technische universität dortmund

DOCTORAL THESIS

Spatial and Spatio-temporal Regression Modelling with Conditional Autoregressive Random Effects for Epidemiological and Spatially Referenced Data

Author:
Dany-Armand DJEUDEU-DEUDJUI

Supervisor:
Prof. Dr. Katja ICKSTADT

Chair: Prof. Dr. Susanne MOEBUS
Referee: Prof. Dr. Katja ICKSTADT
Referee: Prof. Dr. Philipp DOEBLER
Assistant: Prof. Dr. Jörg RAHNENFÜHRER

A thesis submitted in fulfillment of the requirements for the degree of Dr. rer. nat.

in the

Department of Statistics

July 19, 2022

This thesis is dedicated to the memory of my lovely brother Siewe Deudjui Emmanuel Therry.

Contents

Declaration of Authorship

I, Dany-Armand DJEUDEU-DEUDJUI, declare that this thesis titled, "Spatial and Spatio-temporal Regression Modelling with Conditional Autoregressive Random Effects for Epidemiological and Spatially Referenced Data" and the work presented there in are my own. I confirm that:

- This work was carried out wholly or mainly while in my candidature for a research degree at this University.
- If any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institutions, I have stated it clearly wherever appropriate.
- Where I have consulted the published works of others, this is always clearly referenced in the thesis.
- Where references were made from the works of other authors, the source is always given.
- I have acknowledged all main sources of help.
- Where the thesis is based on works done in cooperation with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Abstract

Regression models are suitable to analyse the association between health outcomes and environmental exposures. However, in urban health studies where spatial and temporal changes are of importance, spatial and spatio-temporal variations are usually neglected. This thesis develops and applies regression methods incorporating latent random effects terms with Conditional Autoregressive (CAR) structures in classical regression models to account for the spatial effects for cross-sectional analysis and spatio-temporal effects for longitudinal analysis. The thesis is divided into two main parts.

Firstly, methods to analyse data for which all variables are given on an areal level are considered. The longitudinal Heinz Nixdorf Recall Study is used throughout this thesis for application. The association between the risk of depression and greenness at the district level is analysed. A spatial Poisson model with a latent CAR structured-Random effect is applied for selected time points. Then, a sophisticated spatio-temporal extension of the Poisson model results to a negative association between greenness and depression. The findings also suggest strong temporal autocorrelation and weak spatial effects. Even if the weak spatial effects are suggestive of neglecting them, as in the case of this thesis, spatial and spatio-temporal random effects should be taken into account to provide reliable inference in urban health studies.

Secondly, to avoid ecological and atomic fallacies due to data aggregation and disaggregation, all data should be used at their finest spatial level given. Multilevel Conditional Autoregressive (CAR) models help to simultaneously use all variables at their initial spatial resolution and explain the spatial effect in epidemiological studies. This is especially important where subjects are nested within geographical units. This second part of the thesis has two goals. Essentially, it further develops the multilevel models for longitudinal data by adding existing random effects with CAR structures that change over time. These new models are named MLM tCARs. By comparing the MLM tCARs to the classical multilevel growth model via simulation studies, we observe a better performance of MLM tCARs in retrieving the true regression coefficients and with better fits. The models are comparatively applied on the analysis of the association between greenness and depressive symptoms at the individual level in the longitudinal Heinz Nixdorf Recall Study. The results show again negative association between greenness and depression and a decreasing linear individual time trend for all models. We observe once more very weak spatial variation and moderate temporal autocorrelation. Besides, the thesis provides comprehensive decision trees for analysing data in epidemiological studies for which variables have a spatial background.

Acknowledgements

I would like to express my gratitude to Prof. Dr. Katja Ickstadt and Prof. Dr. Susanne Moebus for their invaluable advice, continuous support, and patience during my PhD study and for giving me this opportunity. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. I am grateful to all members of the chair of mathematical statistics with applications in Biometrics of the faculty of statistics of TU Dortmund University and the members of the institute of urban public health of the University Hospital Essen for the helpful discussions and support in the past few years. Also, I would like to thank my family and friends for always supporting me in a way I cannot even express. In particular, I thank my parents and wife for their unconditional support.

List of Figures

List of Tables

List of Abbreviations

CAR	**C**onditional **A**utoregressive
HNRS	**H**einz **N**ixdorf **R**ecall **S**tudy
RCAR	**R**estricted **C**onditional **A**utoregressive
MLM	**M**ulti**l**evel **M**odel
MLM CL2	**C**lassical **M**ulti**l**evel **M**odel, **T**wo levels
MLM CL3	**C**lassical **M**ulti**l**evel (growth) **M**odel, **T**hree levels
MLM CONV	**M**ulti**l**evel **C**onvolution **M**odel
MLM CARANOVA	**M**ulti**l**evel **CARANOVA** **M**odel
MLM tCARs	**M**ulti**l**evel **C**onditional **A**utoregressive **M**odels for **L**ongitudinal **D**ata
MLM CAR	**M**ulti**l**evel **C**onditional **A**utoregressive **M**odel
MLM RCAR	**M**ulti**l**evel **R**estricted **C**onditional **A**utoregressive **M**odel
MLM CARs	**M**ulti**l**evel **C**onditional **A**utoregressive **M**odels for cross-sectional **D**ata
MRF	**M**arkov **R**andom **F**ields
DIC	**D**eviance **I**nformation **C**riterion
SAS	**S**tatistical **A**nalysis **S**ystem
R	The **R** Project for Statistical Computing
AIC	**A**kaike **I**nformation **C**riterion
BIC	**B**ayesian **I**nformation **C**riterion
SIR	**S**tandardized **I**ncidence **R**atio
SIRs	**S**tandardized **I**ncidence **R**atios
INLA	**I**ntegrated **N**ested **L**aplace **A**pproximation
CES-D	**C**enter for **E**pidemiologic **S**tudies **D**epression Scale
ICC	**I**ntra-class **C**orrelation **C**oefficient
HLM	**H**ierarchical **L**inear **M**odels
NDVI	**N**ormalized **D**ifference **V**egetation **I**ndex
BMI	**B**ody **M**ass **I**ndex

Chapter 1

Introduction

1.1 Motivation

In several observational epidemiological studies, regression models are applied to analyse the association between risk factors and health outcomes (Weber et al., 2019, Waller and Gotway, 2004, Sondermann et al., 2020). One prominent example in this thesis might be in particular the longitudinal Heinz Nixdorf Recall Study (HNRS). The presence of some environmental risk factors with spatial background induces spatial variation. Regression models should provide much more substantial conclusions than could be obtained using just visualisation of the spatial distribution of health outcomes. It allows us to adjust analyses for important covariate information. They also allow us to more explicitly use spatial exposure measurements to help describe the spatial distribution of public health events. For the data of the HNRS, longitudinal as well as cross-sectional analyses are carried out (Weber et al., 2019, Tzivian et al., 2016). Cross-sectional studies offer a snapshot of a single moment in time and can be done very quickly. Cross-sectional analyses are sometimes preferred for simplicity, applicability, and accessibility in many software packages. However, this type of study design may not provide definite information about cause-and-effect relationships. A longitudinal design unlike a cross-sectional design may allow the direct assessment of changes in the response variable over time and help examine if the changes are related to the selected covariates. That's why researchers might start with a cross-sectional study to first establish whether there are links or associations between certain variables. Then they would set up a longitudinal analysis for a more meaningful and robust result. In either case, as for the HNRS, risk factors are collected both at the individual level (e.g., age, Body Mass Index) as well as group or areal level (e.g., noise, unemployment in district). It is important to model all the characteristics and influencing variables.

For a quick overview, simplicity, or other analysis goals, data are sometimes aggregated or disaggregated to the same spatial resolution in order to apply usual regression methods. Treating for instance group-level variables as a repeated measurement, whereby all individuals within the same group are assigned identical values of group variable, will lead to inefficient model estimation and incorrect statistical inference (Julian, 2001). Also for simplicity, all area-level variables are sometimes left aside to concentrate on the individual-level covariates to explain the individual-level health outcomes. This approach has the drawback of ignoring the potential importance of group-level attributes in influencing individual-level health outcome. In the same way, relying only on group-level characteristics to explain group-level (aggregated) health outcomes may lead to bias in the analysis. Just as studies examining differences between groups may need to take into account possible differences in group compositions (i.e. characteristics of individuals within them), studies of individuals may need to take into account differences in the properties of the groups to which individuals belong (Roux, 1998).

For the usual regression methods with data on the same spatial resolution as well as methods that consider different spatial resolutions, spatial and spatio-temporal variation should be taken into account.

1.2 Spatial and spatio-temporal variation

A common belief is that the presence of residual spatial autocorrelation in classical regression (Ordinary Least Squares) leads to inflated significance levels in regression coefficients (Smith and Lee, 2012). More precisely, the failure to account for the spatial variation in the model causes the p-values, as well as the standard errors of the parameter estimates to be artificially smaller, showing a stronger association between exposures and health outcomes than it would really be (in the presence of spatial dependency). Furthermore, environmental resources are often unequally (spatially) distributed across different socioeconomic groups and this may cause grouping effects that affect the behavior of regression analysis. There are mainly two types of spatial effects when it comes to investigating the effects of the built environment or any exposure on health outcomes: The first is spatial autocorrelation, closely related to the first law of geography (Tobler, 1979, Waller and Gotway, 2004). Spatial autocorrelation can arise for a number of reasons, for instance, due to unmeasured/unavailable confounding. Spatial autocorrelation occurs when a spatially patterned risk factor for the response variable is not included in a regression model, and hence its omission induces spatial structure into the residuals. Because confounding might be a complex and unknown function of many covariates, there is rarely enough information in nonexperimental data to allow one to construct an acceptably accurate estimate of exposure's effect from the data alone (Robins and Greenland, 1986). Other causes of spatial autocorrelation that need to be accounted for include neighborhood effects, where the behaviors of individuals in a spatial unit are influenced by individuals and characteristics in adjacent units. The second type of spatial effect (variation) that needs to be accounted for is the grouping effect or spatial heterogeneity, where groups of people with similar behaviors choose to live close together; for example, you might find a higher prevalence of smoking or overweights in one neighbourhood compared to another one, because of the interrelations and preferences of the people in one neighbourhood or in other words: individuals with similar lifestyles and preferences might choose to live closer together. The individuals in one neighbourhood often experience the same environmental conditions, like noise or air pollution. Sometimes, the difference between the two types of spatial effects is not obvious.

1.3 Aims and outline

The aims of this thesis are twofold: the first is to encourage using sophisticated regression models that account for spatial and spatio-temporal variation to analyse the association between health outcomes and environmental exposures. We particularly suggest including a spatial and spatio-temporal latent random effect with Conditional Autoregressive (CAR) structure. The spatio-temporal CAR structure should account for the change of spatial structure in health outcomes. This will firstly be performed in our application for analyses with all variables given at an areal level. As data aggregation and disaggregation may lead to ecological and atomic fallacies (Wakefield, 2009), respectively, we wish to use all data on their finest level given. Thus, the second goal of this thesis consists of developing a method that combines the advantages of multilevel models and the CAR models for longitudinal studies. Since this work is the first for the HNRS and many longitudinal studies that may account for the spatial effects (either spatial heterogeneity or spatial autocorrelation), we always perform an exploratory spatial analysis (for particular time points) prior to the spatio-temporal analysis to better understand the structure of the spatial random effect before investigating the structure of the spatio-temporal random effects.

The chapters of this thesis are divided into three parts. Part I, partly based on the paper (Djeudeu et al., 2020), with two chapters concerns methods and analyses at an areal level. Part II, mainly based on the paper (Djeudeu, Moebus, and Ickstadt, 2022), with two chapters considers methods to deal with data on their initial spatial resolution. Part III with one chapter concerns practical recommendations based on the first two parts. In more details, Chapter 2 (first chapter of Part I) is devoted to methods for cross-sectional data. Spatial models for disease mapping in general are of interest, particularly, models with latent random effects with CAR structured prior to account for spatial effects. In this Chapter 2, we apply a spatial Poisson model to the analysis of the risk of depression in the districts of the HNRS and the association with greenness. In Chapter 3 (second chapter of Part I), a sophisticated spatio-temporal extension of the spatial Poisson model in Chapter 2 is applied on the longitudinal HNRS to the analysis of the association between greenness and depression. The model accounts for the spatio-temporal effect and the dynamic over time, when analysing disease risks at an areal level. Part II with two chapters is dedicated to multilevel models to use all variables at their finest spatial resolution. In Chapter 4 (first chapter of Part II), multilevel CAR models for cross sectional data (MLM CARs), already developed in the literature, are considered, compared to the classical multilevel model in a simulation Study. In Chapter 5 (second chapter of Part II), multilevel CAR models for longitudinal data are developed by combining the multilevel models and some particular Markov Random Fields (MRF) models for the area-level random effect that accounts for the change of spatial effect over time. Part III consists of Chapter 6 that describes a decision tree to help choose the appropriate methods of analysis when participants are nested within geographical units. To end, an overall conclusion is provided in Chapter 7 to summarize the thesis and provide outlook for future analyses. Additional materials including supplementary figures and tables, R-codes used in this thesis and some full conditionals for regression coefficients for Gibbs sampler of the developed model are provided at the end of the thesis.

Part I

Spatial regression models for areal data

Chapter 2

Spatial analysis of the risk of depression at district-level and association with greenness

2.1 Introduction

With increasing urbanization the importance of in-depth knowledge of the spatial distribution of diseases and risk factors on a local or regional level is gaining acceptance worldwide (Galea and Vlahov, 2005, WHO, 2016). Places where people live can be of great importance in identifying patterns of disease and associations with risk factors (Waller and Gotway, 2004), particularly when environmental risk factors are involved. An increasing number of studies investigating effects of the built environment on health (Orban et al., 2016, Orban et al., 2017, Wu and Jackson, 2017) try to identify etiological risk factors and disease-specific processes in epidemiological studies. For instance, according to the WHO (World Health Organization), depressive disorders are among the most important common non-communicable diseases and a leading cause of disability worldwide (WHO, 2018). Urban green is an essential part of the urban ecosystem, promoting health and well-being (WHO, 2018). Recent studies suggest evidence of the importance of green spaces for mental health. Green spaces improve well-being or reduce physiological stress indicators (Orban et al., 2016, White et al., 2013, Tomita et al., 2017). However, (substantial) spatial analyses of disease to identify areas with elevated or low disease risk are still not currently in practice in many studies. At the same time, spatial variations are usually neglected in regression models to analyse the association between health outcomes and exposures with a spatial background such as greenness and air pollution.

Statistical methods for spatially referenced data should consider the spatial autocorrelation as well as the spatial heterogeneity. This will provide more accurate and meaningful conclusions. Ignoring the spatial structure in the data runs the risk of violating the usual assumption of independent observations in ordinary regression analysis and may cause bias in the covariate effects (Cressie, 1993, Pfeiffer et al., 2008).

In this chapter, we would like to address this issue by using as an example the analysis of the effect of urban greenness on depressive symptoms from the data of the German HNRS. The analysis is performed exploratory prior to the spatio-temporal analysis in Chapter 3. The districts are the spatial level of interest. In the HNRS, individual and district level based data are available for different time points. As we aggregate individual-level depression data on the district level, we are aware of the possibility of a loss of information, particularly if the individual-level association is explored with aggregated data, this will lead to an "ecological fallacy". However, analysis at the individual level is not always preferable (Wen et al., 2001). For instance, aggregating data might avoid the so-called "atomic fallacy", the result of improper interpretation and inference about aggregate level association on the basis of associations at the individual level.

Previous analyses of the association between mental health outcome (e.g. depression) and greenness were mainly performed on individual-level variables, with spatial level covariates disaggregated to the individual level (Song et al., 2019, Beyer et al., 2014). The few ecological analyse on aggregated level did not use the spatial random effects to account for the unexplained spatial variation (Nutsford, Pearson, and Kingham, 2013). For the HNRS, linking health with greenness was analyzed in Orban et al., 2017. This analysis was done at the individual level and the spatial correlation was also not accounted for.

The main objective of this chapter is to exploratory analyse the risk of depression for some selected time points of the HNRS. Then we apply a refined spatial Poisson model on the analysis of the association between greenness and depression on the district level in the HNRS. Our approach would help to understand the distribution of the risk of depression and its influencing factors. The method takes simultaneously into account both spatial autocorrelation and spatial heterogeneity. Accordingly, spatial random effects are introduced while smoothing disease risks at a spatial level.

We organized our chapter as follows: We firstly describe of our underlying data set, data aggregation, and imputation. The description for the data set is general, since the same data set will be used in upcoming chapters for application. Secondly, we deal with statistical methods in which we proceed step by step: we firstly present a method (**Besag** and **Newell** method Besag and Newell, 1991) to identify local clusters of elevated risk of depression in the study area. Then, we shortly present the **Moran**'s I statistic (Moran, 1950) to assess (global) clustering. A spatial model for disease mapping is then described within a Bayesian hierarchical model formulation (the **Besag-York-Mollié** model (Pfeiffer et al., 2008)) to estimate and smooth the risk of depression, accounting for covariate effects. Section 2.7 is dedicated to the results. In section 2.8, we discuss the results compared with other studies and future insights for analysis.

2.2 Data of the Heinz Nixdorf Recall Study

The HNRS is a population-based, prospective cohort study of the comparative value of modern risk stratification techniques for "hard" cardiac events to define appropriate methods for identifying high-risk subgroups in the general urban population. It is designed and powered to define the relative risk associated with the specific extent of coronary atherosclerosis measured using electron-beam computed tomography (EBCT)-derived coronary calcium quantities for myocardial infarction and cardiac death. Additionally, the predictive values of conventional cardiovascular risk factors, new candidate and socioeconomic risk factors, certain genetic polymorphisms, and direct signs of subclinical disease are examined with the ankle-brachial index, resting, and stress electrocardiograms, and determination of carotid artery intima-media thickness. Prospective clinical risk-benefit and health economic analyses are an inherent part of the study. Study findings with established clinical significance are reported to the participants, but the EBCT findings are withheld until the conclusion of the study. The **study population** of the HNRS was from the cities Essen, Mülheim and Bochum of the metropolitan Ruhr Area in Western Germany. The study design has been described in detail in Schmermund et al., 2002. The baseline was performed between 2000 and 2003, including 4814 randomly selected men and women aged 45 to 75 years. Individuals were eligible if their address was valid, they were not institutionalized, had sufficient knowledge of the German language, were not severely ill, and were able to be interviewed. Participants were invited to the study center three times each after 5 years (t_0, $2000 - 2003$; t_5, $2006 - 2008$; t_{10}, $2011 - 2015$). Data assessment included standardized computer-assisted personal interviews, clinical examination, comprehensive laboratory tests, and self-administered questionnaires. In addition, a yearly postal follow-up between the examinations was provided, which allows for more than 18 years of observational data. The HNRS was approved by the local ethics committees, and all participants gave informed consent before participation.

The HNRS included several health outcomes such as diabetes mellitus, asthma, tumor, blood pressure, depressive symptoms, obesity. One research area is focusing on detecting the link between the level of calcification and other diseases. There are indications that the level of calcification is generally not only an indication of an unhealthy lifestyle, but also of unfavourable environmental conditions and genetic factors.

Our main outcome in this thesis is **depressive symptoms**. For the current chapter, we consider **depression in districts**. Depressive symptoms during the previous week were assessed using the 15-item short-form questionnaire of the Center for Epidemiologic Studies Depression Scale (CES-D) (Hautzinger and Bailer, 2012, Radloff, 1977). The CES-D is a screening tool for measuring depressive symptoms; it has been validated in different populations and settings and is frequently used in health research (Radloff, 1977). The questionnaire was filled out by participants at the study center and by mailed follow-ups. For our analyses we used CES-D data of eight measurements, assessed between 2000 and 2013. Scores for the 15-item version range from 0 to 45, with a higher score indicating more and/or more frequent depressive symptoms. We used a CES-D score of ≥ 17 as cut point to classify participants as depressive (Radloff, 1977). For each district where HNRS participants were living, we counted the number of cases of depression. In this way we obtained a count at the district level: depression in districts.

The HNRS includes several risk factors and covariates such as smoke status, family status, age, citizenship, marital status, quality of life, socioecnomic status (income, education, ...). The exposure in this thesis is greenness. For this chapter, we particularly consider **greenness measured at the district level**. Exposure to green space is commonly measured either as surrounding greenness or access to green space (e.g. distance to nearest park), which are two different concepts. Here, the concept of surrounding greenness was used, which is typically measured within a certain area, often buffer around a residence or on district level (Orban et al., 2017). Either the percent green space derived from land-cover data or the Normalized Difference Vegetation Index (NDVI) derived from satellite imagery (Rhew et al., 2011) is used. We base our analysis on three-time points of satellite imagery data, 2006, 2009, 2013, and calculated respective the NDVI for all districts of the study area. NDVI is a measure of vegetation level based on reflectance and absorbance of red and near-infrared solar radiation. Values of NDVI range from -1 to $+1$. Those around -1 generally correspond to water, while values near 0 represent bare surfaces, e.g. rock, rooftops and roads, or very scarce or dry vegetation. Values approaching 1 represent dense vegetation, e.g. rainforest. Thus, a 1-unit difference in NDVI corresponds to the difference between a barren area of rock or stone and a rainforest.

All **covariates** in this chapter will be considered at the district level. Some are directly measured at the district level like the unemployment rate in districts, obtained from the local census authorities of the respective cities of Bochum, Essen, and Mülheim. The unemployment rate was calculated by dividing the number of unemployed in the area by the number of the economically active population (unemployed + working population). Unemployment is a strong indicator for deprivation (Reineveld, 1998) in a neighborhood and was used as an indicator of neighborhood-level socioeconomic status (SES). Other covariates are aggregated from individual level to district level: Socioeconomic (e.g., income), demographic (e.g., age), and medical history. Income was measured as the monthly household equivalent income, which was calculated by dividing the total household net income by a weighting factor for each household member, and was divided into four groups using sex-specific quartiles. Information on whether participants ever had a myocardial infarction, heart failure, stroke, diabetes mellitus, emphysema, asthma, cancer, rheumatism, slipped disc, or migraine (yes/no) at baseline was used to create a categorical variable indicating the number of comorbidities (0, 1, or $\geq$ 2). The body mass index (BMI) was calculated as [weight in kilograms/(height in meters)^2]. The effect of greenness on any health outcome may depend on the time participants spent in their district. The percentage of relocations (percentage of participants who moved during the observation period) is a covariate that accounts for that.

Our data are arranged in a particular way. They consists of $K = 108$ districts and $N = 8$ time points: the time points are indexed years 0, 5, 7, 8, 9, 10, 11, 12, respectively. For the response, the $K \times N$ matrix $\mathbf{Y}$ is filled by the number of cases of depression for the given district and the given year. We also have the expected count in each district, which is the count calculated under the constant risk hypothesis in the complete study area. NDVI measures of 2006 were assigned to years 0, 5, 7, NDVI measures of 2009 were assigned to years 8, 9, 10 and NDVI measures 2013 assigned to years 11, 12. This may make sense as the values of NDVI are not supposed to be changing quickly during a small time span.). For some covariates, like BMI and income, data from baseline visit (year 0), first (year 5), and second (year 10) follow up are available. For each of the mailed follow-up surveys, the value of the previous follow-up (year 0, year 5, or year 10) were used. For all other covariates, baseline data was used for analysis. We also choose simplicity for imputation. The missing covariate values (for a given district) are simply imputed by the mean of the available values for the given years. We note that the missing values for any variable were at most 2%. Table 2.1 describes the number of participants for each examination year with a depression score.

	Year 0	Year 5	Year 7	Year 8	Year 9	Year 10	Year 11	Year 12
Participants with depression score	4645	4193	2118	3478	3223	2948	2508	2091

TABLE 2.1: Number of participants with value of depression score for each examination year for the participants of the Heinz Nixdorf Recall Study.

2.3 Formal notations

$Y = (Y_1, \ldots, Y_K)$ resp. $E = (E_1, \ldots, E_K)$ denotes the vector of observations (count of depression) resp. expected counts for the K non-overlapping districts of the study region of the HNRS. $\bar{Y} = \frac{1}{K}\sum_{k=1}^{K} Y_k$ is the arithmetic mean of the $(Y_k)_{k=1\ldots K}$. N_k is the population size in district k. $N = \sum_{k=1}^{K} N_k$ is the population under risk in the HNRS for the selected time point.

$W = (w_{ij})_{i,j=1,\ldots,K}$ is the adjacency matrix. There are several types of adjacency matrices. In this thesis, we use the binary adjacency matrix which is more popular for CAR models. It is defined by:

$$w_{ij} = \begin{cases} 1, & \text{if district } i \text{ and district } j \text{ share a common border and } i \neq j\,; \\ 0, & \text{else.} \end{cases} \tag{2.1}$$

δ_i is the set of districts adjacent to district i. $X_k = (x_{k1}, \ldots, x_{kp})$ is a vector of p known covariates for district k.

