=list(return.marginals.predictor=TRUE))

model_int_type_II <- inla(formula_int_type_II,family="poisson",
                          data=BD_final_distrito_MI,
                          offset=log(Total_inhabitants),
                          control.predictor=list(link=1,compute=TRUE),
                          control.fixed=list(mean=0,prec=0.00001),
                         control.compute=list(return.marginals.predictor=TRUE))

par(mfrow=c(1,2))

#SPATIO-TEMPORAL WITHOUT INTERACTION
plot(BD_final_distrito_MI$N_admissions,
     model_year_rw2$summary.fitted.values$mean,
     xlab="Observed values",ylab="Mean Post. Pred. Distr.")


predicted.p.value<-c()
for(i in (1:90)) {
 predicted.p.value[i]<-inla.pmarginal(q=BD_final_distrito_MI$N_admissions[i],
                         marginal=model_year_rw2$marginals.fitted.values[[i]])
}
hist(predicted.p.value,main="",
     xlab="Posterior predictive p-value",
     breaks=5,xlim=c(0,1.0),ylim=c(0,50))

#SPATIO-TEMPORAL WITH INTERACTION
plot(BD_final_distrito_MI$N_admissions,
     model_int_type_II$summary.fitted.values$mean,
     xlab="Observed values",ylab="Mean Post. Pred. Distr.")
```

```
predicted.p.value<-c()
for(i in (1:90)) {
 predicted.p.value[i]<-inla.pmarginal(q=BD_final_distrito_MI$N_admissions[i],
                     marginal=model_int_type_II$marginals.fitted.values[[i]])
}
hist(predicted.p.value,main="",
     xlab="Posterior predictive p-value",
     breaks=5,ylim=c(0,35))
```

D.8 Spatio-temporal analysis without interaction at municipality level

D.8.1 Models

```
ID.area.conc<-rep(1:278,5)
ID.year.conc<-rep(1:5,each=278)
ID.year1.conc<-ID.year.conc

formula_year_iid_conc<-N_admissions~scale(Density_pop)+
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Hospitals_1000)+
  scale(Prop_residents_65)+
  f(ID.area.conc,model="besag",graph = Conc.adj)+f(ID.year.conc,model="iid")
model_year_iid_conc <- inla(formula_year_iid_conc,family="poisson",
                                data=BD_final_concelhos_MI,
                                offset=log(Total_inhabitants),
                                control.predictor=list(compute=TRUE),
                                control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_year_iid_conc)

formula_year_rw1_conc<-N_admissions~scale(Density_pop)+
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Hospitals_1000)+
  scale(Prop_residents_65)+
  f(ID.area.conc,model="iid",graph = Conc.adj)+f(ID.year.conc,model="rw1")
model_year_rw1_conc <- inla(formula_year_rw1_conc,family="poisson",
                            data=BD_final_concelhos_MI,
                            offset=log(Total_inhabitants),
                            control.predictor=list(compute=TRUE),
                            control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_year_rw1_conc)

formula_year_rw2_conc<-N_admissions~scale(Density_pop)+
```

```
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Hospitals_1000)+
  scale(Prop_residents_65)+
  f(ID.area.conc,model="besag",graph = Conc.adj)+f(ID.year.conc,model="rw2")
model_year_rw2_conc <- inla(formula_year_rw2_conc,family="poisson",
                            data=BD_final_concelhos_MI,
                            offset=log(Total_inhabitants),
                            control.predictor=list(compute=TRUE),
                            control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_year_rw2_conc)

formula_year_iid_rw1_conc<-N_admissions~scale(Density_pop)+
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Hospitals_1000)+
  scale(Prop_residents_65)+
  f(ID.area.conc,model="besag",graph = Conc.adj)+f(ID.year.conc,model="rw1")+
  f(ID.year1.conc,model="iid")
model_year_iid_rw1_conc <- inla(formula_year_iid_rw1_conc,family="poisson",
                                data=BD_final_concelhos_MI,
                                offset=log(Total_inhabitants),
                                control.predictor=list(compute=TRUE),
                                control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_year_iid_rw1_conc)

formula_year_iid_rw2_conc<-N_admissions~scale(Density_pop)+
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Hospitals_1000)+
  scale(Prop_residents_65)+
  f(ID.area.conc,model="besag",graph = Conc.adj)+f(ID.year.conc,model="rw2")+
  f(ID.year1.conc,model="iid")
model_year_iid_rw2_conc <- inla(formula_year_iid_rw2_conc,family="poisson",
                                data=BD_final_concelhos_MI,
                                offset=log(Total_inhabitants),
                                control.predictor=list(compute=TRUE),
                                control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_year_iid_rw2_conc)
```

D.8.2 Graphs of the random effect models

```
#Map of spatial random effect
id_maps <-rep(0:277)
posterior_spatial_random_years_conc<-
        model_year_iid_conc$summary.random$ID.area$mean
```

```
BD_spatial_random_effect_years_conc<-data.frame(id_maps=as.factor(id_maps),
        posterior_spatial_random_years_conc)
BD_spatial_random_effect_years_conc_maps<-left_join(Mun_mapas_dt,
        BD_spatial_random_effect_years_conc,by=c("id"="id_maps"))
summary(posterior_spatial_random_years_conc)

ggplot(data=BD_spatial_random_effect_years_conc_maps,aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=posterior_spatial_random_years_conc),
               colour="black")+
  xlab("Longitude") + ylab("Latitude") +
  labs(fill = "")+
  scale_fill_gradientn(colours=tim.colors(99),limits=c(-0.60,0.65))+
  coord_fixed(1.1)

#Map of temporal random effect
years<-c(2010,2012,2014,2016,2018)

plot(years,model_year_iid_conc$summary.random$ID.year$mean,
     xlab="t",ylab=expression(paste(phi[t])),cex=0,cex.lab=1.5,cex.axis=1.25)
lines(years,model_year_iid_conc$summary.random$ID.year$mean,
      xlab="t",ylab=expression(paste(phi[t])),cex=0)
```

D.8.3 Graphs of the observed and fitted values

```
rate_admission_observed_temp<-(BD_final_concelhos_MI$N_admissions/
                               BD_final_concelhos_MI$Total_inhabitants)*1000

rate_admission_fitted_temp<-(model_year_iid_conc$summary.fitted.values$mean/
                               BD_final_concelhos_MI$Total_inhabitants)*1000

rate_observed_fitted_without_int<-c(rate_admission_observed_temp,
                                    rate_admission_fitted_temp)
BD_observed_fitted_rate_temp<-data.frame(rep(0:277,10),
                               rep(BD_final_concelhos_MI$Municipalities,2),
                               c(rep(c("2010_observed","2012_observed",
                                       "2014_observed","2016_observed",
                                       "2018_observed"),each=278),
                                 rep(c("2010_fitted","2012_fitted",
                                           "2014_fitted","2016_fitted",
                                           "2018_fitted"),each=278)),
                               rate_observed_fitted_without_int)

names(BD_observed_fitted_rate_temp)<-c("id_maps","Municipalities","Year",
```

```
                                        "Rate_admission")

BD_observed_fitted_rate_temp_maps<-left_join(Mun_mapas_dt,
                                             BD_observed_fitted_rate_temp,
                                             by=c("id"="id_maps"))

ggplot(data=BD_observed_fitted_rate_temp_maps, aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=Rate_admission),
               colour="black")+
  coord_equal() +
  labs(fill = "")+
  scale_fill_gradientn(colours=tim.colors(99))+
  facet_wrap(~BD_observed_fitted_rate_temp_maps$Year)
```

D.9 Spatio-temporal analysis with interaction at municipality level

D.9.1 Models

```
ID.area.year.conc<-1:(278*5) #278 municipalities x 5 years
ID.area.int.conc<-ID.area.conc
ID.year.int.conc<-ID.year.conc

formula_int_type_I<-N_admissions~Doctors_1000+Mort_rate_circ+Nurses_1000+
  Mort_rate_diabetes+Density_pop+Hospitals_1000+Pharmacies_1000+Prop_males+
  Prop_residents_65+f(ID.area,model="iid",graph = Conc.adj)+
  f(ID.year,model="rw2")+f(ID.area.year,model="iid")
model_int_type_I <- inla(formula_int_type_I,family="poisson",
                         data=BD_final_concelhos_MI,
                         offset=log(Total_inhabitants),
                         control.predictor=list(compute=TRUE),
                         control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_int_type_I)