2.4 Local cluster detection

Several methods to detect local clusters of elevated risks exploratory are available (see Waller and Gotway, 2004, Pages $200 - 266$). Here we shortly highlight the one used in our application, the **Besag** and **Newell** method (Besag and Newell, 1991), which we prefer because of its interpretability.

This approach assumes that the number of cases Y_k and the population size N_k, $k = 1 \ldots K$, in each district k are known. The first step is to fix the expected cluster size C, $C \geq 2$. The second step is to consider for instance a particular district A_1, then designate the other districts with increasing distance to A_1 with $A_2, \ldots, A_K$ and define the test statistic $M = \min\{j : D_j \geq C\}$, with $D_j = \sum_{k=1}^{j} Y_k$. A small value of M suggests a cluster around district A_1. The probability of whether the specified C cases are observed in fewer districts is evaluated. The following null hypothesis is tested for the area $k \in \{1 \ldots K\}$:

H_0: Y_k is uniformly distributed in the overall population under risk N with probability $r = \frac{Y^+}{N}$, with $Y^+ = \sum_{k=1}^{K} Y_k$. The choice of the cluster size is not unique and it is recommended to use several of them (see Gómez-Rubio, Ferrándiz-Ferragud, and López-Quílez, 2005 for details).

2.5 Global cluster detection

For summarizing the extent of observed spatial similarity between nearby districts, there exist several indices to describe the global strength of the spatial autocorrelation. See Waller and Gotway, 2004, Pages $200 - 266$. Here we shortly highlight the one used in our application, the **Moran**'s I statistic (Moran, 1950), once more because of its interpretability and popularity:

$$I = \left(\frac{1}{s^2}\right) \frac{\sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij}} \tag{2.2}$$

with $s^2 = \frac{1}{K} \sum_{j=1}^{K} (Y_j - \bar{Y})^2$.

With this test statistic I, we test for spatial dependence versus the null hypothesis of no spatial dependence of the $Y_k, k = 1, \ldots, K$. A positive value of I suggests that the pattern is clustered (close districts tend to have identical values) and a negative one implies that the pattern is regular (neighbouring districts tend to have different values). When there is no correlation between neighboring values, the expected value of I is $E(I) = -\frac{1}{K-1}$. See (Lee and Wong, 2001, pp. $79 - 80$) for details.

2.6 Spatial models for areal data including spatial random effects

The basic form of the data involves a set of counts observed (one count for each district) and a matching set of counts expected reporting the number of cases we expect in each district, under the null hypothesis. We keep the notations of Subsection 2.4.

2.6.1 The general model

The general model can be described as:

$$\begin{cases} Y_k & \sim f(\theta_k|X_k) \\ g(\theta_k) & = X_k^T\beta + \psi_k + C \\ \beta & \sim N(\mu_\beta, \Sigma_\beta), \end{cases} \tag{2.3}$$

where θ_k denotes the risk (of depression) in district k. $\beta = (\beta_1, \dots \beta_p)$ is the covariate regression parameter vector, and a multivariate Gaussian prior is assumed with mean μ_β and diagonal variance matrix Σ_β. $\psi = (\psi_1, \dots, \psi_K)$ consists of the random components ψ_k for district k, that is to be precisely described later. C is a constant term.

Possible link functions for g are the natural logarithm, the logit and the identity function, corresponding to the Poisson, Binomial and Gaussian models respectively. To be more concrete, we will consider the Poisson and Binomial model for more details.

2.6.2 CAR-structure prior for latent random effect

We use the **Besag-York-Mollié** model, also called convolution model, which is, e.g. fully described in Pfeiffer et al., 2008, for ψ_k. Here $\psi_k = U_k + V_k$, with unstructured district effect $V = (V_1 \dots V_K)$, $V_k \sim N(0, \tau_v^2)$, $k = 1, \dots K$. $U = (U_1 \dots U_K)$ represents the structured spatial effect between districts and is modelled using CAR priors (Besag, York, and Mollié, 1991). CAR models, a special specification of the MRF models are network-based models, specifically designed to model spatially autocorrelated data based on neighborhood relationships. CAR priors have the advantage of facilitating random-effects analysis under a Markov Chain Monte Carlo (MCMC) sampling approach. Mathematically, it is given as follows:

$$[U_i|U_j = u_j, j \in \delta_i, \tau^2] \sim N(m_i, \tau_i^2), \tag{2.4}$$

where $m_i = \dfrac{\sum_{j\in\delta_i} w_{ij}u_j}{\sum_{j\in\delta_i} w_{ij}}$ and $\tau_i^2 = \dfrac{\tau^2}{\sum_{j\in\delta_i} w_{ij}}$. τ^2 is the general location-independent variance.

For the random effect ψ_k and in contrast with the **Besag-York-Mollié** model, we could have used a single set of random effects to model both the structured and unstructured random effect as given by the **Leroux** model (Leroux, Lei, and Breslow, 2000). The spatio-temporal extension of the **Leroux** model will be applied in Chapter 3 to show the flexibility of the CAR models used.

2.6.3 The spatial Poisson model

For each of the 3 examination visits at the study center, we use a hierarchical Poisson model that accounts for covariate effects and where spatial autocorrelation is modelled via sets of autocorrelated random effects. The model equation is given by

$$\begin{cases} Y_k & \sim Poisson(\theta_k E_k) \\ ln(\theta_k) & = X_k^T \beta + \psi_k \\ \beta & \sim N(\mu_\beta, \Sigma_\beta). \end{cases} \tag{2.5}$$

The term $X_k^T \beta$ represents the covariate effects, while θ_k is the risk of depression in district k. ψ_k is defined as given in Subsection (2.6.2). Neglecting the spatial effects as well as the covariate effects in equation (2.5) results in the classical Poisson model for count data. For the classical Poisson model, the maximum likelihood estimate of the risk is the ratio of observed cases to the expected cases in each district. It is called the Standardized Incidence Ratio (SIR). A SIR significantly greater (resp. smaller) than 1 indicates an increased (resp. decreased) risk of disease. The estimated Risk from model 2.5, including covariate effects and random effects can then be viewed as a smoothed version of the SIR. Model equation (2.5) can be employed to examine risk factors that increase (resp. decrease) the risk of disease in districts. The expected count E_i in district i is computed in our case under the constant risk hypothesis: this means that the risk is the same in each district:

$E_k = \theta N_k$, where $\theta = \frac{\sum_{k=1}^{K} Y_k}{\sum_{k=1}^{K} N_k}$ N_k, $k = 1 \ldots K$, N_k being the population size in district k (in the HNRS).

2.6.4 The spatial Binomial model

Since the counts are relatively small and the classification of each participant in the depressive class can be considered a Bernoulli trial, the Binomial model may be an alternative. Moreover, the maximum number of cases N in the HNRS is known and fixed (from the data). With previous notations the model can be described as

$$\begin{cases} Y_k & \sim Binomial(n_k, \theta_k) \\ ln(\frac{\theta_k}{1-\theta_k}) & = X_k^T \beta + \psi_k \\ \beta & \sim N(\mu_\beta, \Sigma_\beta), \end{cases} \tag{2.6}$$

where n_k denotes the number of trials and θ_k the probability of success in each trial in area k.

The Poisson model is more appropriate in our analysis because the effects of differential age demographics can be accounted for in the expected counts E_i. In the following we just present the results for the Poisson model.

2.6.5 Model building process

We use the data and prior information to critically evaluate our epidemiologic assumptions implied by the model and the statistical assumptions required by the model. Here we use stepwise regression with some forced-in covariates (including the exposure of interest), which are known to influence the outcome of interest (known confounding factors). All other covariates are subject to selection by a forward-selection algorithm. The results will be presented for sets of selected covariates: The first set (Set I) is the set of forced-in covariates. The second set of variables (Set II) includes Set I and all additional variables primarily measured at the individual level and aggregated to a district level. Set III corresponds to Set II and the variables primarily measured at the district level. These sets are depicted in Table 2.2.

We do not aim here to present the whole process of selecting variables. Each set of selected covariates will show the degree of association of the health outcome and the exposure of interest. Variables recognized to be influenced either by the exposure or the disease are firstly led aside. Controlling for such intermediates is added later. To conventionally include the additional covariates, we use a tradeoff between the charge-in-estimate method, statistical significance of the included covariates, and the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) of the model. In the charge-in-estimate method, covariates are selected based on relative or absolute change in the estimated exposure effect. The DIC is a hierarchical modeling generalization of the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

Set I	Set II	Set III
forced-in variables, including NDVI	Set$I$ + multi-morbidity + Body Mass Index + Traffic noise	Set II + Unemployment % in district

TABLE 2.2: Sets of covariates selected to illustrate the analysis of association between greenness and depression at the district level.

2.6.6 Implementation

For the implementation of the **Besag-York-Mollié** model (equation (2.5)), we employed the Integrated Nested Laplace Approximation (INLA) using the R-INLA package since it is a computationally efficient alternative to MCMC in many cases, particularly for CAR models. Rather than aiming at estimating the joint posterior distribution of the model parameters, the focus is on individual posterior marginals of the model parameters. Detailed information on fitting the model with R-INLA is given in (Blangiardo and Cameletti, 2015, Lindgren and Rue, 2015). The R-codes are given in Appendix A. The analysis with the **CARBayes** R-package (Lee, 2013) showed similar results (results not shown in this thesis).

2.7 Results

In this section, results are firstly provided for the identification of local clusters with elevated risk of depression. Secondly, the distribution of the raw SIRs are presented for each of the follow-up visits. Then, the results of the spatial model for disease mapping are presented, where the smoothed risk of depression are depicted for each of the follow-up visits in order to explore the spatial and spatio-temporal effects in the data.

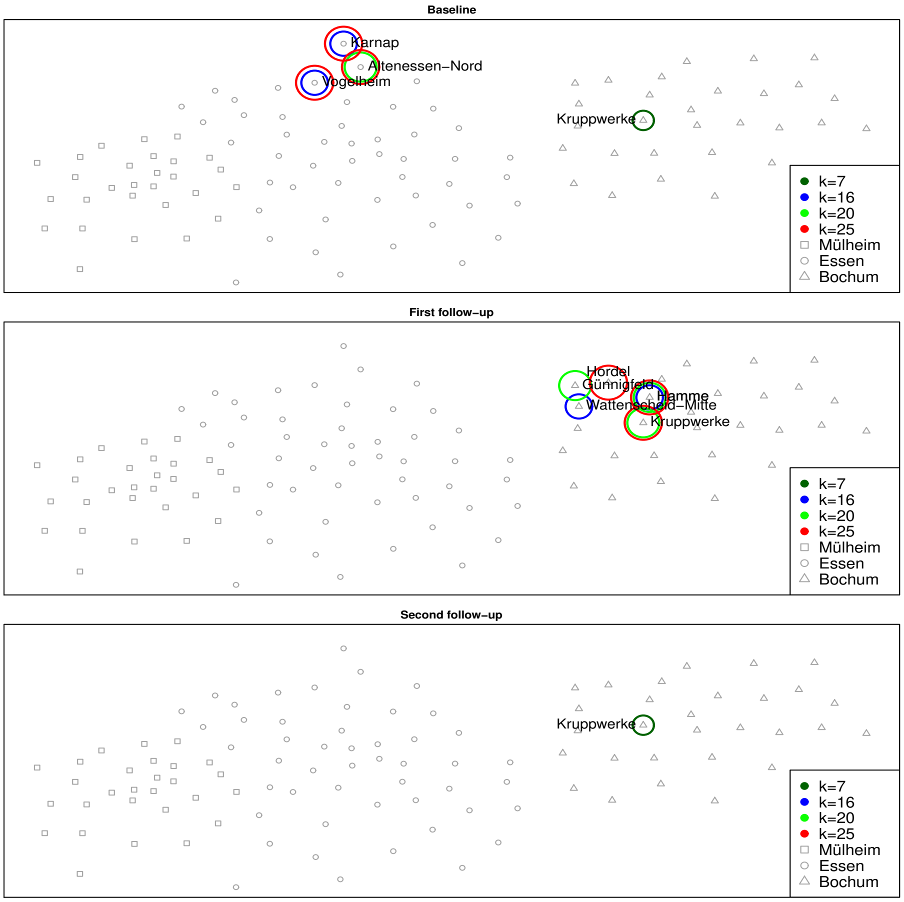

FIGURE 2.1: Districts around which significantly elevated clusters of cases of depression are observed in the study area, by the **Besag-Newell** method for selected cluster sizes k=7, 16, 20, 25, and for selected follow-up examinations.

We start with the **Local cluster detection**. Figure 2.1 displays the districts of the study area for each of the three examination time points at the study center (Years 0, 5, 10) with elevated risk of depression for the selected cluster sizes $k = 7, 16, 20, 25$. These clusters are mainly situated in the northern part of the study area, as indicated in Figure 2.1. Except for the district **Kruppwerke** in Bochum, we identified different sets of districts with clustered cases of depression for the three follow-up examinations considered.

For the **Global indicator of spatial autocorrelation**, we firstly consider $Y_k = SIR_k$ (compare with equation (2.2). $I = 0.09$ at baseline (Year 0). This value is compared with the value expected under the null hypothesis of spatial independence $E(I) = -\frac{1}{K-1} = -\frac{1}{107} = -0.0093$. The p-value is 0.0053 (From R-Funktion Moran.I), which at a selected significance level of 0.05 leads to the rejection of the null hypothesis. Therefore, the spatial distribution of the risks of depressive disorders does not reflect the result of random spatial processes. Even if the overall spatial clustering effect is small, it should not be neglected. Values for I in year 5 and year 10 are different but lead to the same conclusion about the spatial dependence.

Besides, we calculated the **Moran**'s I statistic for the residuals of the classical Poisson model with covariate effects (model equation (2.5) with ψ set to 0) for each of the time points considered to examine the appropriateness to use a spatial model. The estimated **Moran**'s I is 0.17 and the p-value is less than 0.05 for Year 0, suggesting evidence of unexplained spatial autocorrelation in the residuals from the first follow-up after accounting for the covariate effects. The Morans'I for the residuals of the classical Poisson model with covariate effects for the other two follow-up examinations at the study center are 0.06 and -0.08 respectively, with the null hypothesis of spatial independence not rejected. This shows a changing structure of the spatial effect over time.

The spatial distribution of the raw SIRs is shown in the first row of Figure 2.2. Neither clear patterns of elevated or low SIRs nor clear association with the distribution of greenness at the district level (second row of Figure 2.2) for the three follow-up periods is observable.

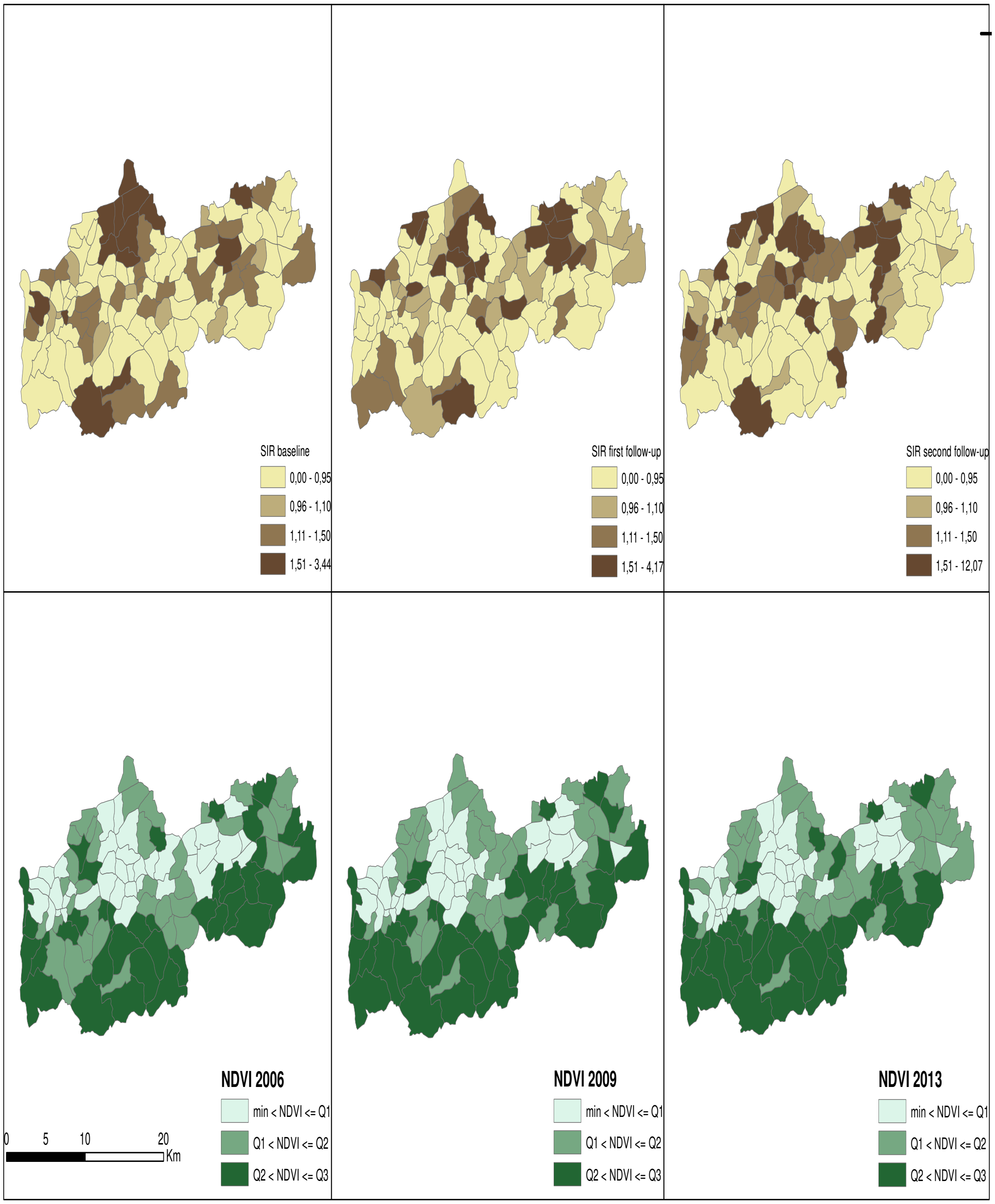

FIGURE 2.2: Standardized Incidence Ratios (SIRs) displayed in the first row of the figure for each of the 3 examination visits selected, compared to the spatial distribution of greenness given in the second row for the corresponding years.

The results of the smoothed risks are displayed in Figure 2.3. There is a pattern of elevated risk in the northern part of the study area for all 3 follow-up visits. This is in accordance with the identification of significant clusters of elevated risk in Figure 2.1 where clusters were identified in the northern part. Compared to the spatial distribution of greenness, elevated risks seem to occur in areas with a low level of greenness. It is then of great importance to examine the spatial pattern of the risk of depression as well as relationship between greenness and the risk of depression over time.

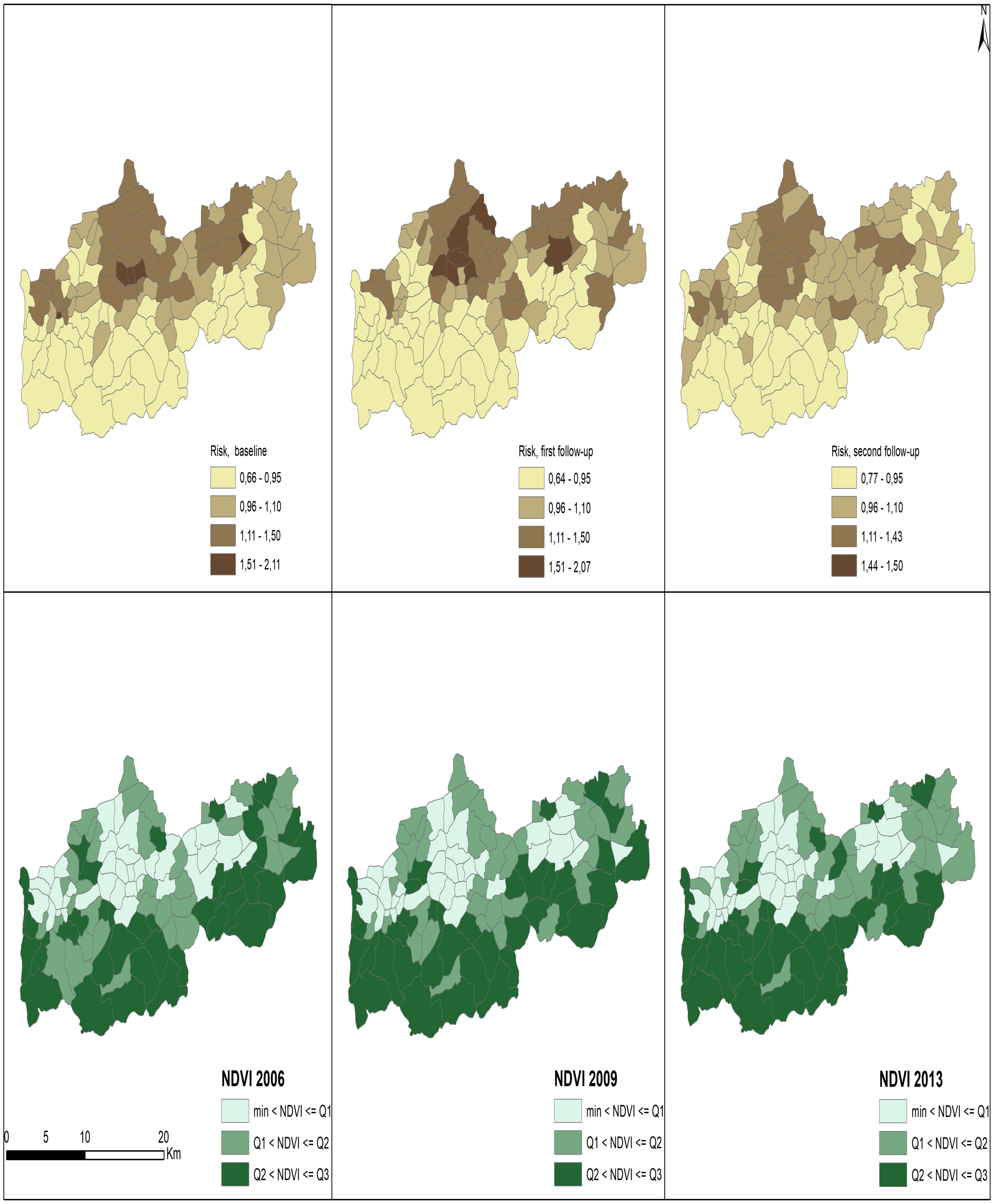

FIGURE 2.3: Risk estimate from the convolution model displayed in the first row of the figure for each of the 3 examination visits selected, compared to the spatial distribution of greenness given in the second row for the corresponding years.

2.8 Discussion

Our aim was to examine the risk of depression exploratory in the HNRS and analyse the effect of greenness on depression, prior to the spatio-temporal analysis. After applying the **Besag- Newell** method to identify local clusters of elevated risk of depression in the study area, a global indicator of spatial clustering was applied. A spatial Poisson model for disease mapping was then described within a Bayesian hierarchical model formulation (the **Besag-York-Mollié** model (Pfeiffer et al., 2008)) to estimate and smooth the risk of depression, accounting for covariate effects. The risk estimates from the spatial Poisson model were regarded as a smoothed version of the SIRs. The risk estimates from the spatial Poisson model were compared exploratory to that of SIRs and both were compared to the spatial distribution of greenness measured at the district level. The results inferred a spatial pattern of clusters of higher risk in the northern part of the study area. This pattern is explained by the unequal geographical distribution of some environmental risk factors in the study area, including greenness.

To achieve our goals and analyse data on the same spatial resolution, we have aggregated some covariates. Data aggregation has the consequence that some information is lost (Orcutt, Watts, and Edwards, 1968). With available covariates at the individual level as well as group level (district), there are existing methods to examine the effects of these covariates on individual-level outcomes as well as estimating risks at the district level, notably multilevel models (Pickett and Pearl, 2001). There is, for instance, a multilevel function in the CARBayes R-package (Lee, 2013). However, the multilevel model has another analysis goal. Our goal was to assess the risk at a group level and examine the associations at the group level. Using a spatial model for disease mapping that accounts for spatial effect is preferable to the mapping of standard estimates of disease risk like SIR. With the smoothed risk, a spatial pattern for the distribution of the risk of depression has been identified whereas the mapping of SIR showed no pattern.

Adding spatially structured extra-variability in the model fitted to the data when such extra-variability does not exist conditionally on the covariates included in the model (over-fitting) may bias the estimation of the ecological association between covariates and relative risks toward the null. However, in the case where no underlying extra-variability from the Poisson process exists, the simulation results show that models accounting for structured and unstructured residuals do not underestimate the ecological association, unless covariates have a very strong autocorrelation structure as shown in Latouche et al., 2007. Because spatial and non-spatial methods are equivalent in the absence of spatial autocorrelation in the errors (Beale et al., 2010), the precautionary principle suggests that models which incorporate spatial autocorrelation should be fitted by default as computational cost of unnecessary fitting a complex model that includes spatial effect is negligible compared with the danger of ignoring potentially important autocorrelation in the error.

In conclusion, we strongly recommend accounting for the spatial variation in the data when linking health outcomes and environmental exposures. Not only those environmental exposures are spatial in nature and can induce a spatial correlation in classical non-spatial models, but also accounting for spatial correlation when it is not present does not affect regression coefficients when there is no strong correlation between covariates. The methodology can now also be employed for other covariates and health outcomes of the HNRS. A spatio-temporal extension of the analysis is however needed to analyse longitudinal data as a whole and analyse the effects of covariates more precisely. Chapter 3 is dedicated to the spatio-temporal extension of the analyses performed in this chapter.

Chapter 3

Spatio-temporal analysis of the risk of depression at district-level and association with greenness

3.1 Introduction

When studying the effects of an urban exposition on a health outcome, the use of a longitudinal approach is most suitable. Even more important, data sets gathered at several time points over a longer time span can provide valuable information about the temporal dynamics of exposure and outcome. In this case, temporal autocorrelation comes into place as the same variables are measured several times. Spatial effect as discussed in Chapter 2 also arises simultaneously. There are several statistical models available in the literature for longitudinal studies linking health outcomes and urban expositions. For certain reasons as given in Chapter 2, data are aggregated to a common spatial resolution of interest like the districts of the HNRS. For such aggregated data, Poisson regression models are appropriate to estimate the disease risks, whereby accounting for covariate effects. Spatio-temporal variation are usually neglected in such urban health studies.

Ultimately, statistical methods for spatially and temporally referenced data should consider the spatial and temporal autocorrelation as well as the spatial and temporal heterogeneity to provide accurate, meaningful conclusions. Ignoring the spatial and temporal variation in the data runs the risk of violating the usual assumption of independent observations in ordinary regression analysis and may cause bias in the covariate effects (Cressie, 1993, Pfeiffer et al., 2008).

In this chapter, we would like to address this issue of spatio-temporal effect by using as an example the analysis of the effect of urban greenness on depressive symptoms from the data of the German HNRS. In Chapter 2, exploratory analyses showed some local clusters of elevated risks of depression as well as the presence of non-negligible spatial autocorrelation for particular time points. The spatial

Poisson model including the CAR model to account for the spatial effect was applied for particular time points (cross-sectional). Furthermore, previous analyses of the association between mental health outcomes (e.g. depression) and greenness were mainly cross-sectional, performed on individual-level variables, with spatial level covariates disaggregated to the individual level (Song et al., 2019, Beyer et al., 2014). The few ecological analysis on aggregated level did not use the spatial and spatio-temporal random effects to account for the unexplained spatial variation (Nutsford, Pearson, and Kingham, 2013). For the HNRS, linking health with greenness was analyzed in Orban et al., 2017. This analysis was done at the individual level and the spatio-temporal correlation was also not accounted for.

The main objective of this chapter is to apply a sophisticated spatio-temporal Poisson model on the analysis of the association between greenness and depression on the district level in the HNRS. This analysis with the spatio-temporal Poisson model extends the analyses performed in Chapter 2. Our approach will help to understand the distribution of the risk of depression over time and its influencing factors. The method takes simultaneously into account both spatial autocorrelation and spatial heterogeneity as well as change of spatial effect over time. Accordingly, spatio-temporal random effects are included while smoothing disease risks at a spatial level. This work is the first for the HNRS that accounts for the spatio-temporal effects (either spatial heterogeneity or spatial autocorrelation) at an areal level. The main part of this chapter is available in Djeudeu et al., 2020.

We organized our chapter as follows: In the method section (Section 3.2), after some formal notations, we firstly present the spatio-temporal model in Subsection 3.2.2 to describe the evolution of the estimated risk of depression. Then, the spatio-temporal CAR model for the spatio-temporal random effect is addressed in Subsection 3.2.3. A model building process similar to the one in Chapter 2 follows in Subsection 3.2.4. The method section ends with the summary of the implementation methods used to fit our model in Subsection 3.2.5. Section 3.3 is dedicated to the results from all the methods applied. In section 3.4, we discuss the results compared with other studies and future insights for analysis.

3.2 Methods

3.2.1 Formal notations

We keep the notations of Subsection 2.3 from Chapter 2 and consider some extensions for the spatio-temporal model. Data are recorded for each district for $t = 1, \ldots, N$ consecutive time periods, thus available for a $K \times N$ rectangular array with K rows(districts) and N columns (time periods). The response data are denoted by

$\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)_{K\times N}$, where $\mathbf{Y}_t = (Y_{1t}, \ldots, Y_{Kt})$ denotes the $K \times 1$ column vector of observations for all K spatial units for time period t. Next, $\mathbf{E} = (\mathbf{E}_1, \ldots \mathbf{E}_N)_{K\times N}$, where similarly $\mathbf{E}_t = (E_{1t}, \ldots, E_{Kt})$ denotes the $K \times 1$ column vector of expected counts. $X_{kt} = (x_{kt1}, \ldots, x_{ktp})$ is a vector of p known covariates for district k (and can include factors or continuous variables and a column of ones for the intercept term). N_{kt} is the population size in district k at time period t. To model the risk of depression at the district level, thereby accounting for covariate effects and unexplained spatio-temporal trends in the data, we use the hierarchical Poisson model.

3.2.2 The Poisson model, including spatio-temporal random effects

Here, we consider the spatio-temporal extension of the spatial Poisson model in Subsection 2.6.3 from Chapter 2) (see equation (2.5)).

The spatio-temporal model has the following form:

$$\begin{cases} Y_{kt} & \sim Poisson(\theta_{kt}E_{kt}) \\ ln(\theta_{kt}) & = X_{kt}^T\beta + \psi_{kt} \\ \beta & \sim N(\mu_\beta, \Sigma_\beta). \end{cases} \tag{3.1}$$

θ_{kt} denotes the risk (of depression) at time t in district k. $\beta = (\beta_1, \ldots \beta_p)$ is the covariate regression parameter vector, and a multivariate Gaussian prior is assumed with mean μ_β and diagonal variance matrix Σ_β. $\Psi = (\Psi_1, \ldots, \Psi_N)$ with $\Psi_t = (\psi_{1t}, \ldots, \psi_{Kt})$, comprises the random components ψ_{kt} for district k and time period t that is to be precisely described later.

Note that the expected count E_{kt} in district k for time period t is computed under the constant risk hypothesis for our application. This means: $E_{kt} = N_{kt}\xi_t$ (compare with the calculation of E_k in district k in Subsection 2.6.3 from Chapter 2). An age-adjusted expected count would have been possible (Waller and Gotway, 2004, page 203), but we use an overall mean risk estimate over the entire geography of the HNRS for each time period, $\xi_t = \frac{\sum_{k=1}^K Y_{kt}}{\sum_{k=1}^K N_{kt}}$.

3.2.3 CAR-prior for spatio-temporal random effect

With the random effect in equation (3.1), we wish to account for the evolution of the spatial structure of the data over time without forcing it to be the same for each time point. We use a single set of spatially and temporally autocorrelated random effects to allow the temporal and spatial correlation to be jointly modelled. $\Psi_t = (\psi_{1t}, \ldots, \psi_{Kt})$ is the vector of random effects for time period t, which evolves over time via a multivariate first-order autoregressive process with temporal autoregressive parameter ρ_T. The model was proposed by Rushworth, Lee, and Mitchell, 2014:

$$\begin{cases} \Psi_t|\Psi_{t-1} & \sim N(\rho_T \Psi_{t-1}, \tau^2 Q(W, \rho_S)^{-1}), t = 2, \ldots, N \\ \Psi_1 & \sim N(0, \tau^2 Q(W, \rho_S)^{-1}) \\ \tau^2 & \sim Inverse - Gamma(1, 0.01) \\ \rho_S, \rho_T & \sim Uniform(0, 1) \\ Q(W, \rho_S) & = \rho_S[diag(W.\mathbb{1}) - W] + (1 - \rho_S)I. \end{cases} \tag{3.2}$$

$\mathbb{1}$ is the $K \times 1$ vector of ones, while I is the $K \times K$ identity matrix. The random effects are zero-mean centered, while flat and conjugate priors are specified for (ρ_S, ρ_T) and τ^2, respectively. ρ_S and ρ_T are the spatial and temporal autoregressive parameters, respectively. $W = (w_{ij})_{i,j=1\ldots K}$ is the adjacency matrix (see equation (2.1)). The temporal autocorrelation is thus induced via the mean $\rho_T \Psi_{t-1}$. The spatial autocorrelation is induced by the variance $\tau^2 Q(W, \rho_S)^{-1}$. $Q(W, \rho_S)$ is the precision matrix, proposed by Leroux, Lei, and Breslow, 2000, which is a mixture between unstructured (uncorrelated) and structured (correlated) effects: The separation of spatially structured and unstructured variance is controlled by the parameter ρ_S, which defines the degree of the spatial dependency. $\rho_S = 1$ gives an Intrinsic CAR (ICAR), autocorrelation structure. $\rho_S = 0$ corresponds to unstructured (uncorrelated) heterogeneity. Hence, the model allows the degree of smoothing or clustering to be estimated. τ^2 is the location independent parameter controlling the overall magnitude of the prior variance. Further models for ψ are also available (Napier et al., 2019, Knorr-Held, 2000, Lee and Lawson, 2016, Rushworth, Lee, and Sarran, 2017, Bernardinelli et al., 2015, Napier et al., 2016). Note that the **Leroux** model is a direct spatio-temporal extension of the **Leroux** model mentioned in Subsection 2.6.2, not the extension of the **Besag-York-Mollié** model applied. The spatio-temporal extension of the **Leroux** model, although more complex than the spatio-temporal extension of the **Besag-York-Mollié** model, has been preferred because it explains the dynamics of the spatial effect better.

3.2.4 Model building process

All other covariates are subject to selection by a forward-selection algorithm. The results will be presented for sets of selected covariates: The first set (Set I) is the set of forced-in covariates. The second set of variables (Set II) includes Set I and all additional variables primarily measured at the individual level and aggregated to the district level. Set III corresponds to Set II and the variables primarily measured at the district level. These sets are depicted in Table 3.1. Compared to Table 2.2 from subsection 2.6.5, some variables like greenness measured at the district level are here time-varying and others are time-invariant like unemployment in districts for the study period considered. Recall that for the spatial Poisson model, we used the values of all variables at a single time point. Moreover, we do not have the same set of variables for the spatial and spatio-temporal model. In addition to the variables for the spatial model, the spatio-temporal model has the variable **Percentage relocations**. So, (Set II) here includes the variable **Percentage relocations** in addition, compared to Table 2.2.

To conventionally include the additional covariates, we also use a tradeoff between the charge-in-estimate method, statistical significance of the included covariates, and the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) of the model.

Set I	Set II	Set III
forced-in variables, including NDVI	Set$I$ + multi-morbidity + Body Mass Index + Traffic noise+ % of relocation	Set II + Unemployment % in district

TABLE 3.1: Sets of covariates selected to illustrate the analysis of association between greenness and depression at the district level.

3.2.5 Implementation

For the spatio-temporal model (equations (3.1) and (3.2)) we use the R-package CARBayesST, the Spatio-Temporal Areal Unit Modelling package described in Lee, Rushworth, and Napier, 2018. All models are fitted in a Bayesian setting using Markov chain Monte Carlo (MCMC) simulation methods. For all parameters, whose full conditional distributions have a closed-form distribution, the Gibbs sampler is used. This includes the regression parameters β and the random effects as well as the variance parameters. These full conditionals are available under request from the authors of the R-package CARBayesST (Lee, Rushworth, and Napier, 2018). We also used this R-package to fit our model. The remaining parameters are updated using Metropolis or Metropolis-Hastings steps and the random effects in equation (3.2) can be updated via the simple Gaussian random walk Metropolis algorithm or the Metropolis adjusted Langevin algorithm (Roberts and Rosenthal, 1998). The convergence of the samples is checked using the convergence diagnostic check proposed by Geweke, 1992.

3.3 Results

We present the results of the spatio-temporal model to describe the evolution of the estimated risk of depression. For illustration, the estimated risks are displayed for each time point, compared to the distribution of greenness.

The results of the fitted models from equation (3.1) are presented in Tables 3.2, 3.3, 3.4 and 3.5. To ease the interpretation, the displayed coefficient estimates for the Poisson model are transformed: The relative risk for an ϵ unit increase in a covariate with regression parameter β_k is given by the transformation $\exp(\epsilon\beta_k)$.

We observed a strong temporal autocorrelation in all models used ($\rho_T > 0.89$), meaning that the (unexplained) spatial variation is not changing much over time. The location-independent parameter τ^2 is always small, indicating a minor overall spatial effect in our data set. The small values of ρ_S indicate that the type of spatial effect present in the data is mainly the (uncorrelated) heterogeneity. Neglecting the spatial autocorrelation (forcing ρ_S to take value 0) does not change the coefficient estimates substantially, but leads to a slightly worse fit (Compare DICs in Tables 3.2 and 3.3). Thus, accounting for spatial autocorrelation, even with no strong effects as in our data, is better than neglecting it and just accounting for the spatial heterogeneity in the data.

The estimate of greenness in Tables 3.2, 3.3, 3.4, 3.5 shows a negative association between greenness and the risk of depression on district level for all sets of selected covariates. However, adding further potential confounders (from Set I to Set II and from Set II to Set III) decreases the greenness estimates towards null (towards 1 in the Tables with the transformation). The estimates for the other covariates show less changes. Overall, the model fits, measured here by the DIC improves (decreased DIC values) by adding further covariates. Regarding the spatial effect parameters, we neither notice a remarkable change of the spatial autocorrelation coefficient ρ_S (compare Tables 3.2, 3.4, 3.5) nor of the location-independent parameter τ^2 (see Tables 3.2, 3.3, 3.4, 3.5). We also notice no remarkable change in the time dependent autocorrelation coefficient ρ_T (compare Tables 3.2, 3.3, 3.4, 3.5).

We performed a sensitivity analysis on the prior for τ^2. Instead of the inverse-gamma prior as given in equation (3.2), we considered an improper uniform density for τ. The results did not change.

	Posterior quantiles		
	Median	2.5%	97.5%
Intercept	2.57	0.34	12.40
Greenness	0.90	0.85	0.96
Age	0.99	0.96	1.02
Income	0.99	0.99	1.00
τ^2	0.02	0.01	0.04
ρ_S	0.06	0.003	0.16
ρ_T	0.97	0.89	0.99
DIC	2846.26		

TABLE 3.2: Parameter estimates for model (3.1) for covariate Set I.

	Posterior quantiles		
	Median	2.5%	97.5%
Intercept	1.92	0.22	14.01
Greenness	0.91	0.85	0.98
Age	0.99	0.96	1.03
Income	1.00	0.99	1.00
τ^2	0.02	0.01	0.03
ρ_S	0.00	0.00	0.00
ρ_T	0.97	0.90	1.00
DIC	2850.18		

TABLE 3.3: Parameter estimates for model (3.1) for covariate Set I, with uncorrelated heterogeneity only.

	Posterior quantiles		
	Median	2.5%	97.5%
Intercept	0.11	0.00	1.22
Greenness	0.94	0.88	1.01
Age	0.99	0.96	1.02
Income	0.99	0.98	1.00
Percentage relocations	1.00	0.99	1.01
multi-morbidity	1.00	1.00	1.01
Body Mass Index	1.09	1.04	1.15
Traffic noise	1.01	1.00	1.02
τ^2	0.02	0.01	0.04
ρ_S	0.04	0.002	0.16
ρ_T	0.98	0.92	1.00
DIC	2837.45		

TABLE 3.4: Parameter estimates for model (3.1) for covariate Set II.

	Posterior quantiles		
	Median	2.5%	97.5%
Intercept	0.17	0.01	1.31
Greenness	0.96	0.91	1.02
Age	0.99	0.95	1.02
Income	0.99	0.99	1.00
Percentage relocations	1.00	1.00	1.01
multi-morbidity	1.00	1.00	1.01
Body Mass Index	1.06	1.02	1.11
Traffic noise	1.00	0.99	1.02
Unemployment status	1.03	1.02	1.05
τ^2	0.02	0.001	0.04
ρ_S	0.04	0.001	0.21
ρ_T	0.98	0.92	1.00
DIC	2823.95		

TABLE 3.5: Parameter estimates for model (3.1) for covariate Set III.

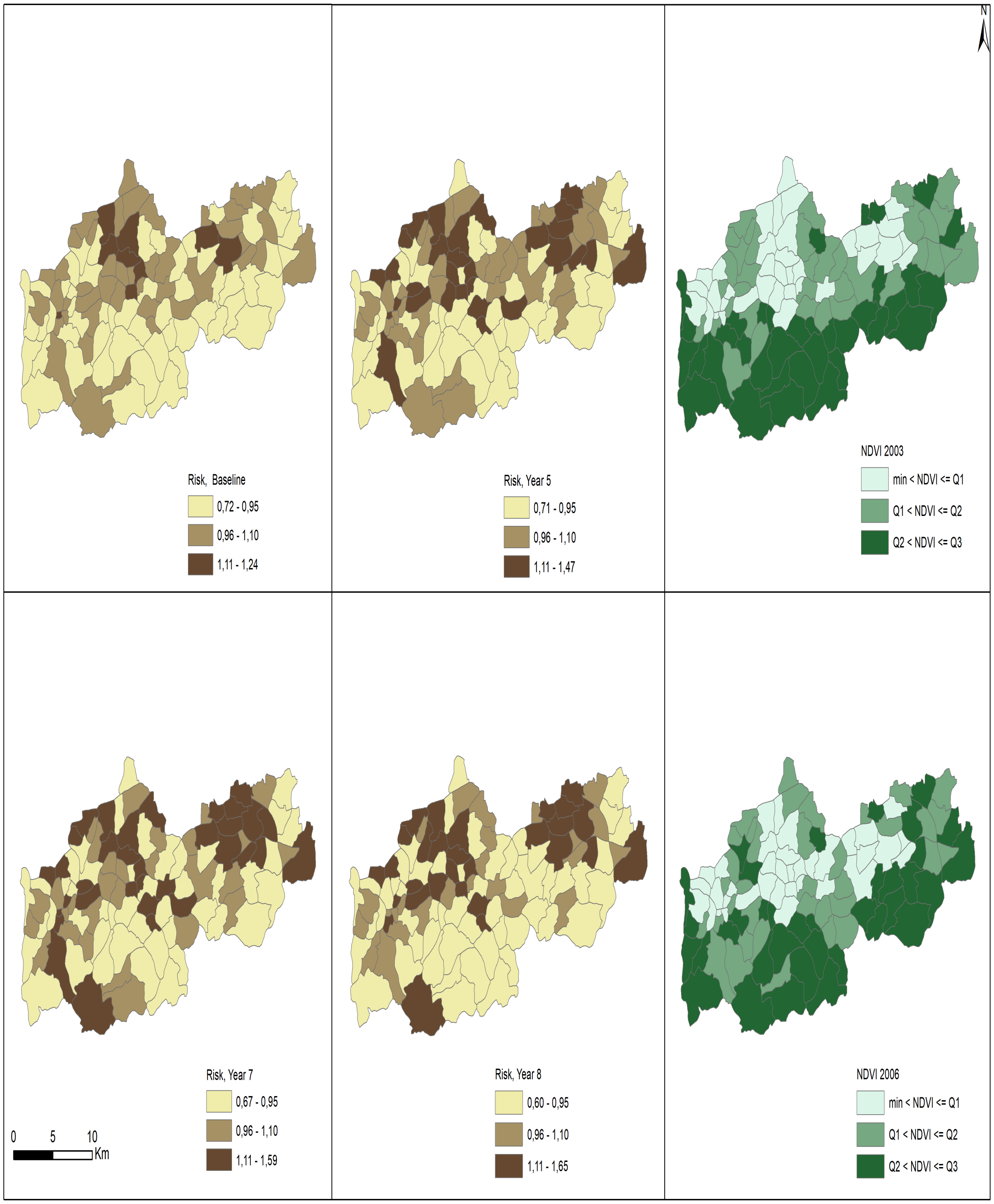

FIGURE 3.1: Risk estimate from the spatio-temporal model displayed in the first and second column of the figure for baseline, year 5, year 7 and year 8 compared to the spatial distribution of greenness given in the third column for the corresponding period of time.

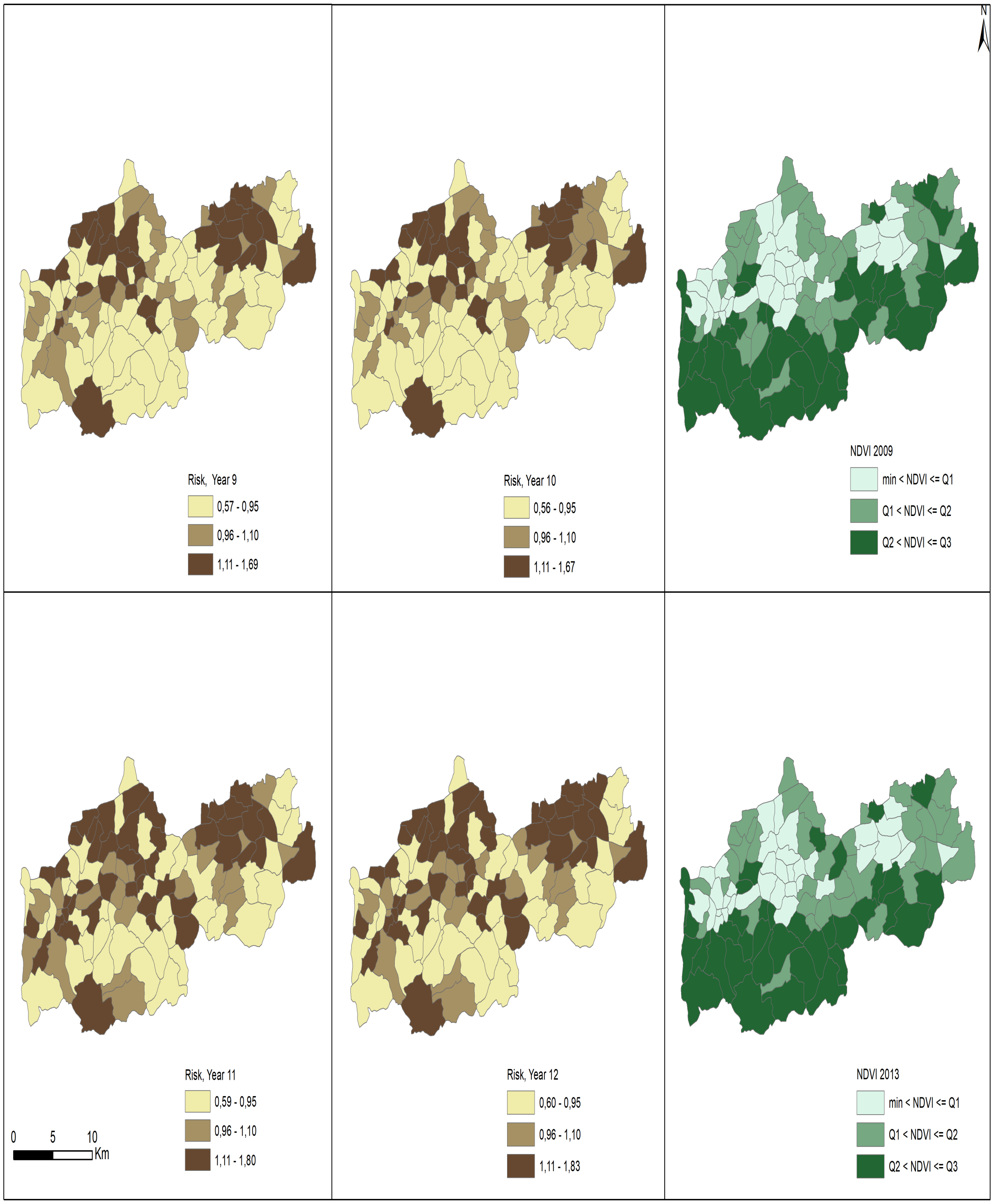

FIGURE 3.2: Risk estimate from the spatio-temporal model displayed in the first and second column of the figure for year 9, year 10, year 11 and year 12 compared to the spatial distribution of greenness given in the third column for the corresponding period of time.

Observing Figures 3.1 and 3.2, we notice that the spatial structure of the risk estimate is not changing very much over time. This is in accordance with the estimated autocorrelation parameter ρ_T in Tables 3.2, 3.3, 3.4 and 3.4.

In the displayed risk estimate over time, a pattern of an increased risk in the northern part of the study area for all time points is observable. Compared to the spatial distribution of greenness, increased risks for depression seem to be accompanied with low levels of greenness. Compare the risk estimates and the spatial distribution of greenness in Figures 3.2 and 3.1.