formula_int_type_II<-N_admissions~Doctors_1000+Mort_rate_circ+Nurses_1000+
  Mort_rate_diabetes+Density_pop+Hospitals_1000+Pharmacies_1000+Prop_males+
  Prop_residents_65+f(ID.area,model="iid",graph = Conc.adj)+
  f(ID.year,model="rw2")+f(ID.area.int,model="iid",group=ID.year.int,
                           control.group = list(model="rw2"))
model_int_type_II <- inla(formula_int_type_II,family="poisson",
                          data=BD_final_concelhos_MI,
                          offset=log(Total_inhabitants),
                          control.predictor=list(compute=TRUE),
                          control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_int_type_II)
```

```
formula_int_type_III<-N_admissions~scale(Density_pop)+
  scale(Unemployment_rate)+scale(Prop_males)+
  scale(Mort_rate_diabetes_1000)+scale(Hospitals_1000)+
  scale(Prop_residents_65)+
  f(ID.area.conc,model="besag",graph = Conc.adj)+
  f(ID.year.conc,model="iid")+f(ID.year.int.conc,model="iid",
                           group=ID.area.int.conc,
                           control.group = list(model="besag",
                           graph=Conc.adj))
model_int_type_III <- inla(formula_int_type_III,family="poisson",
                          data=BD_final_concelhos_MI,
                          offset=log(Total_inhabitants),
                          control.predictor=list(compute=TRUE),
                          control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_int_type_III)

formula_int_type_IV<-N_admissions~Doctors_1000+Mort_rate_circ+Nurses_1000+
  Mort_rate_diabetes+Density_pop+Hospitals_1000+Pharmacies_1000+Prop_males+
  Prop_residents_65+f(ID.area.conc,model="besag",graph = Conc.adj)+
  f(ID.year.conc,model="rw1")+f(ID.area.int.conc,model="besag",
                           graph = Conc.adj,
                           group=ID.year.int.conc,
                           control.group = list(model="rw1"))
model_int_type_IV <- inla(formula_int_type_IV,family="poisson",
                          data=BD_final_concelhos_MI,
                          offset=log(Total_inhabitants),
                          control.predictor=list(compute=TRUE),
                          control.compute=list(dic=TRUE,cpo=TRUE))
summary(model_int_type_IV)
```

D.9.2 Graphs of the random effect models

```
#Map of spatial random effect
id_maps <-rep(0:277)
posterior_spatial_random_int_conc<-
    model_int_type_III$summary.random$ID.area.conc$mean
BD_spatial_random_effect_int_conc<-data.frame(id_maps=as.factor(id_maps),
                                      posterior_spatial_random_int_conc)
BD_spatial_random_effect_int_conc_maps<-left_join(Mun_mapas_dt,
                                       BD_spatial_random_effect_int_conc,
                                       by=c("id"="id_maps"))
summary(posterior_spatial_random_int_conc)
```

```
ggplot(data=BD_spatial_random_effect_int_conc_maps,aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=posterior_spatial_random_int_conc),
               colour="black")+
  xlab("Longitude") + ylab("Latitude") +
  labs(fill = "")+
  scale_fill_gradientn(colours=tim.colors(99),limits=c(-0.37,0.62))+
  coord_fixed(1.1)

#Map of temporal random effect
years<-c(2010,2012,2014,2016,2018)

plot(years,model_int_type_III$summary.random$ID.year.conc$mean,
     xlab="t",ylab=expression(paste(phi[t])),cex=0,cex.lab=1.5,cex.axis=1.25)
lines(years,model_int_type_III$summary.random$ID.year.conc$mean,
      xlab="t",ylab=expression(paste(phi[t])),cex=0)

#Map for interaction random effect
posterior_interaction_random_conc<-
      model_int_type_III$summary.random$ID.year.int.conc$mean

BD_posterior_interaction_conc<-data.frame(rep(0:277,5),
                                          BD_final_concelhos_MI$Municipalities,
                                          BD_final_concelhos_MI$Year,
                                          posterior_interaction_random_conc)
names(BD_posterior_interaction_conc)<-c("id_maps","Municipalities","Year",
                                        "Posterior_interaction")