3.4 Discussion

We aimed to examine the effect of greenness on depression in a longitudinal study by specifically taking into account spatial and temporal (unexplained) variation through random effects in the ongoing longitudinal HNRS. In this way, we wished to investigate in more detail the strength of associations between greenness and depression, to better understand the spatial distribution of the risk of depression, and to provide more accurate estimates. Accordingly, we applied a sophisticated spatio-temporal model using aggregated individual and spatial data on the district level. In epidemiological studies, it is not common so far to analyse longitudinal and spatially referenced data sets while considering simultaneously spatial autocorrelation, heterogeneity as well as spatial variations over time. Nevertheless, spatio-temporal models for disease mapping generally account for the spatio-temporal effect (Lee and Lawson, 2016). In the following, we discuss our results shortly in the light of existing literature, applied statistical models as well as strengths and limitations. Overall, our results suggest that increased levels of greenness sustainably decrease the risk of depression at the district level. This observation is in line with previous studies, where greenness has shown negative associations with mental health both at individual (Song et al., 2019, Beyer et al., 2014) and spatial level (Nutsford, Pearson, and Kingham, 2013). However, unlike ours, these studies are based on cross-sectional data sets. We observed a strong temporal autocorrelation in the (unexplained) spatial variation. In addition to our exposure 'greenness', this points to environmental factors related to depression, that, if time-dependent, only slightly change over time. Apart from the strong temporal autocorrelation, our results indicate minor spatial effects in our HNRS which are mainly based on grouping effects, thus uncorrelated heterogeneity. These results are in agreement with the results from Chapter 2, where the global indicator of spatial autocorrelation (**Moran**'s I) was small for the selected time points. We ascertained however that the model fit worsened without consideration of spatial autocorrelation.

The strength of our space-time methodology is that we can understand the spatial effect and the dynamic thereof. Moreover, we are more confident in the estimation of the association between depression and greenness as the inclusion of the spatio-temporal random effects produces more reliable estimates. Therefore, this approach will have a better predictive ability.

Some limitations of our results have to be considered. (1) Our analyses are likely to be biased by dropouts during follow-up, unmeasured/unmeasurable covariates, and/or imputation of missing values. At least a full case analysis with participants who attended all three examinations at the study center revealed similar results such that bias by dropouts may not be expected to have a crucial role in our analyses. (2) As we aggregate individual data on the district level, we are well aware of the possibility of a loss of information (Orcutt, Watts, and Edwards, 1968). In addition, if the individual-level association is explored with aggregated data, this might lead to the "ecological fallacy". On the other hand, analyses with individual data also have some drawbacks when it comes to spatial issues. For instance, analyses based on individual data are often extrapolated to the spatial unit the individuals are assigned to. The inference of associations identified at the individual level by simply transferring them to an aggregated level can also lead to misinterpretations, which is called "atomic fallacy" (Wen et al., 2001). The focus of our study, however, was to consider longitudinal and spatially referenced data sets simultaneously while addressing spatial autocorrelation, spatial heterogeneity, and both spatial effects over time. Notably multilevel models are able to examine effects of covariates on individual as well as on district-level outcomes (Pickett and Pearl, 2001, Roux, 1998, Wakefield, 2009, Roux, 2000). There is, for instance, a multilevel function in the CARBayes R-package (Lee, 2013). However, this function so far is intended for spatial models and not for spatio-temporal random effect analyses. Our next research step, therefore, is to combine the advantages of multilevel models using data at their finest level given and the advantages of MRF models via their conditional specification to capture the spatial effects and the change of spatial effects over time and to extend the existing approaches accordingly. (3) Adding spatially structured extra-variability to the model when such extra-variability does not exist (over-fitting) may bias the estimation of the association between covariates and relative risks toward the null. However, in the case where no underlying extra-variability from the Poisson process exists, simulation results (Latouche et al., 2007) indicate that models accounting for structured and unstructured residuals do not underestimate associations unless covariates have a very strong autocorrelation structure.

In conclusion, neglecting or not taking into account the spatio-temporal autocorrelation in the analysis would lead to inaccurate estimates. As in Section 2.8 from Chapter 2, we strongly recommend accounting for the spatial variation in the data when linking health outcome and environmental exposures. Here, in addition, we suggest accounting for the change of spatial variation. Not only environmental exposures are spatial in nature and can induce a spatial correlation in regression models, but also accounting for spatial and spatio-temporal effects when it is not present does not affect regression coefficients when there is no strong correlation between covariates. The precautionary principle suggests that models, which incorporate spatial autocorrelation, should be fitted by default as the computational cost of unnecessarily fitting a complex model that includes spatial effect is negligible compared with the danger of ignoring potentially important autocorrelation in the error. The same argument holds for spatio-temporal model when the data shows a dynamic behavior. The methodology can now also be employed for other covariates and health outcomes of the HNRS. Further methodological development towards including individual and aggregated longitudinal data all at their given spatial resolutions in a combined multilevel MRF approach is necessary. Part II of this thesis will focus on the development of such methods.

Part II

Multilevel conditional autoregressive models

Chapter 4

Multilevel conditional autoregressive models for spatially referenced and cross-sectional epidemiological data

4.1 Introduction

Environmental exposure data are spatial in nature, partly determined by where people live and interact. Several issues arise when analysing the association between health outcomes and environmental exposures, with participants nested within spatial areas of interest. In this chapter, we discuss two of the issues.

The first is that risk factors and covariates related to individual level health outcome may be given on different spatial resolutions. It is better to employ data at their finest and initial spatial level, rather than aggregating or/and disaggregating them to a common spatial resolution, to avoid the ecological and atomic fallacy (see, e.g. Wakefield, 2009, Orcutt, Watts, and Edwards, 1968, Wen et al., 2001). The second issue concerns the spatial effect. Two main types of spatial effect should be considered: spatial autocorrelation and spatial heterogeneity. Spatial autocorrelation is closely related to the first law of geography (Tobler, 1979, Waller and Gotway, 2004) and can arise for several reasons, for instance, due to unmeasured/unavailable confounders. The second type is the grouping effect, also called spatial heterogeneity (Duncan, Jones, and Moon, 1998). Sometimes, the difference between the two types of spatial effects is not obvious.

Multilevel modeling (mostly 2-level) offers a framework to take advantage of the hierarchical structure of the data (Roux, 2000, Roux, 1998) and is widely used in many applications in the medical, educational and social science (Bryk and Raudenbush, 1989, Draper, 1995, Goldstein, Browne, and Rasbash, 2002, Nezlek, 2001).

The classical multilevel model (MLM CL2) has multiple advantages, but also some limitations. The inclusion of random effects at the highest hierarchical level helps to adjust for fixed effect estimates, for missing or unavailable covariates with spatial structure. A central assumption of interest in the MLM CL2 is that these random effects are mutually independent across spatial units, modelling spatial heterogeneity only. In the Bayesian framework, this consists of assuming exchangeable unstructured priors on area level random effects. This is equivalent to a global smoothing towards the mean effect. However, for positive covariation between adjacent units, positions of the spatial units are important; local smoothing using adjacency may be more appropriate. Figure A.1 in the supplementary material shows the importance of accounting for spatial autocorrelation instead of spatial heterogeneity only.

Several works in the literature of health geography and spatial epidemiology, mostly for cross-sectional analysis have recognised both the spatial heterogeneity and the spatial autocorrelation and modified the MLM CL2 accordingly: (Browne, Goldstein, and Rasbash, 2001, Hongwei, 2014, Dormann, 2007). However, most of these methods are neither useful when substantial autocorrelation is expected nor do they exploit spatial adjacency. Therefore, structured priors on the area level random effect recognizing adjacency are of interest. MRF models, particularly the CAR models (Besag, York, and Mollié, 1991, Hoef et al., 2018), are suitable for this task. CAR priors have the advantage of facilitating random effects analysis under a Markov Chain Monte Carlo (MCMC) sampling approach. Recent works that combined the advantages of multilevel models and MRF for cross-sectional data include Dong and Harris, 2015, Dong et al., 2015a, Dong et al., 2015b and the S.CARmultilevel() function of the CARBayes package (Lee, 2013). However, spatial confounding (Hodges and Reich, 2010, Reich, Hodges, and Zadnik, 2006) was not explicitly examined in the previous papers. We consider the multilevel model with restricted CAR model (MLM RCAR) to account for this spatial confounding.

The objective of this chapter is to compare the MLM CARs for cross-sectional data to the MLM CL2 in a simulation study. For cross-sectional analysis, these model classes are available in the literature although not common in simulation studies to analyse the consequences of additional random effect terms on the behavior of regression coefficients explicitly, which is the main goal of this chapter. In contrast to existing studies (Dong et al., 2015b), we additionally analyse whether adding the spatially correlated error term (CAR-prior term) to a linear model shrinks or enlarges/inflates the true regression coefficients. The Restricted Multilevel CAR Model (MLM RCAR) (Paddock, Leininger, and Hunter, 2016) accounts for spatial confounding. Thus, we compare the MLM RCAR, the MLM CAR and the MLM CL2 for cross-sectional data.

We organized our chapter as follows: we shortly present the MLM CL2 for cross-sectional data and some advantages and limitations. Then we introduce the MLM CARs for cross-sectional studies. Some selected comparison criteria for epidemiological and public health applications are carried out before defining the simulation strategy for comparing models. A computation section to explain the techniques used for simulation and model fitting ends the method part of the chapter. The results for simulation studies are then presented. For the application of the methods compared in the simulation studies, we consider the analysis of the association between depressive symptoms and greenness at the baseline in Subsection 4.4. The chapter ends with a discussion of the results in Subsection 4.5. we also consider perspectives for future insights.

4.2 The classical multilevel model: advantages and limitations

Multilevel models in their uniqueness help to overcome the problems of spatial heterogeneity and data given on different spatial resolutions. Moreover, multilevel models can help to separate the effects of neighborhood characteristics from the effects of individual-level attributes that persons living in a certain type of area may share. The central statistical model in the multilevel analysis is one of successive sampling from each level of a hierarchical population. However, there is some natural geographical hierarchy in environmental data and the sample data are viewed as a multistage sample from this hierarchical population. The hierarchical model we consider here is presented exemplarily for a better understanding.

Assume we have data from K districts, with a different number of participants n_j ($j = 1, \ldots, K$) in each district. On the individual (participant)-level, we have the outcome variable Y (square root of the depression scores). We only consider one explanatory variable on the individual level, say greenness (X) measured at the individual level, and one district-level variable, unemployment in district (Z).

We firstly set up separate regression equations in each district to predict the outcome variable Y as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}, i = 1, \ldots, n_j, j = 1, \ldots, K. \tag{4.1}$$

β_{0j} is the usual intercept, β_{1j} the usual regression slope for the explanatory variable greenness, and e_{ij} is the usual residual error term, with mean 0 and variance σ^2. The subscript j is for the districts ($j = 1, \ldots, K$) and the subscript i is for participants ($i = 1, \ldots, n_j$). The difference with the usual regression model is that we assume that each district has a different intercept coefficient β_{0j}, and a different slope coefficient β_{1j}. Across all districts, the regression coefficients β_{0j} and β_{1j}

have a distribution with some mean and variance. The next step in the hierarchical model is to explain the variation of the regression coefficients β_{0j} and β_{1j} by introducing explanatory variables at the district level, as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \tag{4.2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j} \tag{4.3}$$

Equation (4.2) predicts the average depression (square root of depression score) in a district (the intercept β_{0j}) by district-level variable Z. Thus, if γ_{01} is positive, the average depression is higher in districts with larger values of Z. Conversely, if γ_{01} is negative, the average depression is lower in districts with larger Z. Equation (4.3) states that the relationship, as expressed by the slope coefficient β_{1j}, between the depression Y and the greenness (X) of participants, depends on the value of unemployment in districts (Z). If γ_{11} is positive, the effect of greenness on depression is larger the larger the values of Z are. Conversely, if γ_{11} is negative, the effect of greenness is smaller when the values of Z are large. Thus, the unemployment in districts Z acts as a moderator variable for the relationship between depression and greenness. The term u_{0j} and u_{1j} in equations (4.2) and (4.3) are random residual error terms at the district level (macro errors), assumed to have a mean of zero, and to be independent from the residual error e_{ij} at the individual level. The variance of the residual errors u_{0j} is $\sigma^2_{u_0}$ and the variance of the residual errors u_{1j} is $\sigma^2_{u_1}$, and the covariance between u_{0j} and u_{1j} is $\sigma_{u_{01}}$. By including an error term in the district-level equations (equations (4.2) and (4.3)), these models allow for sampling variability in the district-specific coefficients (β_{0j} and β_{1j}) and also for the fact that the district-level equations are not deterministic (i.e. the possibility that all district-level variables have been included in the model). The model can be written as a single complex regression equation by substituting equations (4.2) and (4.3) into (4.1):

$$Y_{ij} = [\gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j] + [u_{1j} X_{ij} + u_{0j} + e_{ij}], \tag{4.4}$$

where $[\gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j]$ is the fixed part and $[u_{1j} X_{ij} + u_{0j} + e_{ij}]$ is the random part of the model. The term $X_{ij} Z_j$ is an interaction term that appears in the model as a consequence of modeling the varying regression slope β_{1j} of the individual level covariate X_{ij} with the class level variable Z_j. Thus, the moderator effect of Z on the relationship between the dependent variable Y and the predictor X, is expressed in the single equation version of the model as a cross-level interaction. An important issue is that the traditional multiple regression model assumes "homoscedasticity". In multilevel models, the random error term u_{1j} and the explanatory variable X_{ij} are multiplied, the resulting total error will be different for different values of X_{ij}. This is in contrast with homoscedasticity. We can incorporate additional individual-level and group-level covariates and have the same interpretations. One should keep in mind that group-level variables moderate the effect of individual-level variables.

There are some typical research questions in the analyses of the association between a health outcome and environmental exposures using multilevel models.

How much of the variance in the health outcome (Y) is attributable to individuals and how much to the areal unit (group)? Do we really need a multilevel model? How is the individual-level exposure (X) related to the health outcome (Y) in participants nested in groups, controlling for other confounders? In other words, what is the independent effect of X on Y? What is the independent effect of area-level covariate Z on health outcome Y? What is the cross-level interaction? In other words, how does the group-level variable (unemployment in districts for instance) moderate the effect of the individual-level variable (individual-level greenness for instance) on the outcome Y. Does the effect of individual-level exposure on health outcomes differ across groups? Do groups (districts in our case) differ in average health outcomes after controlling for the characteristics of individuals within the groups? To answer some of these questions, we consider in Chapter 6, Section 6.3.2, the advantages of the classical multilevel model via an application. The application is another example of the analysis of association between a health outcome and an environmental exposure of the HNRS. We show with the classical multilevel model that increased depression is negatively associated with increased neighborhood greenness, even when the hierarchical structure of the data is accounted for and socio-economic risk factors both at individual level and district level are controlled for. Just after controlling for district characteristics (unemployment in districts) the effect of individual-level exposure does not vary across districts: The variability in the effect of greenness across districts can be explained by some district-level covariates and neglecting these district-level covariates can be misleading.

Multilevel models help to address the problems listed above, but multilevel models do not account for spatial autocorrelation. MRF models can, in addition to the advantages of multilevel models, account for the spatial autocorrelation

4.3 Multilevel conditional autoregressive models for cross-sectional data

4.3.1 Model description

For multilevel models for cross-sectional data, the study region is partitioned into K non-overlapping areal units. Data are available on $n = \sum_{j=1}^{K} n_j$ individuals, with n_j individuals within area j, $j = 1, \ldots, K$. The following model equation accounts for both spatial heterogeneity and spatial autocorrelation:

$$\begin{cases} y_{ij} & = X_{ij}^T \beta + \psi_j + e_{ij}, \\ \psi_j | \psi_{-j}, W & \sim N\left(\dfrac{\rho \sum\limits_{k \neq j} w_{jk} \psi_k}{\rho \sum\limits_{k \neq j} w_{jk} + 1 - \rho} \dfrac{\tau^2}{\rho \sum\limits_{k \neq j} w_{jk} + 1 - \rho} \right), \\ e_{ij} & \sim N(0, \sigma_e^2) \ \forall \ i, j, \ \ i = 1, \ldots n, \ \ j = 1, \ldots, K, \end{cases} \tag{4.5}$$

where $\psi_{-j} = (\psi_1, \psi_2, \ldots, \psi_{j-1}, \psi_{j+1}, \ldots, \psi_K)$, σ_e^2, τ^2 follow an inverse Gamma ($IG(a, b)$) distribution, ρ follows a uniform distribution $R(0, 1)$. Here, X_{ij}^T is a $1 \times p$ vector of intercept and $p - 1$ covariates for individual i in area j. X_{ij}^T includes individual level as well as area-level covariates. $W = (w_{kj})_{k,j=1 \ldots K}$ is the binary adjacent matrix. The area-level random effect vector $(\psi_1, \ldots, \psi_K)^T$ has the **Leroux** structure given in equation (4.5) (Leroux, Lei, and Breslow, 2000, Congdon, 2010, p. $181 - 183$). This is the spatial version of the spatio-temporal **Leroux** model of equation (3.2) from Chapter 3. The **Besag-York-Mollié** model from Subsection 2.6.2 from Chapter 2 is also a good alternative for ψ. However, the **Leroux** model offers more flexibility. $\rho = 0$ corresponds to a lack of spatial interdependence, i. e. the classical multilevel model (2 levels, MLM CL2). By contrast, $\rho = 1$ leads to the intrinsic CAR (Congdon, 2010, p. $183 - 184$) model. In common with the sophisticated regression models applied in Chapter 2, the models also consist of adding existing area-level random effect terms with CAR prior specification, this time in the classical multilevel models for cross-sectional studies. But, in contrast to the models in Chapter 2, the random effect is given at an areal level while the outcome variable is at an individual level and covariates at different spatial levels.

Spatial confounding can be interpreted in linear-model terms as a collinearity problem. When spatial confounding is detected, spatial smoothing is restricted to the orthogonal complement of the fixed effect of area-level variables, called the Restricted CAR model (RCAR). This is recommended and described in Hodges and Reich, 2010, Reich, Hodges, and Zadnik, 2006. Choosing the restricted CAR models for the area-level random effect instead of the CAR models leads to the MLM RCAR.

4.3.2 Model comparison criterion for epidemiological studies

Model comparisons for regression analysis generally involve two related but different questions. The first is whether the increment in prediction or explained variability obtained by adding a random effect (structured and unstructured) is important. The second question, equally important, is whether the coefficients that describe the relationships (association) between the independent variable and the predictor variable in the model without random effect differs from the coefficient when the random effect is accounted for. This question cannot be answered in validly with conventional methods used for the assessment of incremental improvement in prediction. Methods suited for the former question need not be valid for considering the latter question (Clogg, Petkova, and Haritou, 1995).

In epidemiology and public health science, the goal is mainly inference rather than prediction. In other words, we are mostly interested in association-based (regression) models applied to observational data. This means that we use statistical regression models for testing causal explanations and this will lead to statistical conclusions in terms of effect sizes and statistical significance (Shmueli, 2010). Best practice in reporting regression analyses in Epidemiology and public health application include the report of the regression coefficients (beta weights) of each explanatory variable and the associated confidence intervals and P values, preferably in a table. The former captures the mean effect (strength) of the association (effect size) and the latter the uncertainty of the coefficient estimate (Litteratur). Due to the spatial arrangement of the data, there might be bias in reported coefficient estimates of model estimation if the appropriate method has not been used. This is why in our comparison, we prefer models for which uncertainty in coefficient estimates are reduced. Arguably, such models are not the best models to reduce model uncertainty (with better predictive ability). We are more interested in answering the following questions: Is the regression relationship between the response and a covariate of interest stable across two specifications? Or, does 'controlling' for additional random effect (random autocorrelation) suppress or enhance the relationship between the response and the covariate?

Since we are more interested in association-based (regression) models applied to observational data, we concentrate on comparison methods for which uncertainty in coefficient estimates are examined.

Let $\hat{\beta}_1^M, \dots, \hat{\beta}_p^M$ and $\hat{sd}_1^M, \dots, \hat{sd}_p^M$ be the estimated regression coefficients and standard deviations respectively, using regression model M to fit the generated data. The bias for the coefficient i, using model M is defined by $|\hat{\beta}_i^M - \beta_i|$, where $\beta_1, \ldots, \beta_p$ are the true regression coefficients. The Root Mean Square Error

$$RMSE(\hat{\beta}_i^M) = \sqrt{(Bias(\beta_i^M))^2 + Var(\hat{\beta}_i^M)}$$

assesses the quality of the estimators of the true regression coefficients using the underlying model M.

We use the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) as an in-sample prediction criterion, which is a reasonable choice for our nested models. When comparing candidate models, smaller values of DIC indicate better models. The DIC should be used together with the posterior log-likelihood before recommendation. The model maximizing the posterior log-likelihood is preferable.

4.3.3 Simulation strategy

We base our simulation on equation (4.5). We simulate two predictor variables: one variable defined at the individual level and the other defined at district level, both described by normal distributions. The individual-level error term is also simulated from a normal distribution. More details on the model equations used for simulating the spatial random effect are given in Appendix A, equation (A.2). ρ is an autocorrelation parameter and lies between 0 and 1. If ρ is closer to 0, then the simulated spatial effect is more similar to spatial heterogeneity. If ρ is closer to 1, then this is more similar to a CAR structure. A value of ρ between 0.4 and 0.6 corresponds to both medium spatial heterogeneity and autocorrelation. The different 9 scenarios are described in Table A.2 in Appendix A. Here we just include 3 scenarios for common data situations. For sensitivity analysis, for each scenario, different values of ρ and τ^2 are given. For each generated data, 3 candidate regression models are fitted and compared using the comparison methods listed in subsection 4.3.2: MLM CL2, MLM CAR and MLM RCAR.

4.3.4 Computation

All models are fitted in the Bayesian setting with Markov Chain Monte Carlo (MCMC) simulation methods. All parameters whose full conditional distributions have a closed-form distribution, i. e. the regression parameters and all area level, individual and observational level variance parameters, are updated using a Gibbs sampler (Gelfand, 2000). These full conditionals are available under request from the authors of the R-package CARBayesST (Lee, 2013).

For better comparability, we relied on the expert system of BUGS (Best et al., 2012), particularly run from within the R software using the R-package R2WinBUGS (Sturtz, Ligges, and Gelman, 2005). We used devices such as centering of the covariates as well as hierarchical centering of the random effects (Gelfand, Sahu, and Carlin, 1995) to reduce correlation in the joint posterior and increase Markov Chain Monte Carlo (MCMC) effective sample sizes. To access convergence and consistency of the chains, single as well as two parallel chains initialized at different points were used, and the Geweke diagnostic (Geweke, 1992), Brook & Gelman diagnostic (Brook and Gelman, 1998) and Heidelberger & Welch's diagnostic (Heidelberger and Welch, 2010) were applied. We ran the chains and chose the number of iteration until at least $\hat{R}$, the measure of mixing chains, is less than 1.02 for all parameters and quantities of interest. For the final estimation, we used a single long run after discarding a part of the sampled data: The length of the burn-in period was determined for each model separately. We also thinned the chains by storing only every 10th draw for the MLM CL2, MLM CAR, and MLM RCAR in order to decrease autocorrelation and speed up 'mixing'. For the Markov chains for the MLM CAR and the MLM RCAR, we could not detect departure from convergence after the 15000th iterations. The MLM CL2 stabilized earlier.

The models were run in parallel using the R-package 'batchtools' (Lang, Bischl, and Surmann, 2017), which provides a parallel implementation. The complete R-code for the simulation study is available. The important parts of the code are displayed in Appendix B of the additional materials.

4.3.5 Results for the simulation studies

The summary result is that the RMSE for the area level regression coefficients is larger for the MLM CL2 and the MLM CAR, compared to the MLM RCAR. This depends on the strength of the overall variance and spatial autocorrelation (CAR-structure) in the simulated data. The difference in the RMSE is mainly due to the standard error since there is little bias in the regression coefficients. The individual-level regression coefficients are not influenced very much. For a very small value of the overall spatial variance parameter τ^2, the bias is in general negligible. In more detail, the individual-level regression coefficients are very well retrieved by the three candidate models in any of the scenarios, though the classical model performs worse; compare the RMSE for the individual-level variable in Figure 4.1. The coverage of the 95% CI is 100% for all scenarios and for all candidate models (data not shown). For the area-level variable coefficient in Figure 4.1, there is negligible bias and RMSE for the three models in case of a very small value of τ^2. For moderate and higher values of τ^2, the RMSE is larger for the three models. Again, the RMSE is larger for the MLM CL2, smaller for the MLM RCAR, and larger the larger τ^2 is. The individual-level error variance is better retrieved by the MLM CARs (data not shown). The spatial autocorrelation parameter (data not shown) is also well retrieved by the MLM CAR and MLM RCAR; the MLM CL2 tends to overestimate the corresponding random effects variances.

Observing Figure A.5 (Appendix A), the DIC is smaller for the MLM RCAR compared to the MLM CAR and MLM CL2. The log-likelihoods are about the same for the three models.

With additional explanatory variables (data not shown), we obtain similar results. For a stronger correlation between the area-level variable and the CAR random effect term, the MLM RCAR clearly has better performances.

After centering of the random effects, the computation time for the MLM RCAR is slightly larger than that of the MLM CAR model. This, however, is not considerable compared to the risk of making weak inferences with biased regression coefficients.

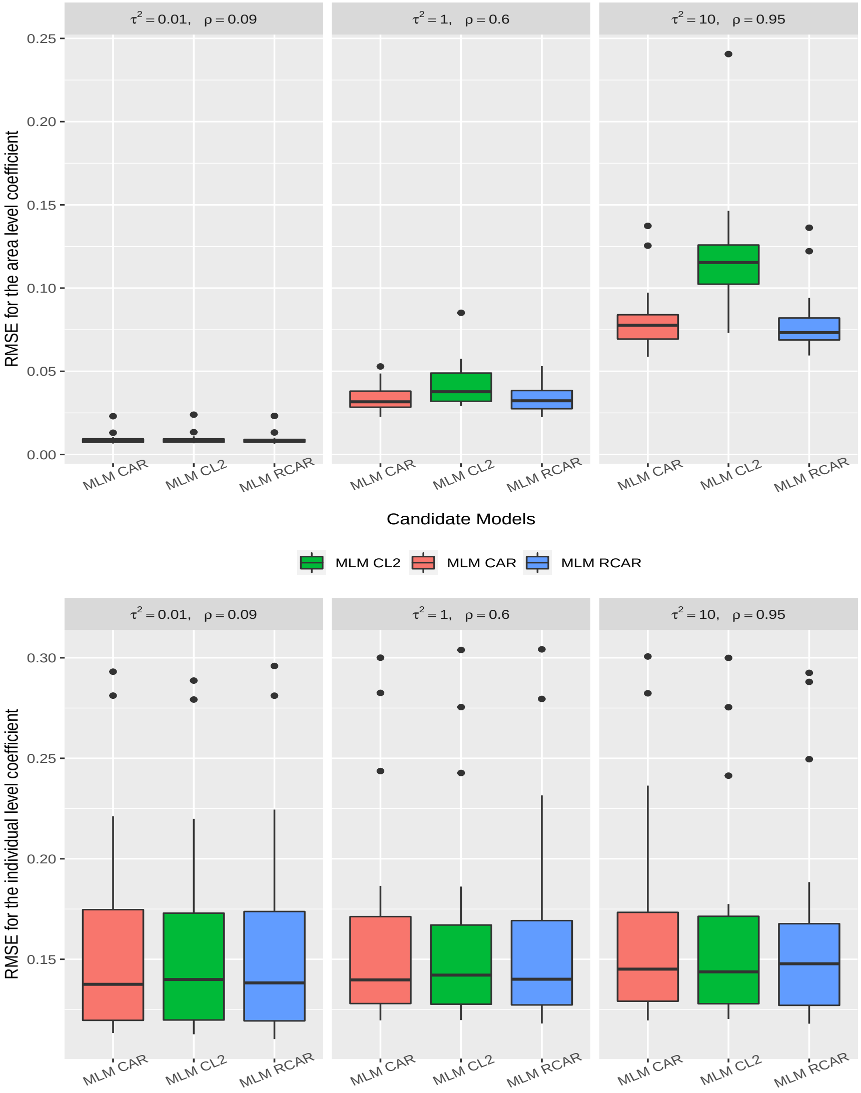

FIGURE 4.1: Comparison of the Root Mean Square Error (RMSE) for the area level and individual level variable coefficients, for a set of selected scenarios of the simulated spatial effect. The true value for the individual level coefficient is -1.50, while the true value for the area level coefficient is 0.14. τ^2 and ρ are the overall spatial variance and autocorrelation parameters from equation (A.2) in Appendix A, respectively.

4.4 Application to the analysis of the association between depression and greenness at baseline

4.4.1 Data

We use the data of the HNRS as described in Chapter 2. Here, all variables are considered at their initial spatial resolution, in contrast to the application in Chapter 2, where some variables are aggregated to the district level. For our analysis we used the CES-D data of the baseline.

Exposure to green space is commonly measured either as surrounding greenness or access to green space. Greenness is considered at the individual level. We base our analysis on baseline measurement of satellite imagery data.

Further covariates are included. Some are directly measured at the district level like the unemployment rate in districts, obtained from the local census authorities of the respective cities of Bochum, Essen, and Mülheim. Unemployment is a strong indicator for material deprivation in a neighborhood and was used as an indicator of neighborhood-level socio-economic status (SES). Other covariates such as socioeconomic (e.g., income), demographic (e.g., age), gender, medical history, and Body Mass Index (BMI) are measured at the individual level.

4.4.2 Analysis and results

We apply the MLM CL2, the MLM CAR, and the MLM RCAR comparatively. The square root of the depression scores leads to models that also better fit the assumptions and requirements of linear models. Square rooted depression is continuous with values between 0 and 6.70, has a mean 2.47, and a standard deviation of 1.21. In order to decrease autocorrelation and speed up the ’mixing’ of the MCMC, we thinned the chains by storing only every 10th draw. The length of the burn-in period was determined for each model separately. For the Markov chains, for the MLM CL2, we could not detect any departure from convergence after the 5000th iteration. For the MLM CAR, the chain stabilized also very quickly, at about 9.000 iterations. For the MLM RCAR, more iterations are needed, e.g., about 20000 to observe no departure from convergence for the spatial autocorrelation parameter.

The results suggest a negative association between greenness and depressive symptoms for all three methods (not significant because 0 is included in the 95% Credible Interval (CI)) as indicated in Table 4.1. For MLM CL2, MLM CAR and MLM RCAR respectively, a unit increase in the value of NDVI leads to a decrease

	MLM CL2			MLM CAR			MLM RCAR		
	Med.	2.5%	97.5%	*Med.*	2.5%	97.5%	*Med.*	2.5%	97.5%
intercept	2.45	2.40	2.50	2.45	2.39	2.52	2.45	2.40	2.51
greenness	-0.13	-0.51	0.23	-0.13	-0.47	0.22	-0.13	-0.47	0.21
female vs male	0.30	0.23	0.36	0.30	0.23	0.32	0.29	0.23	0.32
baseline Age	0.004	0.0003	0.008	0.004	0.0004	0.009	0.004	0.003	0.008
BMI	0.01	0.005	0.02	0.01	0.004	0.02	0.01	0.005	0.02
unempl. in district	0.02	0.01	0.03	0.02	0.01	0.03	0.02	0.01	0.03
τ	–	–	–	2.92	2.46	10.98	2.81	2.12	11.43
ρ_S	–	–	–	0.55	0.06	0.97	0.55	0.09	0.96
σ_{u_0}	0.08	0.05	0.11	–	–	–	–	–	–
σ_e	1.10	1.07	1.12	1.10	1.07	1.12	1.10	1.08	1.12
log-likelihood	-7017.66			-7020.42			-7018.29		
DIC	14060			14060			14060		

TABLE 4.1: Posterior quantiles from equation (4.5) for the MLM CL2, MLM CAR, and MLM RCAR for the analysis of the association between greenness and depressive symptoms of the Heinz Nixdorf Recall Study at baseline.

of the root of depression score, on average by -0.13 (with respective 95% CI given by $[-0.51; 0.23]$, $[-0.47; 0.22]$, $[-0.47; 0.21]$). We also notice that the coefficient estimates for the area level variables are the same for the three models. This is also in line with the results of the simulation study, where in the case of medium spatial autocorrelation, the three compared methods give approximately the same results for coefficient estimates and credible intervals.

A medium value of the spatial autocorrelation parameter $\rho_S = 0.55$ is indicative of a medium spatial dependence. Both the posterior log-likelihood and DIC are similar for the 3 models. This is expected because the spatial autocorrelation parameter ρ_S suggests medium spatial heterogeneity and medium spatial autocorrelation, see the results of the simulation studies displayed in Figure A.5 from Appendix A. The total 20000 iterations took 1.5 minutes for the MLM CL2, 254.75 minutes for the MLM CAR, and 3652.6 minutes for the MLM RCAR (just for comparison) on an ordinary personal computer.

4.5 Discussion

In the analysis of the association between health outcomes and environmental exposures, multilevel models should be widely used because of their simplicity and availability in several software packages. The ease of interpretation as well as the fact that the hierarchical structure of the data is used as an advantage make it very popular. Ignoring the spatial nested structure of the data can lead to biased results. In our application, increased greenness is associated with decreased depression scores in the HNRS, although it is not significant in the frequentist sense. Increased depression is negatively associated with increased neighborhood greenness, even when the hierarchical structure of the data is accounted for and socio-economic risk factors both at the individual level and the district level are controlled for. Just after controlling for district characteristics (unemployment in districts) the effect of individual-level exposure does not vary across districts: The variability in the effect of greenness across districts can be explained by some district-level covariates and neglecting these district-level covariates can be misleading.

Despite all the advantages of the MLM CL2, the method is not spatial in nature and is unable to capture spatial autocorrelation. MLM CARs models (MLM CAR and the MLM RCAR) were considered in a cross-sectional simulation study in comparison to the MLM CL2. The MLM CARs models are obtained by combining multilevel models and the properties of MRF in order to borrow strength in adjacent areas. The simulation studies were performed for several scenarios of the spatial effect. In summary, the results indicated that neglecting the spatial effect may lead to slightly larger RMSEs in coefficient estimates and slightly worse model fit in general. Multilevel CAR models for cross-sectional studies although available in some software packages are not spread and used in epidemiological studies routinely. This chapter may help consider the advantages of multilevel CAR models.

For the application, the three models showed the negative association between greenness and depressive symptoms, though not significant. This analysis of the association at the individual level is also perfectly in line with the analysis by Djeudeu et al., 2020, and the analyses in Chapter 2 and Chapter 3, which were performed on the HNRS at an aggregated level. Moreover, the current analysis has the added value that all spatial level variables are used at their finest level. This avoids the risk of the ecological and the atomic fallacy by aggregating or disaggregating them. The overall spatial effect was medium and the spatial autocorrelation was also medium. This explains why the models showed almost the same behaviours for coefficient estimates as expected from the scenarios of the simulation study.

As longitudinal analysis may add some values to the cross-sectional analysis, we are looking forward to extend the multilevel CAR models for longitudinal data.

Chapter 5

Multilevel conditional autoregressive model for spatially referenced and longitudinal epidemiological data

5.1 Introduction

Longitudinal studies are certainly expensive and time-consuming compared to cross-sectional studies but also offer additional strength. Several issues arise when analysing epidemiological data as introduced in Chapter 4. One issue that we could not discuss for cross-sectional studies concerns the dynamic effect for longitudinal data. Not only do individuals change over time, but (characteristics of) geographical units as well. These changes are the consequences of unavailable/unmeasured area-level covariates that may be changing over time and should, therefore, be considered (Steele, 2008, Bauer et al., 2013). Modelling and inference should exploit the dynamic and the spatial effect.

To examine the change over time of the health outcome in longitudinal studies, classical 3-level growth models (also MLM CL3) are widely used when participants are nested within geographical units. This consists of including area-level random effects that are independent across spatial units in the model equation, to account for the spatial effect. In addition to the aforementioned limitations of MLM CL2 for cross-sectional analyses, the geographical units are conceived here as entities that exert an effect that changes systematically with time. It is more realistic to assume that areas undergo structural and functional changes over time that are more stochastic in nature. These changes are not well captured by the classical model MLM CL3 (Bauer et al., 2013).

The main objective of this chapter is to further develop Multilevel CAR models for longitudinal data (MLM tCARs) by combining some already existing models:

The classical 3-level growth model and some CAR models to account for spatio-temporal random effects. We compare the developed MLM tCARs in the longitudinal setting to the classical (3-level) multilevel growth model (MLM CL3) in terms of accuracy and stability in the coefficient estimates under the presence and absence of spatial effects in data via a simulation study.

We organized our chapter as follows: In section 5.2 we introduce the developed MLM tCARs for longitudinal data. The simulation strategy follows. The method section ends with a computation subsection, which summarizes the techniques used for simulation and model fitting. In section 5.3 we outline the results of the simulation studies. In section 5.4 we apply the MLM CL3 and MLM tCARs for longitudinal studies comparatively, to the analysis of the association between depression and greenness in the longitudinal HNRS. Section 5.5 is dedicated to the discussion of the results of the simulation studies and the application.

5.2 Methods

5.2.1 Multilevel conditional autoregressive models for longitudinal data

Our model combines the advantages of multilevel models and the properties of the MRF. The goal is to accurately model the random effects in a multilevel growth model. Considered as a longitudinal version of model equation (4.5) from Chapter 4, it is defined as follows:

$$\begin{cases} y_{tij} & = X_{tij}^T\beta + \psi_{tj} + r_{0ij} + r_{1ij}g(t) + e_{tij}, \\ \psi_{tj} & \sim \text{(refer to equations (5.2) and (5.3) for candidate models)}, \\ (r_{0ij}, r_{1ij})^T & \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{r_0} & \sigma_{r_{01}} \\ \sigma_{r_{01}} & \sigma_{r_1} \end{pmatrix} \right) \\ e_{tij} & \sim N(0, \sigma_e^2) \;\; \forall \;\; t, i, j, \quad t = 1, \ldots, N, \;\; i = 1, \ldots n, \;\; j = 1, \ldots, K. \end{cases} \tag{5.1}$$

y_{tij} is a continuous outcome for individual i in the spatial unit j at the measurement occasion t. X_{tij}^T is a $1 \times p$ vector of intercept and $p-1$ covariates for individual i in area j at the measurement occasion t. X_{tij}^T includes individual-level and time-varying variables, individual-level and time-invariant variables, area-level and time-varying variables, and area-level and time-invariant variables. It also includes a deterministic function of time g, defining the individual growth. Note that g could be defined differently for each individual. $\beta = (\beta_0, \ldots, \beta_{p-1})^T$ with prior $\beta \sim N(0, \Sigma_\beta)$ is the vector of regression coefficients. There are three levels represented by sources of random variation at the area level, the individual level, and the observation level in the random effect part. We assume the outcome for individuals i in spatial areas j and time t are conditionally independent and normally distributed. r_{0ij} and r_{1ij}, independent of e_{tij}, are random effects for the individuals to account for person-to-person differences in the repeated measures and how they change over time. $(\psi_{t1}, \ldots, \psi_{tK})$ is the vector of random effects for time period t, which evolves over time and makes use of temporal and spatio-temporal dynamics. Independent of e_{tij}, r_{0ij} and r_{1ij}, it decomposes the spatial effects into spatial heterogeneity and spatial autocorrelation. The models for $(\psi_{t1}, \ldots, \psi_{tK})$ have the same structure like the spatio-temporal model applied in Chapter 3 (model equation (3.2)), where all data were applied at the district level. There are existing models for ψ_{tj} in the literature with this structure, mostly used in disease mapping models so far, including (Knorr-Held, 2000, Lee and Lawson, 2016, Bernardinelli et al., 2015). Such models also help to understand the dynamics of the spatial effect. We select three models for ψ_{tj} for our model in equation (5.1) and for our simulation study: the CAR ANOVA model, the convolution model, and the classical model. This lead to the MLM CAR ANOVA, MLM CONV und MLM CL3, respectively. Together, the MLM CAR ANOVA and MLM CONV are called MLM tCARs.

The CAR ANOVA model for ψ is given by

$$\begin{cases} \psi_{tj} & = \phi_j + \delta_t + \omega_{tj}, \\ \phi_j | \phi_{-j}, W & \sim N\left(\dfrac{\rho_S \sum\limits_{k \neq j} w_{jk}\phi_k}{\rho_S \sum\limits_{k \neq j} w_{jk} + 1 - \rho_S}, \dfrac{\tau_S^2}{\rho_S \sum\limits_{k \neq j} w_{jk} + 1 - \rho_S} \right), \\ \delta_t | \delta_{-t}, D & \sim N\left(\dfrac{\rho_T \sum\limits_{l \neq t} d_{tl}\delta_l}{\rho_T \sum\limits_{l \neq t} d_{tl} + 1 - \rho_T}, \dfrac{\tau_T^2}{\rho_T \sum\limits_{l \neq t} d_{tl} + 1 - \rho_T} \right), \\ \omega_{tj} & \sim N(0, \sigma_\omega^2), \quad t = 1, \ldots, N, \; j = 1, \ldots, K, \end{cases} \tag{5.2}$$

where $\phi_{-j} = (\phi_1, \phi_2, \ldots, \phi_{j-1}, \phi_{j+1}, \ldots, \phi_K)$, $\delta_{-t} = (\delta_1, \delta_2, \ldots, \delta_{t-1}, \delta_{t+1}, \ldots, \delta_T)$. $\rho_S, \rho_T \sim R(0,1)$, as well as $\tau_S^2, \tau_T^2, \sigma_\omega^2 \sim IG(a,b)$, see Knorr-Held, 2000. The model decomposes the spatio-temporal variation in the data into 3 components: an overall spatial effect common to all time points, an overall temporal trend common to all spatial units, and a set of independent space-time interactions. This is an ANOVA-type decomposition. This model is appropriate if the goal is to estimate overall time trends and spatial patterns. Here, the spatio-temporal autocorrelation is modelled by a common set of spatial random effects $\phi = (\phi_1, \ldots, \phi_K)$ and a common set of temporal random effects $\delta = (\delta_1, \ldots, \delta_T)$. Both are modelled by the CAR prior proposed by Leroux, Lei, and Breslow, 2000. W is the same as in subsection 4.3 while

$D = (d_{tj})$ is the binary $N \times N$ temporal adjacency matrix defined by $d_{tl} = 1$ if $|l - t| = 1$ and $d_{tl} = 0$ otherwise, $t, l = 1, \ldots, N$. Additionally, the model can incorporate an optional set of independent space-time interactions $\omega = (\omega_{11}, \ldots, \omega_{NK})$. For the inverse Gamma prior for the variance components $IG(a, b)$, values for a and b could be $a = 1, b = 0.01$.

The convolution model for ψ is defined by

$$\begin{cases} \psi_{tj} & = \phi_{tj} + \omega_{tj}, \\ \phi_{tj}|\phi_{-tj}, W & \sim N\left(\bar{\phi}_{tj}, \sigma^2_{\phi_{tj}}\right), \quad \bar{\phi}_{tj} = \dfrac{\sum\limits_{k \neq j} w_{jk}\phi_{tk}}{\sum\limits_{k \neq j} w_{jk}}, \quad \sigma^2_{\phi_{tj}} = \dfrac{\tau^2_t}{\sum\limits_{k \neq j} w_{jk}}, \\ \omega_{tj} & \sim N\left(0, \sigma^2_{\omega t}\right), \quad t = 1, \ldots, N, \;\; j = 1, \ldots, K, \end{cases} \tag{5.3}$$

where τ^2_t, $\sigma^2_{\omega t} \sim IG(a, b)$. $\phi_{-tj} = (\phi_{t1}, \ldots, \phi_{t(j-1)}, \phi_{t(j+1)}, \ldots, \phi_{tK})$. It could be considered as a direct spatio-temporal extension of the **Besag-York-Mollié** model in Subsection 2.6.2 from Chapter 2. The spatial autocorrelation parameter τ^2_t as well as the spatial heterogeneity parameter $\sigma^2_{\omega t}$ are allowed to vary over time. i.e., the model produces a separate effect for each area and each time point.

The classical linear growth model is the (3-level) model for which $\psi_{tj} = u_{0j} + g(t)u_{1j}$, reducing to $\psi_{tj} = u_{0j}$, depending on the goal of the analysis. Here, $u_{0j} \sim N(0, \sigma^2_{u_0})$ and $u_{1j} \sim N(0, \sigma^2_{u_1})$ capture the area level random effect or unexplained spatial variation for the intercept and slope respectively.

5.2.2 Simulation strategy

The simulation study is motivated by data situations typically observed in spatial epidemiology or health geography, where data are collected on different spatial levels. The main goal of the simulation study is to examine, how well the true regression coefficients for the candidate models are retrieved for simulated spatial effects. We use the geography of the HNRS.

We start with the simulation study for longitudinal data, based on equation (5.1). We simulate the spatio-temporal random effect ψ. Then, we simulate two covariates at the individual level, one of which is time-varying, from normal distributions. One time-varying variable is simulated at the area level, from a normal distribution. We consider a linear individual time trend $g(t) = t$. We hold predictor variables fixed as well as the individual level error term. After simulating the area level spatio-temporal random effect ψ_{tj} for each scenario, we generate the dependent variable. More details on the model equations used for simulating the spatio-temporal random effect are given in Appendix A, equation (A.1). ρ_S and ρ_T

are spatial and temporal autocorrelation parameters respectively, with values in the unit interval $[0,1]$. τ_S^2 and τ_T^2 are overall spatial and temporal variation parameters, respectively. The values for these parameters define the strength of the simulated spatio-temporal effect. For ρ_S and ρ_T we consider low, medium and high values, corresponding to 0.09, 0.5 and 0.9, respectively. For τ_S^2 and τ_T^2, we also consider low, medium and high values, corresponding to 0.009, 0.8 and 3 (large enough for this problem), respectively. Overall, there are $3^4 = 243$ possible scenarios. For presentation, we consider just a few selected ones for τ_S, τ_T and ρ_S, ρ_T as described in Table A.1 in Appendix A. For sensitivity analyses about the structure of the spatio-temporal random effect, we vary the values of τ_S, τ_T, ρ_S and ρ_T for each scenario. We also simulated different types of spatial models, all of which include spatial heterogeneity and spatial autocorrelation, whose strength changes over time. Additionally, we have also simulated several independent variables instead of just 3 for sensitivity analysis (data not shown). For each generated data, 3 candidate regression models are fitted and compared using the comparison methods listed in subsection 4.3.2: MLM CL3, MLM CAR ANOVA, and MLM CONV.

5.2.3 Computation

All models are fitted in the Bayesian setting with Markov Chain Monte Carlo (MCMC) simulation methods. All parameters whose full conditional distributions have a closed-form distribution, i. e. the regression parameters and all area level, individual and observational level variance parameters, are updated using a Gibbs sampler (Gelfand, 2000). The spatial and temporal parameter ρ_S and ρ_T for the MLM CAR ANOVA are updated using the slice sampler (Neal, 1997). Full conditional distributions for parameters of interest are calculated in Appendix A by applying Lindley and Smith, 1972. For better comparability, we relied on the expert system of BUGS (Best et al., 2012), particularly run from within the R software using the R-package R2WinBUGS (Sturtz, Ligges, and Gelman, 2005). We used devices such as centering of the covariates as well as hierarchical centering of the random effects (Gelfand, Sahu, and Carlin, 1995) to reduce correlation in the joint posterior and increase Markov Chain Monte Carlo (MCMC) effective sample sizes. To access convergence and consistency of the chains, single as well as two parallel chains initialized at different points were used, and the Geweke diagnostic (Geweke, 1992), Brook & Gelman diagnostic (Brook and Gelman, 1998) and Heidelberger & Welch's diagnostic (Heidelberger and Welch, 2010) were applied. We ran the chains and chose the number of iteration until at least $\hat{R}$, the measure of mixing chains, is less than 1.02 for all parameters and quantities of interest. For the final estimation, we used a single long run after discarding on of the samples: The length of the burn-in period was determined for each model separately.

For the longitudinal analyses, the MLM CL3 (3 levels) and the MLM CONV stabilized earlier at about 8000 iterations. For the MLM CAR ANOVA, we needed up to 250000 simulations for the chains for some parameters to stabilize. We thinned the chain by storing only every 10th draw.

The models were run in parallel using the R-package 'batchtools' (Lang, Bischl, and Surmann, 2017), which provides a parallel implementation. The complete R-code for the simulation study is available upon request.

5.3 Results for the simulation studies

In summary, the MLM CONV model and the MLM CAR ANOVA model perform much better than the MLM CL3 model. This is particularly important in the case of a strong spatial variation and a changing spatial structure over time. Otherwise, the MLM tCARs should be used cautiously to avoid overfitting, for instance.