BD_posterior_interaction_conc_maps<-left_join(Mun_mapas_dt,
                                              BD_posterior_interaction_conc,
                                              by=c("id"="id_maps"))
summary(BD_posterior_interaction_conc)

ggplot(data=BD_posterior_interaction_conc_maps, aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=Posterior_interaction),
               colour="black")+
  coord_equal() +
  labs(fill = "")+
  scale_fill_gradientn(colours=tim.colors(99))+
  facet_wrap(BD_posterior_interaction_conc_maps$Year)
```

D.9.3 Graphs of the observed and fitted values

```
rate_admission_observed_int<-(BD_final_concelhos_MI$N_admissions/
                                BD_final_concelhos_MI$Total_inhabitants)*1000

rate_admission_fitted_int<-(model_int_type_III$summary.fitted.values$mean/
                        BD_final_concelhos_MI$Total_inhabitants)*1000

rate_observed_fitted_without_int<-c(rate_admission_observed_int,
                                     rate_admission_fitted_int)
BD_observed_fitted_rate_int<-data.frame(rep(0:277,10),
                              rep(BD_final_concelhos_MI$Municipalities,2),
                                 c(rep(c("2010_observed","2012_observed",
                                 "2014_observed","2016_observed",
                                 "2018_observed"),each=278),
                                 rep(c("2010_fitted","2012_fitted",
                                 "2014_fitted","2016_fitted",
                                 "2018_fitted"),each=278)),
                              rate_observed_fitted_without_int)

names(BD_observed_fitted_rate_int)<-c("id_maps","Municipalities","Year",
                                     "Rate_admission")

BD_observed_fitted_rate_int_maps<-left_join(Mun_mapas_dt,
                                            BD_observed_fitted_rate_int,
                                            by=c("id"="id_maps"))

ggplot(data=BD_observed_fitted_rate_int_maps, aes(x=long,y=lat))+
  geom_polygon(aes(group=group, fill=Rate_admission),
               colour="black")+
  coord_equal() +
  labs(fill = "")+
  scale_fill_gradientn(colours=tim.colors(99))+
  facet_wrap(~BD_observed_fitted_rate_int_maps$Year)
```

D.10 Diagnosis of the spatio-temporal models chosen at municipality level

```
model_year_iid_conc <- inla(formula_year_iid_conc,family="poisson",
                            data=BD_final_concelhos_MI,
                            offset=log(Total_inhabitants),
                            control.predictor=list(link=1,compute=TRUE),
                            control.fixed=list(mean=0,prec=0.00001),
                    control.compute=list(return.marginals.predictor=TRUE))
```

```
model_int_type_III <- inla(formula_int_type_III,family="poisson",
                           data=BD_final_concelhos_MI,
                           offset=log(Total_inhabitants),
                           control.predictor=list(link=1,compute=TRUE),
                           control.fixed=list(mean=0,prec=0.00001),
                     control.compute=list(return.marginals.predictor=TRUE))

par(mfrow=c(1,2))

#SPATIO-TEMPORAL WITHOUT INTERACTION
plot(BD_final_concelhos_MI$N_admissions,
     model_year_iid_conc$summary.fitted.values$mean,
     xlab="Observed values",ylab="Mean Post. Pred. Distr.")


predicted.p.value<-c()
for(i in (1:1390)) {
 predicted.p.value[i]<-inla.pmarginal(q=BD_final_concelhos_MI$N_admissions[i],
                   marginal=model_year_iid_conc$marginals.fitted.values[[i]])
}
hist(predicted.p.value,main="",
     xlab="Posterior predictive p-value",
     breaks=11,xlim=c(0,1.0),ylim=c(0,600))

#SPATIO-TEMPORAL WITH INTERACTION
plot(BD_final_concelhos_MI$N_admissions,
     model_int_type_III$summary.fitted.values$mean,
     xlab="Observed values",ylab="Mean Post. Pred. Distr.")