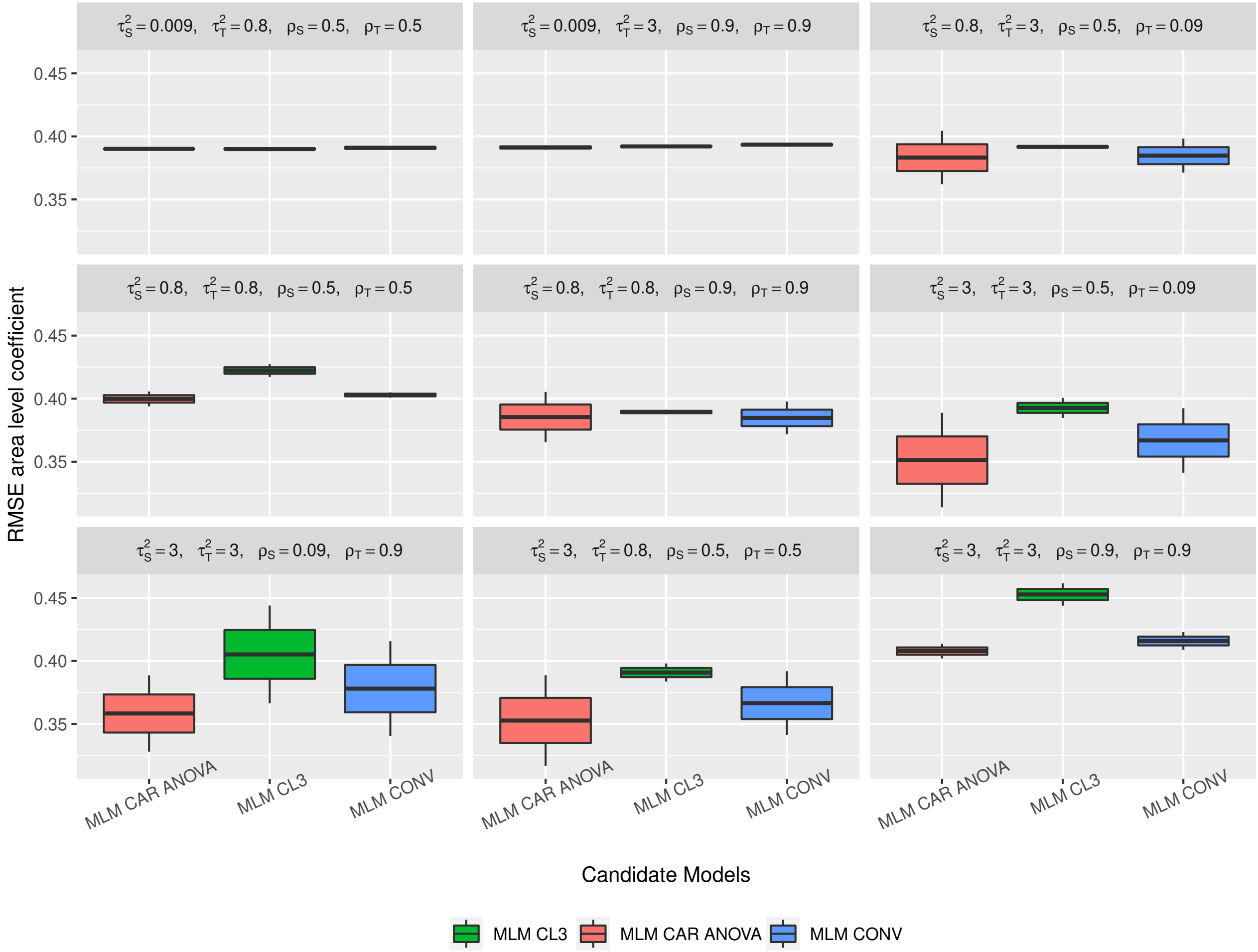

FIGURE 5.1: Comparison of the Root Mean Square Error (RMSE) for the area level variable coefficient, for the set of selected scenarios of the simulated spatio-temporal effect. The true value for the area level coefficient is 0.39. τ_S^2 and ρ_S, τ_T^2 and ρ_T are overall variance and auto-correlation parameters from equation (A.1) in Appendix A, for space and time respectively.

In more detail, the individual level regression coefficients are very well retrieved by the 3 candidate models in any of the scenarios as indicated in Figure A.3. However, the RMSE is more pronounced for the MLM CL3 compared to the MLM tCARs in general. The coverage of the 95% CI is 100% for all scenarios and for all candidate models (data not shown). For the area-level variable coefficient in Figure 5.1, the RMSE is not negligible for the three models. The bias and, therefore, the RMSE is larger in case of larger values of τ_S^2. The RMSE is still larger for the

MLM CL3. The time coefficient is almost equally retrieved for the three methods, and the RMSE is larger when the value of ρ_T larger is (see Figure A.2 in Appendix A).

Observing Figures 5.2 and A.4 (see Appendix A), the DIC and log-likelihood, respectively, for the MLM CAR ANOVA and MLM CONV are larger than that of the MLM CL3 in general. The DIC is smaller for the MLM CONV compared to the MLM CAR ANOVA, and smaller for the MLM CAR ANOVA compared to the MLM CL3. This indicates a better fit for the MLM tCARs in general. Note from the results of the sensitivity analysis for the model parameters and model goodness of fit that very small simulated observational level variances result in negative DIC values (results not shown) even though fitting a correct model.

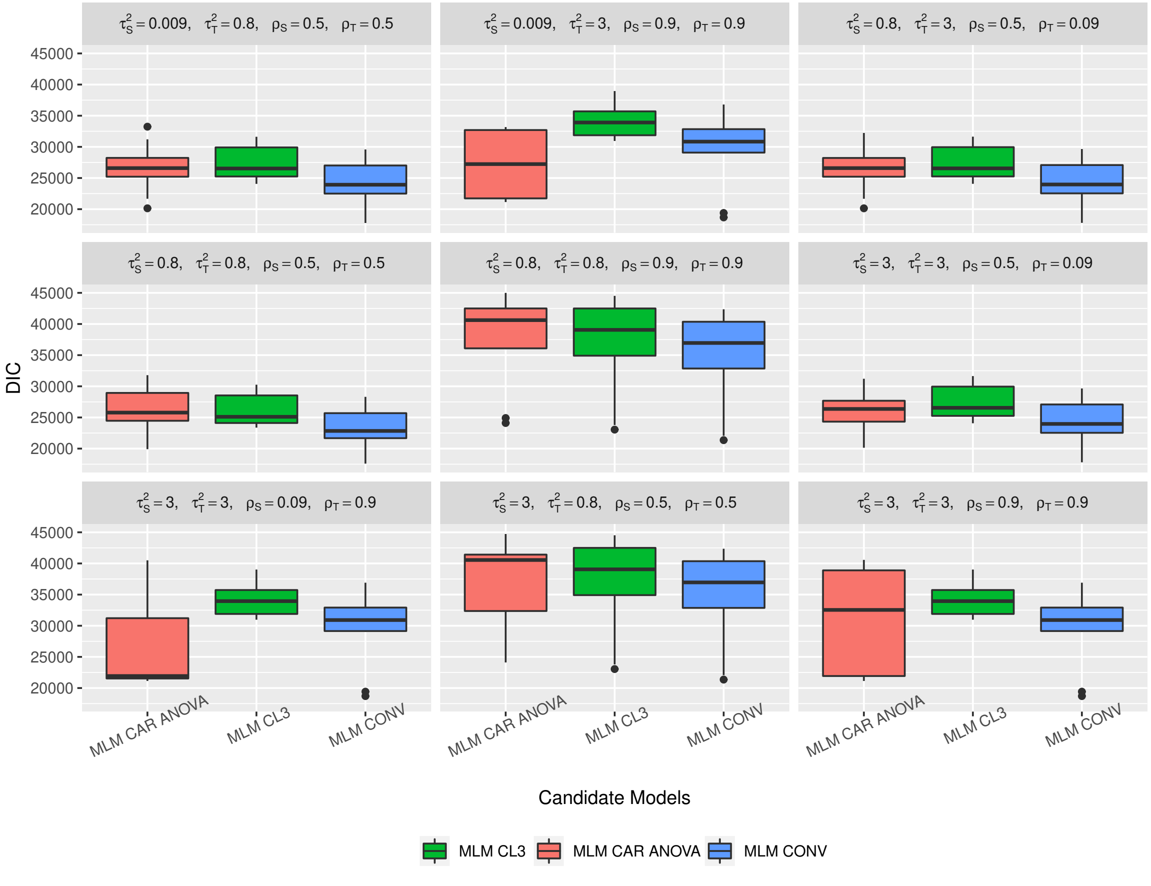

FIGURE 5.2: Comparison of the DIC, for a set of selected scenarios of the simulated spatio-temporal effect, longitudinal. τ_S^2 and ρ_S, τ_T^2 and ρ_T are overall variance and autocorrelation parameters from equation A.1, for space and time respectively.

In any case, the DIC values for the MLM tCARs are smaller in general. Moreover, small observational level (simulated) variances lead to very small bias in the individual level regression coefficients. The coverage of the 95% CI is very good for the three models in case of small values for τ_S^2 and τ_T^2.

For a very small value of spatial and temporal parameters, the MLM CL3 has a smaller RMSE values for the regression coefficients in general compared to the MLM CAR ANOVA. This is understandable as the MLM CAR ANOVA may be too complicated to fit such a simple structure. The MLM CAR ANOVA in comparison to the MLM CONV models seems to be more complex. The MLM CONV model shows better results for the RMSE for coefficient estimates and model fit in almost all scenarios.

For sensitivity analyses, different values for the fixed effect parameters for simulation were investigated. The results were similar regarding the behaviors of regression coefficients.

5.4 Application to the analysis of the association between depression and greenness

5.4.1 Data

We use the data of the HNRS as described in Chapter 2. But here, in contrast to the application in Chapter 2 and Chapter 3, data are used for eight measurement time points. In common with Chapter 4, all variables are considered at their initial spatial resolution. For our analysis, we used CES-D data of eight measurements assessed between 2000 and 2013.

Greenness is considered at the individual level. We base our analysis on eight time points of satellite imagery data.

Further covariates are included. Some are directly measured at the district level like the unemployment rate in districts, obtained from the local census authorities of the respective cities of Bochum, Essen, and Mülheim. Other covariates in the model such as socioeconomic (e.g., income), demographic (e.g., age), gender, medical history, and Body Mass Index (BMI) are measured at the individual level. In contrast to the NDVI, further covariates are time-invariant.

5.4.2 Analysis and results

We apply the MLM CL3, the MLM CAR ANOVA, and the MLM CONV comparatively. The square root of the depression scores leads to models that better fit the assumptions and requirements of linear models like in Section 4.4 from Chapter 4. Square rooted depression is continuous with values between 0 and 6.70, has a mean 2.47, and a standard deviation of 1.20.

Individual profiles of depression data with average trend line (Figure A.7, Appendix A) are used exploratorily to detect a slightly linear decreasing trend. We also analysed the spatial confounding at baseline. In order to decrease autocorrelation and speed up the 'mixing' of the MCMC, we thinned the chains by storing only every 10th draw. The length of the burn-in period was determined for each model separately. For the Markov chains, for the MLM CL3, we could not detect any departure from the convergence after the 8000th iteration. For the MLM CONV, the chain stabilized also very quickly, at about 9.000 iterations. For the MLM CAR ANOVA, more iterations are needed, e.g., about 1000000 to observe no departure from convergence for the spatial autocorrelation parameter.

Figure A.6 in the supporting materials shows the mixing of the chains for some fixed effect parameters of the MLM CONV. The trace plots look similar for other methods. Trace plots, density plots and further diagnostics for all parameter estimates as well as variance components are available upon request.

The results suggest a negative association between greenness and depressive symptoms for all three methods (0 is not included in the 95% CI) as indicated in Table 5.1. For MLM CL3, MLM CONV and MLM CAR ANOVA respectively, a unit increase in the value of NDVI leads to a decrease of the root of depression score, on average by -0.32 (with 95% CI $[-0.47; -0.16]$, $[-0.48; -0.16]$, $[-0.47; -0.15]$, respectively). The time slope is on average -0.02 ($[-0.03; -0.02]$) for all three methods (0 not included in the CI). This indicates a slightly linear decreasing trend, which is in perfect accordance with the exploratory individual profile plot of the data in Figure A.7 of Appendix A. We also notice that the coefficient estimates for the area-level variables are the same for the three models. The intercept is slightly different for the MLM CAR ANOVA. This is also in line with the results of the simulation study, where, in case of a very small overall spatial effect, not changing much over time, and a medium spatial autocorrelation, the three compared methods give approximately the same results for coefficient estimates and credible intervals. In fact, the spatial heterogeneity parameter for the MLM CONV σ_{ω_t} is very small (median value 0.09), and the spatial autocorrelation parameter for the MLM CONV σ_{φ_t} as well (median value 0.09). Both are not changing much over time. This is also in accordance with the MLM CAR ANOVA, where the overall spatial and temporal parameters τ_S and τ_T are 0.07 and 0.06. A medium value of the spatial autocorrelation parameter $\rho_S = 0.68$ is indicative of a medium spatial dependence. A large

	MLM CL3			MLM CAR ANOVA			MLM CONV		
	Med.	2.5%	97.5%	*Med.*	2.5%	97.5%	*Med.*	2.5%	97.5%
intercept	2.45	2.40	2.49	2.75	0.85	6.75	2.45	2.40	2.51
greenness	−0.32	−0.47	−0.16	−0.32	−0.48	−0.16	−0.32	−0.47	−0.15
female vs male	0.29	0.23	0.34	0.28	0.23	0.34	0.28	0.23	0.34
baseline Age	0.009	0.005	0.01	0.009	0.005	0.01	0.009	0.006	0.01
BMI	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.02
time	−0.02	−0.02	−0.01	−0.02	−0.02	−0.01	−0.02	−0.02	−0.01
unempl. in district	0.03	0.02	0.03	0.03	0.02	0.03	0.02	0.02	0.03
τ_S	—	—	—	0.07	0.04	0.13			
τ_T	—	—	-	0.07	0.04	0.15			
ρ_S	—	—	—	0.68	0.13	0.99			
ρ_T	—	—	—	0.82	0.03	0.99			
σ_{r_0}	0.85	0.83	0.88	0.85	0.83	0.88	0.85	0.83	0.87
σ_{r_1}	0.09	0.09	0.10	0.09	0.09	0.10	0.09	0.09	0.10
σ_{ω}	—	—	—	0.06	0.04	0.09	—	—	—
σ_{u_0}	0.06	0.04	0.09	—	—	—	—	—	—
σ_e	0.71	0.70	0.72	0.71	0.70	0.72	0.71	0.70	0.72
median $\sigma_{\omega t}$	—	—	—	—	—	—	0.09	0.05	0.16
median τ_t	—	—	—	—	—	—	0.09	0.05	0.22
log-likelihood		−27194.85			−27190.52			−27186.08	
DIC		59570			59570			59560	

TABLE 5.1: Posterior quantiles from equation (5.1) for the MLM CL3, MLM CAR ANOVA and MLM CONV for the analysis of the association between greenness and depressive symptoms of the Heinz Nixdorf Recall Study.

value of the temporal autocorrelation parameter $\rho_T = 0.82$ is indicative of a spatial effect that is not changing much over time. The spatio-temporal interaction parameter $\tau_\gamma = 0.06$ is also very weak. Both the posterior log-likelihood and DIC are better for the MLM CAR ANOVA and MLM CONV, particularly the MLM CONV, compared to the MLM CL3. The total 30000 iterations took 57.26 minutes for the MLM CL3, 57.22 minutes for the MLM CONV, and 57.98 minutes for the MLM CAR ANOVA (just for comparison) on an ordinary personal computer.

Similar results for fixed effect estimates are also expected from the decision tree in Chapter 6 after having a look at the spatial structure of the data. For a weak spatial effect, not changing much over time, the MLM tCARs were suggested, with the MLM CL3 as an alternative.

5.5 Discussion

We aimed to explore the effect of considering, or neglecting spatio-temporal random effects, on regression coefficients in epidemiological data, with participant nested within geographical units. We noticed from Chapter 4 that MLM CL2 fail to capture the proximity effect in epidemiological studies, where subjects are nested within geographical units. MLM CARs are alternatives to help explain the spatial effect better. They have been developed for cross-sectional studies but not for longitudinal studies so far. This chapter has further developed the multilevel (growth) models for longitudinal data by adding existing area-level random effect terms with CAR prior specification, whose structure is changing over time. We named these models MLM tCARs for longitudinal data.

The MLM tCARs models are obtained by combining multilevel models and the properties of MRF to borrow strength in adjacent areas. We considered regression models for longitudinal studies in simulation studies in several scenarios of the spatio-temporal effects. We introduced new models named MLM tCARs (MLM CAR ANOVA and MLM CONV). Such models are not common in the literature, at least not for epidemiological analyses. These models help to explicitly borrow strength and simultaneously account for individual dynamics as well as area dynamics. After introducing the MLM tCARs, we compared them to the MLM CL3. In summary, the results indicated that neglecting either the spatial or spatio-temporal effects leads to larger RMSEs in coefficient estimates and worse model fit in general. In the end, we applied the methods to the analysis of the association between depression and greenness in the HNRS. In the following, we discuss our results, pointing out the strengths and limitations.

To achieve our goals, we had to simulate the spatial and spatio-temporal random effects. The simulation of random effects from CAR models is not common and should be taken cautiously (Hodges, 2014). However, CAR-type random effects are successfully used in many works and applications to simulate the spatial structure in order to show the correctness and advantages of some CAR-like spatial models (Lee, Rushworth, and Napier, 2018). We used the geography of the HNRS, but the results can be generalized to other geographical structures.

For the application, the three models showed the negative association between greenness and depressive symptoms. In contrast to the analysis of the association at baseline in Section 4.4 from Chapter 4, the association is significant. This is understandable since cross-sectional studies offer just a snapshot of a single moment in time, whereas longitudinal studies allow the direct assessment of changes in the response variable over time. We also noticed a linear decreasing individual trend. This analysis of the association at the individual level is also perfectly in line with the analysis by Djeudeu et al., 2020, which was performed on the HNRS at an aggregated level. Moreover, the current analysis has the added value that all spatial level variables are used at their finest level. This avoids the risk of the ecological and the atomic fallacy by aggregating or disaggregating them. Though the analysis of the association and growth showed the same result for all three models for the coefficient estimates of interest, the MLM CAR ANOVA, as well as the MLM CONV, showed an overall better fit compared to the MLM CL3. Moreover, using the MLM CAR ANOVA or the MLM CONV, we were able to separate the individual time trend and the spatial time trend. The spatial effect was not changing very much over time. The overall spatial effect was medium and the temporal autocorrelation was rather strong. This explains why the models showed almost the same behaviours for coefficient estimates as expected from the scenarios of the simulation study and as indicated in the decision tree in Figure 6.2.

The MLM tCARs have the limitation that using them unnecessarily for non-complex data structures may lead to overfitting and slightly time-consuming fits. However, the damage of neglecting the spatial structure is more important than the unnecessary computational burden to fit the model. The bias for individual-level coefficients is generally negligible for all methods. However, not accounting for the spatio-temporal dependence and dynamic could lead to large standard deviations for regression coefficients and therefore larger RMSEs for classical models. The estimates of the standard errors determine the 'significance' of the fixed effect parameters (frequentist). Not choosing the right model to account for the spatio-temporal effect will lead to (falsely) small p-values (frequentist) and, therefore, false 'significant' associations between health outcome and exposure in some epidemiological studies. Therefore, it is recommended to consider a model that accounts for the spatial effect as well as the dynamic of the residual variation rather than using the MLM CL3.

The spatio-temporal CAR-prior should, however, be used cautiously. In the situation of spatial confounding, we recommend to use the restricted CAR models. This will avoid the fixed and random area-level effects to compete to explain common variation, which could distort estimates of the fixed effects area level regression coefficients and unduly inflate their posterior variances. This was considered for our cross-sectional analyses. The MLM CAR and MLM RCAR were just a bit more time-consuming compared to the MLM CL2. Although this seems to be a huge disadvantage, in absolute practical terms in a multitasking computing environment, it makes little impact. Relative to the time taken to collect the data (sometimes more than 10 years) it is irrelevant. Restricted MLM tCARs models still need to be fully developed and were not part of our simulation study.

Using MLM tCARs and MLM CARs from Chapter 4, spurious inferences regarding fixed effects parameters can be avoided. This is particularly important when the primary inferential focus is on fixed effect estimates as in several epidemiological analyses. If the goal of the analysis was a model comparison in terms of prediction, we would have to use some sort of hold-out of data. See White, Gelfand, and Utlaut, 2017 for analysing spatial data with the goal of spatial prediction. Methods suited for inference may not always be appropriate for comparing models in terms of prediction Clogg, Petkova, and Haritou, 1995.

We noticed that the MLM CONV has a less complex spatio-temporal structure and performs a bit better than the MLM CAR ANOVA in most of the scenarios, also in the application. MLM tCARs with different spatio-temporal CAR prior structures for the area level random effects can also be compared in future works to further improve the decision tree. The MLM CAR ANOVA did not retrieve the intercept properly for certain scenarios. This could be a result from spatial and temporal confounding that should jointly be addressed in future studies. The overall results in this chapter are developed for Gaussian likelihood models and could easily be extended to other likelihood models (generalized linear models). A preliminary simulation study with the logit-link function for cross-sectional data showed similar results for coefficient estimates that we described here.

For the growth modelling, we considered a linear time trend in this thesis. A model with nonlinearity in time could have been considered Grimm, Ram, and Hamagami, 2011.

Part III

Decision trees

Chapter 6

Decision trees

6.1 Introduction

Regression methods are needed in several fields of studies involving health for different goals depending on the fields of research or application. In social science, for instance, it is important to identify social determinants of health and promote interventions on these factors to improve population health. In spatial epidemiology, there is a need to examine and describe the spatial distribution of disease, risk factors for disease, and the intersection of the two both visually and statistically using geographically-referenced data. A wide range of risk factors including demographic, environmental, behavioral, socioeconomic, genetic, and infectious risk factors are generally available. Most of these variables have a spatial background and sophisticated regression models are needed. Most of the time, non-statisticians or non-experienced users of spatial data analyse data in order to reach the goals listed above. Although there exist many advanced methods available for analysing health data within a geographical context, simpler methods are often called upon to obtain a ′quick′ solution. Thus the choice of simplicity may lead sometimes to erroneous conclusions.

We aim in this chapter to provide decision trees to help experienced as well as non-experienced users of spatial data decide on the appropriate methods when spatial and spatio-temporal effects are suspected in linking health outcomes and exposures with a spatial background.

We primarily display the first decision tree in section 6.2.1 in Figure 6.2, after a short description of the methods used to construct the decision tree in Section 6.2. This first decision tree is restricted to the case where a natural hierarchy arises from the data like participants nested within geographical units. Then, we consider the case where all variables in the data are measured at an areal level or some variables aggregated in order to analyse data on a unique spatial resolution in Section 6.2.2. For the sake of completeness, we consider giving general tips for the implementation of the methods described in the decision tree in software packages in section

6.3. A more detailed example is provided to fit the MLM CL2 in Section 6.3.2. The main Codes to fit multilevel CAR models are provided in the Additional materials in Appendix B of the additional materials.

6.2 Methods

For data given on disparate spatial resolutions, the suggestion of the methods is based on the results of our simulation studies. We use a tradeoff between accuracy to retrieve regression coefficients, model's goodness of fit, and time needed to complete the fit, in order to suggest models. We consider methods for cross-sectional data as well as methods for longitudinal data. spatial methods, as well as non-spatial methods, are also of concern. The methods are summarized in Figure 6.2. We recommend the MLM tCARs for longitudinal data instead of the MLM CL3 when a strong spatial effect is expected in the data. Some of the MLM tCARs, like The MLM CONV, can be used routinely, no matter how strong the spatial effect is. For cross-sectional data, we recommend using the MLM RCAR and MLM CAR instead of the MLM CL2 except when a very weak spatial effect is expected in the data.

For data given on the same spatial level, the applications in Chapter 2 and Chapter 3, as well as the literature therein, helped to propose the decision tree. It is recommended to use generalized linear models that simultaneously account for spatial heterogeneity and spatial autocorrelation in order to produce more accurate results. CAR models are appropriate to account for such spatial effects. For longitudinal data, it is better to use models that account for the dynamics of the spatial effect over time.