predicted.p.value<-c()
for(i in (1:1390)) {
 predicted.p.value[i]<-inla.pmarginal(q=BD_final_concelhos_MI$N_admissions[i],
                    marginal=model_int_type_III$marginals.fitted.values[[i]])
}

hist(predicted.p.value,main="",
     xlab="Posterior predictive p-value",
     breaks=11,ylim = c(0,400))
```

Attachments

A Maps of Portugal

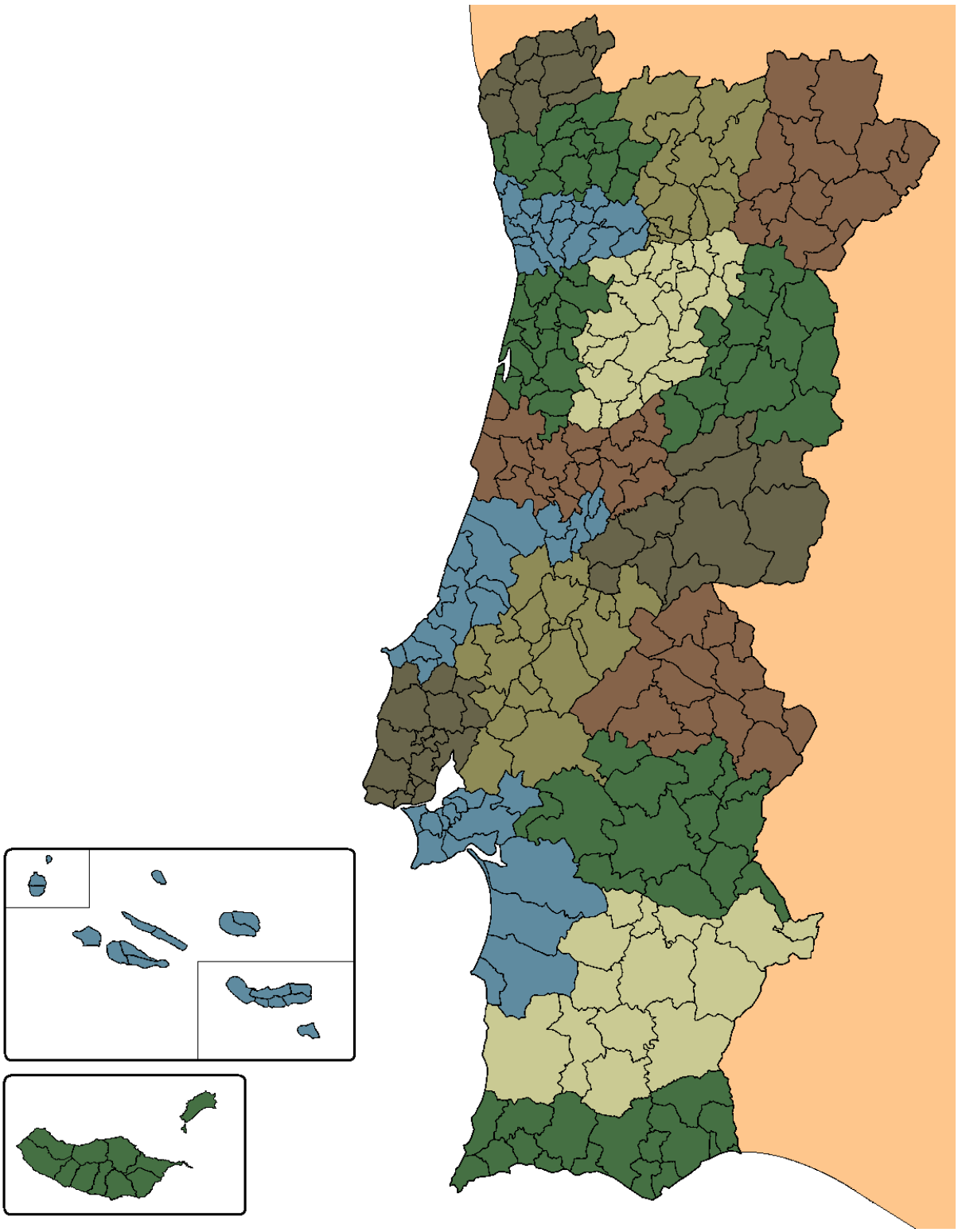

Figure 17: Municipalities of Portugal.