6.2.1 Decision tree for data given on disparate spatial level

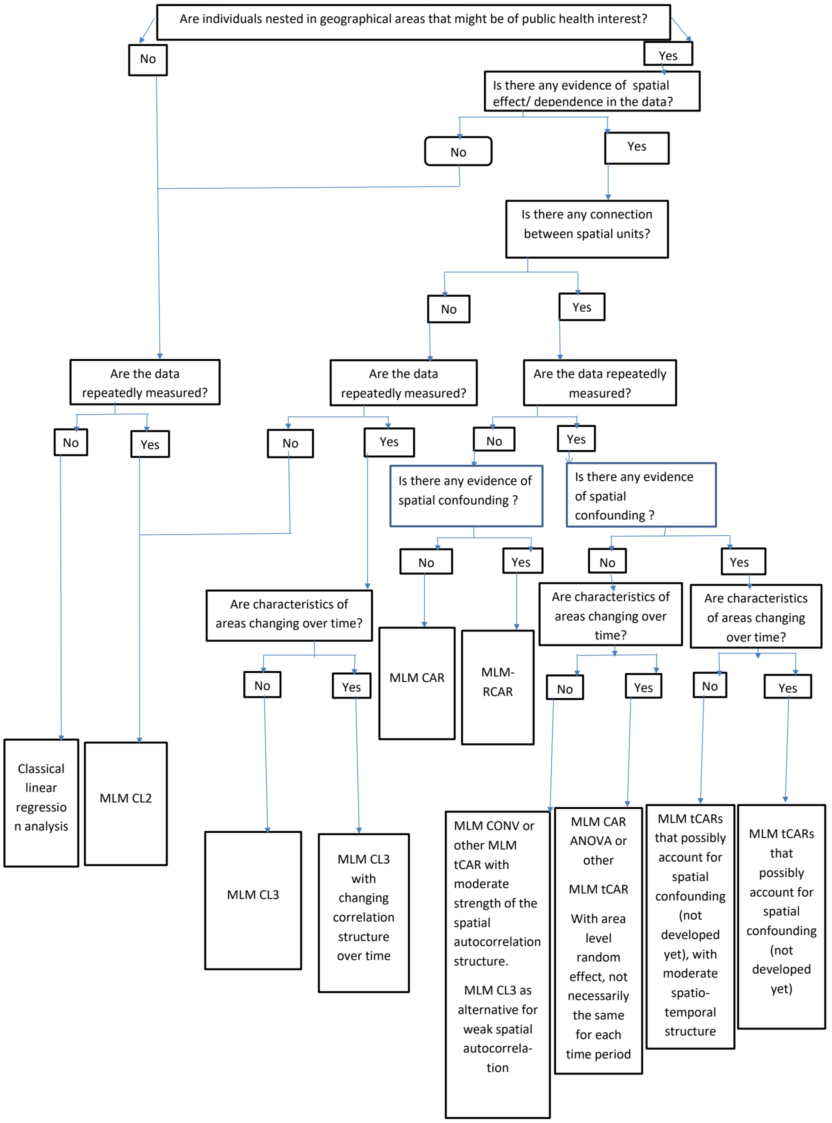
Are individuals nested in geographical areas that might be of public health interest?
No
Yes
Is there any evidence of spatial effect/ dependence in the data?
No
Yes
Is there any connection between spatial units?
No
Yes
Are the data repeatedly measured?
Are the data repeatedly measured?
Are the data repeatedly measured?
No
Yes
No
Yes
No
Yes
Is there any evidence of spatial confounding ?
Is there any evidence of spatial confounding ?
No
Yes
No
Yes
Are characteristics of areas changing over time?
Are characteristics of areas changing over time?
Are characteristics of areas changing over time?
No
Yes
No
Yes
No
Yes
MLM CAR
MLM-RCAR
Classical linear regressio n analysis
MLM CL2
MLM CL3
MLM CL3 with changing correlation structure over time
MLM CONV or other MLM tCAR with moderate strength of the spatial autocorrelation structure.
MLM CL3 as alternative for weak spatial autocorrela-tion
MLM CAR ANOVA or other
MLM tCAR
With area level random effect, not necessarily the same for each time period
MLM tCARs that possibly account for spatial confounding (not developed yet), with moderate spatio-temporal structure
MLM tCARs that possibly account for spatial confounding (not developed yet)

FIGURE 6.1: Decision tree for users, to analyse data for which participants are nested within geographical areas that might be of public health interest. The method suggestion is based on a tradeoff between accuracy to retrieve regression coefficients, model's goodness of fit, and time needed to complete the fit.

6.2.2 Decision tree for data given on the same spatial level

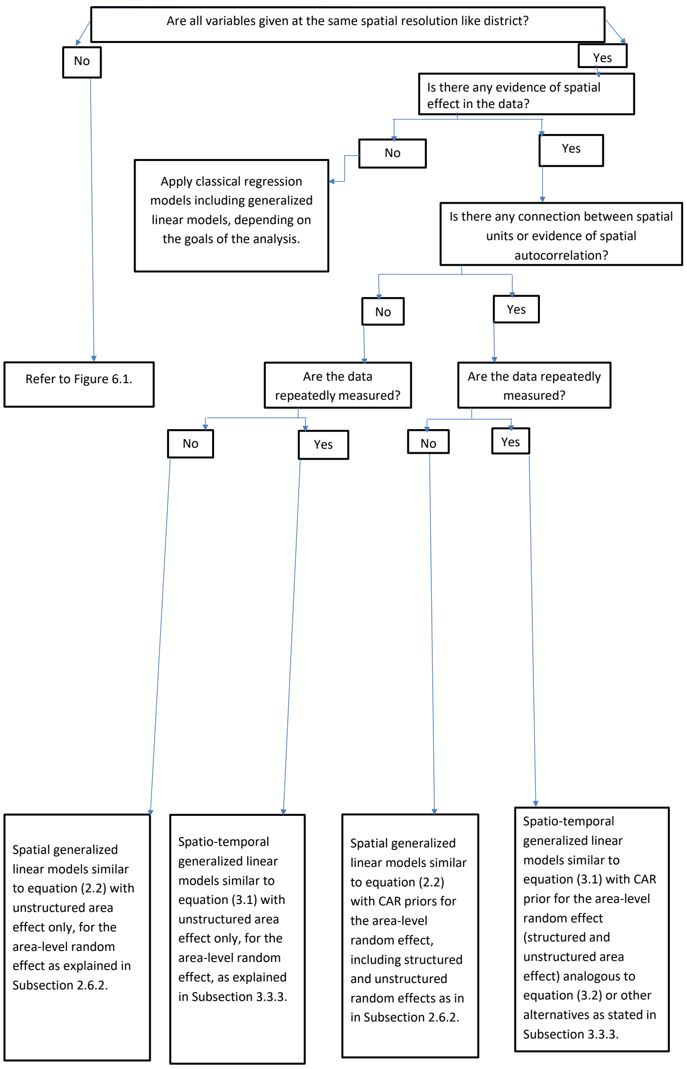
Are all variables given at the same spatial resolution like district?
No
Yes
Is there any evidence of spatial effect in the data?
No
Yes
Apply classical regression models including generalized linear models, depending on the goals of the analysis.
Is there any connection between spatial units or evidence of spatial autocorrelation?
No
Yes
Refer to Figure 6.1.
Are the data repeatedly measured?
Are the data repeatedly measured?
No
Yes
No
Yes
Spatial generalized linear models similar to equation (2.2) with unstructured area effect only, for the area-level random effect as explained in Subsection 2.6.2.
Spatio-temporal generalized linear models similar to equation (3.1) with unstructured area effect only, for the area-level random effect, as explained in Subsection 3.3.3.
Spatial generalized linear models similar to equation (2.2) with CAR priors for the area-level random effect, including structured and unstructured random effects as in in Subsection 2.6.2.
Spatio-temporal generalized linear models similar to equation (3.1) with CAR prior for the area-level random effect (structured and unstructured area effect) analogous to equation (3.2) or other alternatives as stated in Subsection 3.3.3.

FIGURE 6.2: Decision tree for users, to analyse data for which all variables are initially measured at an areal level or aggregated to a common spatial resolution that might be of public health interest. The method suggestion is based on a tradeoff between accuracy to retrieve regression coefficients, model's goodness of fit, and time needed to complete the fit.

6.3 Software overview

6.3.1 Introduction

For software implementations, the classical linear model, the MLM CL2, the MLM CL3 and some MLM CARs are available in software packages. However, the implementation for MLM tCARs is not always given in software packages. See Appendix B for the implementation of MLM tCARs developed in this thesis. The detailed information on the implementation of the methods from the decision trees in software packages are summarized in Table 6.1. It concentrates more on methods for data on their initial spatial resolutions, but most of the methods could be used for aggregated data as well. The implementation of the MLM CL2 is common in the R software R, 2021. We wish here to consider the implementation of the MLM CL2 using the Statistical Analysis System (SAS) SAS, 1985) to properly account for spatial heterogeneity is given in Section 6.3.2.

6.3.2 Implementation of the classical multilevel model in SAS: An example

For the following example we use the study population described in Chapter 2. Here, the outcome, exposure and covariates are identical to the one from section 5.4 of Chapter 5. In contrast to Chapter 5, we only consider the data from the baseline for this application.

All analyses are performed with the SAS 9.4 software (SAS, 1985). We use PROC MIXED for the linear mixed model (see Appendix B.6) and PROC GLMMIX for discrete responses, particularly the multilevel or random logistic model. SAS PROC MIXED is a flexible program suitable for fitting hierarchical or multilevel linear models. We do not aim to give detailed information on the mixed procedure of the SAS software. For a more comprehensive documentation, we refer to Gamst, Meyers, and Guarino, 2009 to get started with the SAS software, Singer, 1998 for PROC MIXED and Boykin et al., 2010 for PROC GLMMIX. Without the already and carefully developed procedures in SAS, few users would fit the models we would like in environmental epidemiology (McArdle, 2015, Sullivan and Greenland, 2014). However, as the model specification is more important, we give some key recommendations to non-experienced users of the multilevel or mixed models in order to easily perform accurate, reliable, and interpretable results from multilevel models, according to the research question of interest. Multilevel models can be formulated in two ways: the first is by presenting separate equations for each level like equations 4.1 for the first level then 4.2 and 4.3 for the higher level.

	Statistical Model	Usual R packages, implementation in R	Usual implementation in SAS
Cross-sectional design, classical non spatial methods	Classical multivariate linear, Poisson, Binomial, negative Binomial, log normal regression. There are non-parametric and semi-parametric methods as well.	Frequentist approach includes R-packages stats (lm function for linear and glm for generalized linear model), MASS (glm.nb), gamm4 (gamm, gamm4), nlme (lme) lme4 (lmer, lmer2, glmer), glmmADMB (glmmadmb) for mixed effect, gee (gee) for marginal models. Bayesian approach includes R package MCMCglmm, R2WinBUGS, RINLA, rstan. The list is not exhaustive.	Frequentist approach includes proc reg, proc glm for linear regression, and proc logistic, proc genmod, proc glimmix, proc nlmixed for generalised linear models. Bayesian approach include proc mcmc and built-in capabilities in the genmod procedure.
Cross-sectional design, classical spatial methods	Classical MLM (MLM CL2) Fixed Effect Methods (FEM) Genaralized Estimating Equations (GEEs)	Frequentist approach includes gamm4 (gamm, gamm4) lme4 (function lme), mgcv (gam, semi-parametric). Bayesian approach includes R packages MCMCglmm, R2WinBUGS, RINLA, rstan.	Frequentist approach includes proc mixed for linear regression, proc logistic, proc genmod and proc GEE for fixed or marginal effect, proc glimmix and proc nlmixed for generalized ME, Proc GAM for semi-parametric models Bayesian approach includes proc MCMC.
Cross-sectional design, spatial heterogeneity and spatial autocorrelation	Multilevel CAR (MLM CAR), Multilevel resricted CAR (MLM RCAR), Semi-parametric regression (generalized additive models).	Bayesian approach include R package MCMCglmm, R2WinBUGS, RINLA, rstan, CARBayes, HSAR.	Bayesian approach includes proc mcmc
Longitudinal design, classical methods	Univariate and multivariate ANOVA, Growth curve models, multilevel regression models (MLM CL3), SEM (Structural Equation Modeling) for longitudinal data, Fixed Effect Models (FEM), Generalized Estimating Equations (GEEs), Semi-parametric models.	Frequentist approach includes R packages gamm4 (gamm, gamm4) lme4 (lme) geepack and gee for Generalized Estimating Equation, multgee for multinomial response, CRTgeeDR. Semi-parametric methods includes mgcv (gam). Bayesian approach includes R package MCMCglmm, R2WinBUGS, RINLA, rstan.	Frequentist approach includes proc mixed for linear regression, proc logistic, proc genmod and proc gee for fixed or marginal effect, proc glimmix and proc nlmixed for generalized MLM. Bayesian approach includes proc mcmc
Longitudinal design, changing spatial effect, spatial autocorrelation and spatial heterogeneity	MLM tCARs for longitudnal data	Implementation using WinBUGS run from within R via the R-package R2WinBUGS	

TABLE 6.1: Summary of methods (and software packages) for the analysis of spatial data in cross-sectional and longitudinal design in epidemiology.

The second formulation is to combine all equations by substitution into a single model-equation, like equation 4.4. It is extremely important to notice that SAS PROC MIXED and PROC GLMMIX use the single equation representation, thus the substituted model. The first recommendation is then to write out (or have it written out) the substituted model mathematically before writing SAS code to fit the model. The research questions are generally easier to present in the hierarchical form (equations 4.1 for the first level then 4.2 and 4.3 for the higher level) but the SAS code should be written in the substituted form. As models get more complex, it is not always obvious how to parameterize the model so that the output can be used directly to answer your research question. Experience suggests that proceeding directly to PROC MIXED and PROC GLMMIX syntax is likely to produce output that is not what the user intended. Another important point is the centering of variables. Binary and Categorical variables do not need to be centered. For continuous variables, if we have a complicated random part, including random components for regression slope (varying exposure effect across units), we should think carefully about the scale (centered or standardized) of our explanatory variables. We may estimate the unstandardized results, including the random part of the model, and reanalyse the data using standardized (or centered) variables and compare. Unlike some specialized software programs like HLM 8 Software (Raudenbush et al., 2019) which ask whether you want to center variables, the data analyst must be proactive when using PROC MIXED and PROC GLMMIX.

About variable selection or the selection of the set of confounders, the user should decide which effect should vary across higher spatial units. For our analysis, we choose the intercept and the exposure of interest. The analysis is performed with selected risk factors entered sequentially. We always start with an unconditional model (i.e., a model that has no predictor) as this is used to compute the Intraclass Correlation Coefficient(ICC), the proportion of district-level variance compared to the total variance, which estimates how much variation in the outcome exists between level-2 units (districts) and gradually estimating more complex models (adjusting for some covariate) while checking for model improvement in model fit after each model is estimated. We use the data as well as prior information to critically evaluate our epidemiologic assumptions implied by the model and the statistical assumptions required by the model. Unfortunately, all model selections methods are subject to errors, and no optimal method for selecting the best model form is known (Greenland, 1989). Once more we choose simplicity and use stepwise regression with some forced-in covariates (some variables of interest, not subject to variable selection, including the exposure of interest), which are recognized (from previous studies or expert point of view) to influence the outcome of interest (known confounding factors) and the remaining available covariates are subject to selection by a forward-selection algorithm. To conventionally include the additional covariates, we use a tradeoff between charge-in-estimate method, in which covariates are selected based on the relative or absolute change in the estimated exposure effect, the statistical significance of the included covariates, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). statistical selection procedures based on significance tests alone, such as stepwise regression, can be particularly misleading (Greenland and Neutra, 1980).

Table 6.2 describes our model building process. Model 1 with no predictor will be used to quantify the variance explained at the district level and compute the mean outcome in the case of MLM CL2. Models 2 with additional individual level covariates and Model 3 with additional area level covariates are generally simultaneously applied.

	model 1	model 2	model 3
Construction	no predictor, just random effect for the intercept	model 1 + level-1 fixed effects+random coefficient for the exposure of interest	model 2 + level-2 fixed effects and cross-level interaction with exposure
Output, Interpretation	output used to compute the Intra-Class Correlation (ICC) and grand-mean value of the outcome	Indicate the relationship between level-1 covariates and health outcome and if the exposure varies across spatial units	reveal if level-2 predictors are responsible for the variability of the exposure effect across spatial units and the independent effect of level-2 covariates on the outcome

TABLE 6.2: Summary of the model building for the analysis of the association between depression and greenness, using the classical multilevel model (MLM CL2).

We find out how much of the variance in depression score is attributable to individuals and districts. Table 6.3 summarizes the variance explained at district level by district-level covariates.

	Bochum	Essen	Muelheim	Complete Area
Male	0%	0%	0.3%	0%
Female	0.5%	5%	0.9%	2%
Complete sample	0.00%	2%	2%	1%

TABLE 6.3: Percentage of variances explained at the highest level of the hierarchy for the analysis of the association between depressive symptoms and greenness.

The (spatial) Intra-Class correlation is particularly "small" in the male group. In the female population in Essen, the (spatial) intra class correlation is 5%. This tells us that there is a fair bit of clustering of depression scores within districts in the female population, suggesting the importance of contextual variables (see Table 6.3).

After including risk factors sequentially as indicated in table 6.2 of the model building process, the general result is that Greenness is negatively associated to depressive symptoms, after accounting for the spatial heterogeneity of the data. The association is stronger in the female population. Not considering the spatial structure of the data can lead to bias in the covariate effects, particularly in the female, where the intra-class correlation is non-zero, even if not substantial. Taking advantage of the spatial correlation can improve the model fit when the spatial correlation is present in the data, given by the AIC and BIC.

6.4 Discussion

The goal of this chapter was to provide decision trees useful in epidemiology, particularly when spatial effect is suspected. This was based on simulation studies when participants are nested within geographical units and data used on their initial spatial resolution. We also provided some tips to help choose the appropriate software packages when analysing data using regression analysis, particularly when a spatial effect is possibly present. In this thesis, we have applied different software packages to show the flexibility in implementing the regression models applied. The tips for software implementations were mostly suggested after several applications in different software packages. The most important part in the regression modelling involving spatial and spatio-temporal effects is the model specification rather than model fitting itself. An example of the implementation of the classical multilevel model is given in a popular software package, SAS. Choosing an appropriate software depends on the ability and background of each user. However, if the method has been correctly specified, software implementation becomes a secondary task. Software implementations are being developed very quickly, but the methods described here are still helpful when spatial and spatio-temporal effects are involved.

Chapter 7

Overall conclusion and outlook

This Ph.D. thesis addressed the problem of spatial effects when linking health outcomes to exposures with a spatial background like environmental exposures. Regression methods are appropriate to linking health outcome and exposure, accounting for other covariates. These regression methods were considered both for cross-sectional as well as for longitudinal studies. The thesis was divided into two parts. Part I focused on methods for data given on an areal level or aggregated to an areal level. Depending on the goals of the analysis, data may be aggregated to the same spatial resolution. In Chapter 2 (First Chapter of Part I), we considered exploratory analysis for aggregated data, cross-sectional analysis. As an example, we investigated spatial variation in analyses of the effects of urban greenness on depression using the data of the longitudinal HNRS, for some selected time points. The goal was to identify spatial clusters of elevated risks in the districts of the HNRS and analyse the dependence of the risk on covariates. Data were aggregated and methods to identify local clusters of elevated risk of depression in the study area as well as a method of the global indicator of spatial clustering were firstly applied exploratively. Then, a sophisticated spatial model for disease mapping within a Bayesian hierarchical model formulation was then described to estimate and smooth the risk of depression, accounting for covariate effects. The results suggested negative associations between greenness and depression as well as weak spatial effects. In Chapter 3 we considered a spatio-temporal extension of the spatial model in Chapter 2 to analyse longitudinal data. We investigated spatio-temporal variation in analyses of effects of urban greenness on depression by including spatio-temporal random effect terms in a Poisson model on the district level. The spatio-temporal model is available in the literature of spatial statistics but rarely applied to longitudinal studies. With this class of spatio-temporal models, we were able to accurately smooth the risk at an areal level, explain the association between health outcome (depression) and environmental exposure (greenness) and explain the spatial effect and its dynamic over time. The results in accordance with the results of Chapter 2 showed negative associations between greenness and depression. The findings suggest strong temporal autocorrelation and weak spatial effects. Even if the weak spatial effects are suggestive of neglecting them, as in our case, spatio-temporal random effects should be taken into account to provide reliable inference in urban health studies. We are aware that aggregating the data in Chapters 2 and 3 to the same spatial resolution could lead to ecological fallacy and some loss of information. This is why we considered methods to use data on their finest and initial level in Part II. In Chapter 4, the advantages of the MLM CL2 models were presented before pointing out some limitations, particularly in the presence of residual spatial autocorrelation. MLM CARs models were then introduced and compared to the MLM CL2 models in simulation studies, where different scenarios of the spatial effect were simulated. The results suggested in general that the MLM CAR and MLM RCAR performed better compared to the MLM CL2 in retrieving the estimated regression coefficients. We applied the three models comparatively on the analysis of the association between depressive symptoms and greenness. In contrast to Chapter 2, all data were used at their initial spatial resolution. The results showed a negative association between greenness and depressive symptoms, although not significant. In Chapter 5, we wished to extend the MLM CARs models for longitudinal data. Combining the advantages of MLM CL3 models and the properties of the MRF models with a structure changing over time, capable of capturing the dynamic of the spatio-temporal effects was the focus of Chapter 5. The CAR-prior models used in Chapter 5 are the same in the sophisticated models in Chapter 3. The difference is that the models in Chapter 3 are exclusively applied on areal level data while the spatio-temporal CAR models in Chapter 5 are combined with data at different spatial resolutions. From our knowledge, these models were not developed in the literature, at least in Epidemiological studies to account for the spatial effect and produce more accurate coefficient estimates of the association between individual-level health outcome and exposures, with covariates given at disparate spatial resolutions. We compared the developed MLM tCARs to the MLM CL3 via simulation studies in common spatial data situations. The results indicated the better performance of the MLM tCARs, to retrieve the true regression coefficients and with better fit in general. The MLM tCARs and MLM CL3 were also applied comparatively to the analysis of the association between greenness and depressive symptoms. Unlike the application in Chapter 4, all data were used at their initial spatial resolution for eight time points. The results showed a negative association between greenness and depression and a decreasing linear individual time trend. We also observed very weak spatial variations and moderate temporal autocorrelation. In Part III, Chapter 6 was dedicated to produce decision trees based on the results of our simulation studies in Chapters 4 and 5 as well as the applications of sophisticated methods from Chapter 2 and

Chapter 3. The decision trees were intended to help the users analyse epidemiological data for which participants are nested within geographical areas/units more comfortably. We also considered the case where data are exclusively given on an areal level or some data are aggregated to a common spatial resolution. In Chapter 6 from Part III, we also provided software tips for the methods described in the decision trees. In this thesis, we have applied different software packages to show the flexibility in implementing the regression models used. The most important part in regression modelling involving spatial and spatio-temporal effects is the model specification rather than model fitting itself.

To sum up, in the analysis of the association between health outcomes and environmental exposures for epidemiological studies with participants nested within geographical units and data given on disparate spatial resolutions, multilevel models are appropriate. Classical multilevel models are widely used because of their simplicity and ease of interpretation as well as the fact that the hierarchical structure of the data is used as an advantage, whereas ignoring it can lead to biased results compared to classical regression analysis (classical linear models for instance). However, classical multilevel models, either for cross-sectional or longitudinal studies do not account for possible residual spatial and/or spatio-temporal autocorrelation. MLM CARs and MLM tCARs models offer a good alternative and help to explain the spatial effect better. Implementing and interpreting MLM CARs and MLM tCARs models require a much greater amount of computational resources. Nevertheless, the choice of simplicity to automatically apply classical methods could lead to erroneous conclusions on the analysis of the association between health outcomes and environmental exposure. Relative to the effort taken to collect the data (sometimes more than 10 years), the effort of implementing and interpreting a more complicated model is irrelevant. The developed MLM tCARs models also have the advantages of explaining the dynamic of spatial effects over time, in addition to the advantages of multilevel CAR models for cross-sectional data. The advantage of multilevel models is particularly pronounced for epidemiological analysis where confidentiality is important, compared to spatial models that make use of the individual coordinates of participants. The individual spatial coordinates of the participants like the house number may not be made available for confidentiality, rather, only their membership to a spatial area like district, town, or postal code is available. MLM CARs and MLM tCARs models would still be used in these situations. We know that aggregating the data may lead to ecological fallacy and some loss of information. Nevertheless, depending on the goals of the analysis, data must be collected or aggregated to the same spatial resolution. An example is the analysis of the risk of depression in the HNRS. The first analytical goal was to identify clusters (districts) of elevated risks. A sophisticated hierarchical Poisson model incorporating spatial random effect using a CAR-prior specification was needed for a better inference. We recommend using automatically such models accounting for spatio-temporal random effects, even if the weak spatial effect is suggestive of neglecting as in the case of the HNRS.

Restricted MLM tCAR models for longitudinal data still need to be fully developed and were not part of our simulation study. In future analysis, we plan to consider the restricted MLM tCAR models in simulation studies. The simulation studies performed in Chapter 4 and Chapter 5 were performed for linear models. The methods could easily be extended to generalized linear models. A preliminary work (co-directed bachelor thesis in the faculty of statistics) showed similar results for the logit-link function.

Bibliography

Bauer, D. J., N. C. Gottfredson, D. Dean, and R. A. Zucker (2013). “Analyzing Repeated Measures Data on Individuals Nested within Groups: Accounting for Dynamic Group Effects”. In: *Psychological Methods* 18.1, pp. 1–14.

Beale, C. M., J. J. Lennon, J. M. Yearsley, M. J. Brewer, and D. A. Elston (2010). “Regression analysis of spatial data”. In: *Ecology Letters* 13.2, pp. 246–264.

Bernardinelli, L. et al. (2015). “Bayesian Analysis of Space-Time Variation in Disease Risk”. In: *Statistics in Medicine* 14, pp. 2433–2443.

Besag, J. and J. Newell (1991). “The Detection of Clusters in Rare Diseases”. In: *Journal of the Royal Statistical Society, Series A* 154.1, pp. 143–155.

Besag, J., J. York, and A. Mollié (1991). “Bayesian image restoration, with two applications in spatial statistics”. In: *Annals of the Institute of Statistical Mathematics* 43.1, pp. 1–20.

Best, N., C. Jackson, D. Lunn, D. Spiegelhalter, and A. Thomas (2012). *The BUGS Book : A Practical Introduction to Bayesian Analysis*. CRC Press.

Beyer, K. M. et al. (2014). “Exposure to Neighborhood Green Space and Mental Health: Evidence from the Survey of the Health of Wisconsin”. In: *International Journal of Environmental Research and Public Health* 11.3.

Blangiardo, M. and M. Cameletti (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. John Wiley and Sons.

Boykin, D. et al. (2010). “Generalized linear mixed model estimation using proc glmmix: Results from simulations when the data match, and when the model is misspecified”. In: *Annual Conference on Applied Statistics in Agriculture* 4, pp. 323–355.

Brook, S. P. and A. Gelman (1998). “General Methods for Monitoring Convergence of Iterative Simulations”. In: *Journal of Computational and Graphical Statistics* 7, pp. 434–455.

Browne, W.J., H. Goldstein, and J. Rasbash (2001). “Multiple membership multiple classification (MMMC) models”. In: *Statistical Modelling* 1, pp. 103–124.

Bryk, A. S. and S. W. Raudenbush (1989). *Multilevel analysis of educational data*. Academic Press.

Clogg, C. C., E. Petkova, and A. Haritou (1995). “Statistical Methods for Comparing Regression Coefficients Between Models”. In: *American Journal of Sociology* 100.5, pp. 1261–1293.

Congdon, P. D. (2010). *Applied Bayesian Hierarchical Methods*. CRC Press, New York.

Cressie, N. (1993). *Statistics for Spatial Data, revised edition*. John Wiley and Sons, inc., pp. 14–15.

Djeudeu, D., M. Engel, K.-H. Jöckel, S. Moebus, and K. Ickstadt (2020). “Spatio-temporal analysis of the risk of depression at district-level and association with greenness based on the Heinz Nixdorf Recall Study”. In: *Spatial and Spatio-temporal Epidemiology* 33, p. 100340.

Djeudeu, D., S. Moebus, and K. Ickstadt (2022). “Multilevel Conditional Autoregressive models for longitudinal and spatially referenced epidemiological data”. In: *Spatial and Spatio-temporal Epidemiology* 41, p. 100477.

Dong, G. and R. Harris (2015). “Spatial Autoregressive Models for Geographically Hierarchical Data Structures”. In: *Geographical Analysis* 47, pp. 173–191.

Dong, G., R. Harris, K. Jones, and J. Yu (2015a). “Multilevel Modelling with Spatial Interaction Effects with Application to an Emerging Land Market in Beijing, China”. In: *PLoS ONE* 10.6, e0130761.

Dong, G., J. Ma, R. Harris, and G. Pryce (2015b). “Spatial Random Slope Multilevel Modeling Using Multivariate Conditional Autoregressive Models: A Case Study of Subjective Travel Satisfaction in Beijing”. In: *Annals of the American Association of Geographers* 106.1, pp. 19–35.

Dormann, C. F. (2007). “Assessing the validity of autologistic regression”. In: *Ecological Modelling* 207.2-4, pp. 234–242.

Draper, D. (1995). “Inference and Hierarchical Modeling in the Social Sciences”. In: *Journal of Educational and Behavioral Statistics* 20.2, pp. 115–147.

Duncan, C., K. Jones, and G. Moon (1998). “Context, composition and heterogeneity: using multilevel models in health research”. In: *Social Science & Medicine* 46, pp. 97–117.

Galea, S. and D. Vlahov (2005). “Urban health: evidence, challenges, and directions”. In: *Annual Review of Public Health* 26.

Gamst, G., L. S. Meyers, and A. J. Guarino (2009). *SAS 9 study guide : preparing for the base programming certification exam for SAS 9*. Cambridge Univ. Press.

Gelfand, A. E. (2000). “Gibbs Sampling”. In: *Journal of the American Statistical Association* 95, pp. 1300–1304.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). “Efficient parametrizations for normal linear mixed models”. In: *Biometrika* 82, pp. 479–488.

Geweke, J. (1992). “Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments”. In: *In: Bernardo JM, Berger JO, Dawid AP and*

Smith AFM, Editors., Bayesian Statistics, Vol. 4, Clarendon Press, Oxford, pp. 169–193.

Goldstein, H., W. Browne, and J. Rasbash (2002). “Multilevel modelling of medical data”. In: *Statistics in Medicine* 21.21, pp. 3291–3315.

Gómez-Rubio, V., J. Ferrándiz-Ferragud, and A. López-Quílez (2005). “Detecting clusters of disease with R”. In: *Journal of Geographical Systems* 7.2, pp. 189–206.

Greenland, S. (1989). “Modeling and variable selection in Epidemiologic analysis”. In: *American Journal of Public Health* 79.

Greenland, S. and R. Neutra (1980). “Control of Confounding in the Assessment of Medical Technology”. In: *International Journal of Epidemiology* 9.

Grimm, K. J., N. Ram, and F. Hamagami (2011). “Nonlinear Growth Curves in Developmental Research”. In: *child development journal* 82, : 1357—1371.

Hautzinger, M. and M. Bailer (2012). *Allgemeine Depressions Skala (ADS) [General Depression Scale; in German]*. Hogrefe Verlag GmbH & Co. KG.

Heidelberger, P. and P. D. Welch (2010). “A spectral method for confidence interval generation and run length control in simulations”. In: *Communications of the ACM* 24, pp. 233–245.

Hodges, J. S. (2014). *Random Effects Old and New in Richly Parameterized Linear Models Additive, Time Series,and Spatial Models Using Random Effects*. CHAPMAN & HALL-CRC.

Hodges, S. J. and B. J. Reich (2010). “Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love”. In: *The American Statistician* 64.4, pp. 325–334.

Hoef, J. M. Ver, E. E. Peterson, M. B. Hooten, E. M. Hanks, and M. J. Fortin (2018). “Spatial autoregressive models for statistical inference from ecological data”. In: *Ecological Monographs* 88, pp. 31–59.

Hongwei, X. (2014). “Comparing Spatial and Multilevel Regression Models for Binary Outcomes in Neighborhood Studies”. In: *Sociological Methodology* 44, pp. 229–272.

Julian, M. W. (2001). “The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling”. In: *Structural Equation Modeling* 8, pp. 325–352.

Knorr-Held, L. (2000). “Bayesian Modelling of Inseparable Space-Time Variation in Disease Risk”. In: *Statistics in Medicine* 19, pp. 2555–2567.

Lang, M., B. Bischl, and D. Surmann (2017). “batchtools: Tools for R to work on batch systems”. In: *The Journal of Open Source Software* 2, 10.21105/joss.00135.

Latouche, A., C. Guihenneuc-Jouyaux, C. Girard, and D. Hémon (2007). “Robustness of the BYM model in absence of spatial variation in the residuals”. In: *International Journal of Health Geographics* 6, p. 39.

Lee, D. (2013). “CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors”. In: *Journal of Statistical Software* 55.13, pp. 1–24.

Lee, D. and A. Lawson (2016). "Quantifying the Spatial Inequality and Temporal Trends in Maternal Smoking Rates in Glasgow". In: *Annals of Applied Statistics* 10, pp. 1427–1446.

Lee, D., A. Rushworth, and G. Napier (2018). "Spatio-Temporal Areal Unit Modeling in R with Conditional Autoregressive Priors Using the CARBayesST Package". In: *Journal of Statistical Software* 84.9.

Lee, J. and D. W. S. Wong (2001). *Statistical analysis with ArcView GIS*. New York: John Wiley and Son. New York: John Wiley and Son.

Leroux, B. G., X. Lei, and N. Breslow (2000). "Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence". In: *In: Halloran M.E., Berry D. (eds) Statistical Models in Epidemiology, the Environment, and Clinical Trials* 116, pp. 179–191.

Lindgren, F. and H. Rue (2015). "Bayesian Spatial Modelling with R-INLA". In: *Journal of Statistical Software* 63.19, pp. 1–25.

Lindley, D. V. and A. F. M. Smith (1972). "Bayes estimates for the linear model (with discussion)". In: *Journal of the Royal Statistical Society, Serie B* 34, pp. 1–41.

McArdle, P. F. (2015). "An Aid to Generating Figures for the American Journal of Epidemiology Using SAS/GRAPH". In: *American Journal of Epidemiology* 182.9, 747–749.

Moran, P. A. P. (1950). "Note on continuous stochastic phenomena". In: *Biometrika* 37, pp. 17–23.

Napier, G., D. Lee, C. Robertson, and A. Lawson (2019). "A Bayesian space-time model for clustering areal units based on their disease trends". In: *Biostatistics* 20, pp. 681–697.

Napier, G., D. Lee, C. Robertson, A. Lawson, and K. Pollock (2016). "A Model to Estimate the Impact of Changes in MMR Vaccination Uptake on Inequalities in Measles Susceptibility in Scotland". In: *Statistical Methods in Medical Research* 25, pp. 1185–1200.

Neal, R. M. (1997). "Slice sampling". In: *Annals of Statistics* 31, pp. 705–767.

Nezlek, J. B. (2001). "Multilevel Random Coefficient Analyses of Event- and Interval-Contingent Data in Social and Personality Psychology Research". In: *Personality and Social Psychology Bulletin* 27.7, pp. 771–785.

Nutsford, D., A. L. Pearson, and S. Kingham (2013). "An ecological study investigating the association between access to urban green space and mental health". In: *Public Health* 127.11.

Orban, E., R. Sutcliffe, N. Dragano, K. H. Jöckel, and S. Moebus (2017). "Residential Surrounding Greenness, Self-Rated Health and Interrelations with Aspects of Neighborhood Environment and Social Relations". In: *Journal of Urban Health* 94.2, pp. 158–169.

Orban, E. et al. (2016). “Residential Road Traffic Noise and High Depressive Symptoms after Five Years of Follow-up: Results from the Heinz Nixdorf Recall Study”. In: *Environ Health Perspect* 124.5, pp. 578–585.

Orcutt, G. H., H. W. Watts, and J. B. Edwards (1968). “Data Aggregation and Information Loss”. In: *The American Economic Review* 58.4, pp. 773–787.

Paddock, S. M., T. J. Leininger, and S. B. Hunter (2016). “Bayesian Restricted Spatial Regression for Examining Session Features and Patient Outcomes in Open-Enrollment Group Therapy Studies”. In: *Statistics in Medicine* 35, pp. 97–114.

Pfeiffer, D. U. et al. (2008). *Spatial Analysis in Epidemiology*. Oxford University Press. ISBN: 978-0198509899.

Pickett, K. E. and M. Pearl (2001). “Multilevel analyses of neighbourhood socioeconomic context and health outcome: a critical review”. In: *Journal of Epidemiology and Community Health* 55.2, pp. 111–122.

R, C. T. (2021). “R: A Language and Environment for Statistical Computing”. In: URL: https://www.R-project.org/.

Radloff, L. S. (1977). “The CES-D Scale: a self-report depression scale for research in the general population”. In: *Applied Psychological Measurement* 1.3, pp. 385–401.

Raudenbush, S. W., A. S. Bryk, Y. F. Cheong, R. T. Congdon, and M. D. Toit (2019). *HLM 8, Hierarchical Linear and Nonlinear Modeling*. Scientific Software International , Inc.

Reich, J. B., J. S. Hodges, and V. Zadnik (2006). “Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models”. In: *Biometrics* 62.4, pp. 1197–1206.

Reineveld, S. A. (1998). “The impact of individual and area characteristics on urban socioeconomic differences in health and smoking”. In: *International Journal of Epidemiology* 27.1, pp. 33–40.

Rhew, I. C., A. V. Stoep, A. Kearney, N. L. Smith, and M. D. Dunbar (2011). “Validation of the Normalized Difference Vegetation Index as a measure of neighborhood greenness”. In: *Annals of Epidemiology* 21.12, pp. 946–952.

Roberts, G. and J. Rosenthal (1998). “Optimal scaling of discrete approximations to the Langevin diffusions”. In: *Journal of the Royal Statistical Society, Series B* 60.1, pp. 255–268.

Robins, J. M. and S. Greenland (1986). “The role of model selection in causal inference from nonexperimental data”. In: *American Journal of Epidemiology* 123.3, pp. 392–402.

Roux, A. V. Diez (1998). “Bringing context back into epidemiology: variables and fallacies in multilevel analysis”. In: *American Journal of Public Health* 88.2, pp. 216–222.

— (2000). “Multilevel analysis in public health research”. In: *Annual Review of Public Health* 21, pp. 171–192.

Rushworth, A., D. Lee, and R. Mitchell (2014). "A Spatio-Temporal Model for Estimating the Long-Term Effects of Air Pollution on Respiratory Hospital Admissions in Greater London". In: *Spatial and Spatio-temporal Epidemiology* 10, pp. 29–38.

Rushworth, A., D. Lee, and C. Sarran (2017). "An Adaptive Spatio-Temporal Smoothing Model for Estimating Trends and Step Changes in Disease Risk". In: *Journal of the Royal Statistical Society C* 66, pp. 141–157.

SAS, I. (1985). *SAS user's guide: Statistics*. Vol. 2. Sas Inst.

Schmermund, A. et al. (2002). "Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL Study". In: *American Heart Journal* 144.2, pp. 212–218.

Shmueli, G. (2010). "To Explain or to Predict?" In: *Statistical Science* 25.

Singer, J. D. (1998). "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models". In: *Journal of Educational and Behavioral Statistics* 4, pp. 323–355.

Smith, T. E. and K. L. Lee (2012). "The effects of spatial autoregressive dependencies on inference in ordinary least squares: a geometric approach". In: *Journal of Geographical Systems* 14, 91—124.

Sondermann, W. et al. (2020). "Psoriasis, Cardiovascular risk factors and metabolic disorders: sex-specific findings of a population-based study". In: *Journal of the European Academy of Dermatology and Venereology* 34, pp. 779–786.

Song, H. et al. (2019). "Association between Urban Greenness and Depressive Symptoms: Evaluation of Greenness Using Various Indicators". In: *International Journal of Environmental Research and Public Health* 16.2.

Spiegelhalter, D., N. Best, B. Carlin, and A. Van der Linde (2002). "Bayesian Measures of Model Complexity and Fit". In: *Journal of the Royal Statistical Society, Series B* 64, pp. 583–639.

Steele, F. (2008). "Multilevel models for longitudinal data". In: *Journal of the Royal Statistical Society, Series A* 171.1, pp. 5–19.

Sturtz, S., U. Ligges, and A. Gelman (2005). "R2WinBUGS: A Package for Running WinBUGS from R". In: *Journal of Statistical Software* 12, pp. 1–16.

Sullivan, S. G. and S. Greenland (2014). "Bayesian regression in SAS software". In: *Int J Epidemiol* 43, pp. 1667–8.

Tobler, W. (1979). *Cellular geography, pages 379-386*. Reidel, Dordrecht, Holland.

Tomita, A. et al. (2017). "Green environment and incident depression in South Africa: a geospatial analysis and mental health implications in a resource-limited setting". In: *Lancet Planet Health* 1.4, pp. 152–162.

Tzivian, L. et al. (2016). "Long-Term Air Pollution and Traffic Noise Exposures and Mild Cognitive Impairment in Older Adults: A Cross-Sectional Analysis of the

Heinz Nixdorf Recall Study". In: *Environmental Health Perspectives* 124, pp. 1361–1368.

Wakefield, J. (2009). "Multi-level modelling, the ecologic fallacy, and hybrid study designs". In: *International Journal of Epidemiology* 38.2, pp. 330–336.

Waller, L. A. and C. A. Gotway (2004). *Applied Spatial Statistics for Public Health Data*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471662679.

Weber, T. et al. (2019). "Cross-sectional analysis of pulsatile hemodynamics across the adult life span: reference values, healthy and early vascular aging: the Heinz Nixdorf Recall and the MultiGeneration Study". In: *Journal of hypertension* 37, pp. 2404–2413.

Wen, S. W., K. Demissie, D. August, and G. G. Rhoads (2001). "Level of aggregation for optimal epidemiological analysis: the case of time to surgerey and unnecessary removal of normal appendix". In: *Journal of Epidemiology and Community Health* 55, pp. 198–203.

White, M. P., I. Alcock, B. W. Wheeler, and M. H. Depledge (2013). "Would you be happier living in a greener urban area? A fixed-effects analysis of panel data". In: *Psychological Science* 24.6, pp. 920–928.

White, P., A. Gelfand, and T. Utlaut (2017). "Prediction and model comparison for areal unit data". In: *Spatial Statistics* 22, pp. 89–106.

WHO (2016). "Urban green spaces and health, a review of evidence". In: *World Health Organization Regional Office for Europe, Copenhagen. http://www.euro.who.int/__data/assets/pdf_file/0005/321971/Urban-green-spaces-and-health-review-evidence.pdf?ua=1 (accessed 29 November 2019).*

— (2018). "Depression". In: *World Health Organization, https://www.who.int/en/news-room/fact-sheets/detail/depression (accessed 29 November 2019).*

Wu, J. and L. Jackson (2017). "Inverse relationship between urban green space and childhood autism in California elementary school districts". In: *Environment International* 107, pp. 140–146.

Appendix A

Additional materials

A.1 Tables and figures

Scenario No	τ_S^2	ρ_S	τ_T^2	ρ_T	Meaning/Interpretation
1	0.09	0.5	0.8	0.5	weak spatial effect, medium spatial heterogeneity and autocorrelation, medium temporal effect
2	0.009	0.9	3	0.9	weak spatial effect, mainly spatial autocorrelation, strong temporal effect
3	0.8	0.5	3	0.09	medium spatial effect, medium spatial heterogeneity and medium spatial autocorrelation, strong temporal effect
4	0.8	0.5	0.8	0.5	medium spatial effect, medium spatial heterogeneity, medium temporal effect
5	0.8	0.9	0.8	0.9	moderate spatial effect, mainly spatial autocorrelation, medium temporal effect
6	3	0.5	3	0.09	strong spatial effect, medium spatial autocorrellation, strong temporal effect
7	3	0.09	3	0.9	strong spatial effect, mainly spatial autocorrellation, strong temporal effect
8	3	0.5	0.8	0.5	strong spatial effect, medium spatial autocorrellation, medium temporal effect
9	3	0.9	3	0.9	strong spatial effect, mainly spatial medium spatial autocorrellation, strong temporal effect

TABLE A.1: The selected scenarios of the simulated spatio-temporal effect.

Scenario number	Value of τ^2	Value of ρ	Meaning/Interpretation
1	1	0.95	moderate spatial effect, mainly spatial autocorrellation
2	1	0.09	moderate spatial effect, mainly spatial heterogeneity
3	1	0.6	moderate spatial effect, medium spatial heterogeneity and medium spatial autocorrelation
4	10	0.09	strong spatial effect, mainly spatial spatial heterogeneity
5	10	0.95	strong spatial effect, mainly spatial autocorrellation
6	10	0.6	strong spatial effect, medium spatial heterogeneity and medium spatial autocorrelation
7	0.01	0.95	weak spatial effect, mainly spatial autocorrellation
8	0.01	0.09	weak spatial effect, mainly spatial heterogeneity
9	0.01	0.6	weak spatial effect, medium spatial heterogeneity and medium spatial autocorrelation

TABLE A.2: The different scenarios of the simulated spatial effect. The scenarios with gray color are the ones presented in the thesis.

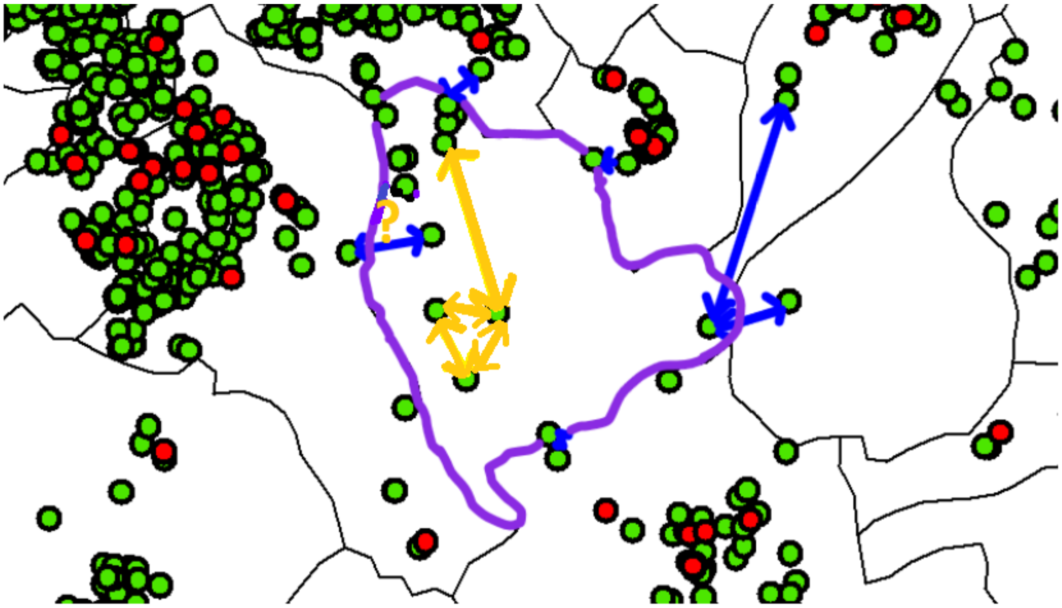

FIGURE A.1: Description of spatial effects for the example of the Heinz Nixdorf Recall Study: The red/green points represent participants' positions with high/no high value of depression score, the yellow arrows indicate the spatial proximity for participants in the same geographical unit, while the blue ones are for participants in adjacent units. Two participants in adjacent units may be closer and have more similar outcome than participants in the same spatial unit (yellow arrows). This imply that both spatial heterogeneity and spatial autocorrelation should be accounted for.

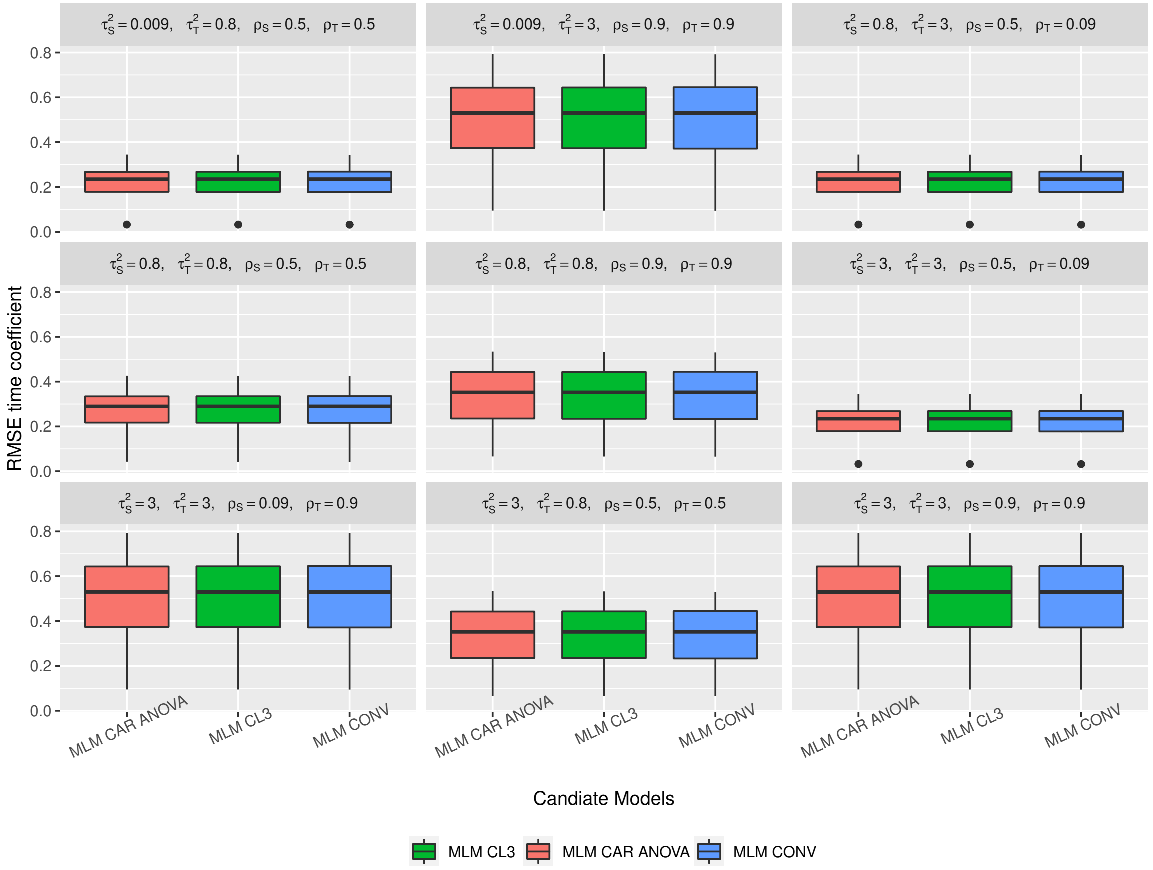

FIGURE A.2: Comparison of the Root Mean Square Error (RMSE) for the time variable coefficient (growth), for the set of selected scenarios of the simulated spatial effect. The true value for the time coefficient is -0.1. τ_S^2 and ρ_S, τ_T^2 and ρ_T are overall variance and autocorrelation parameters from equation (A.1), for space and time respectively.

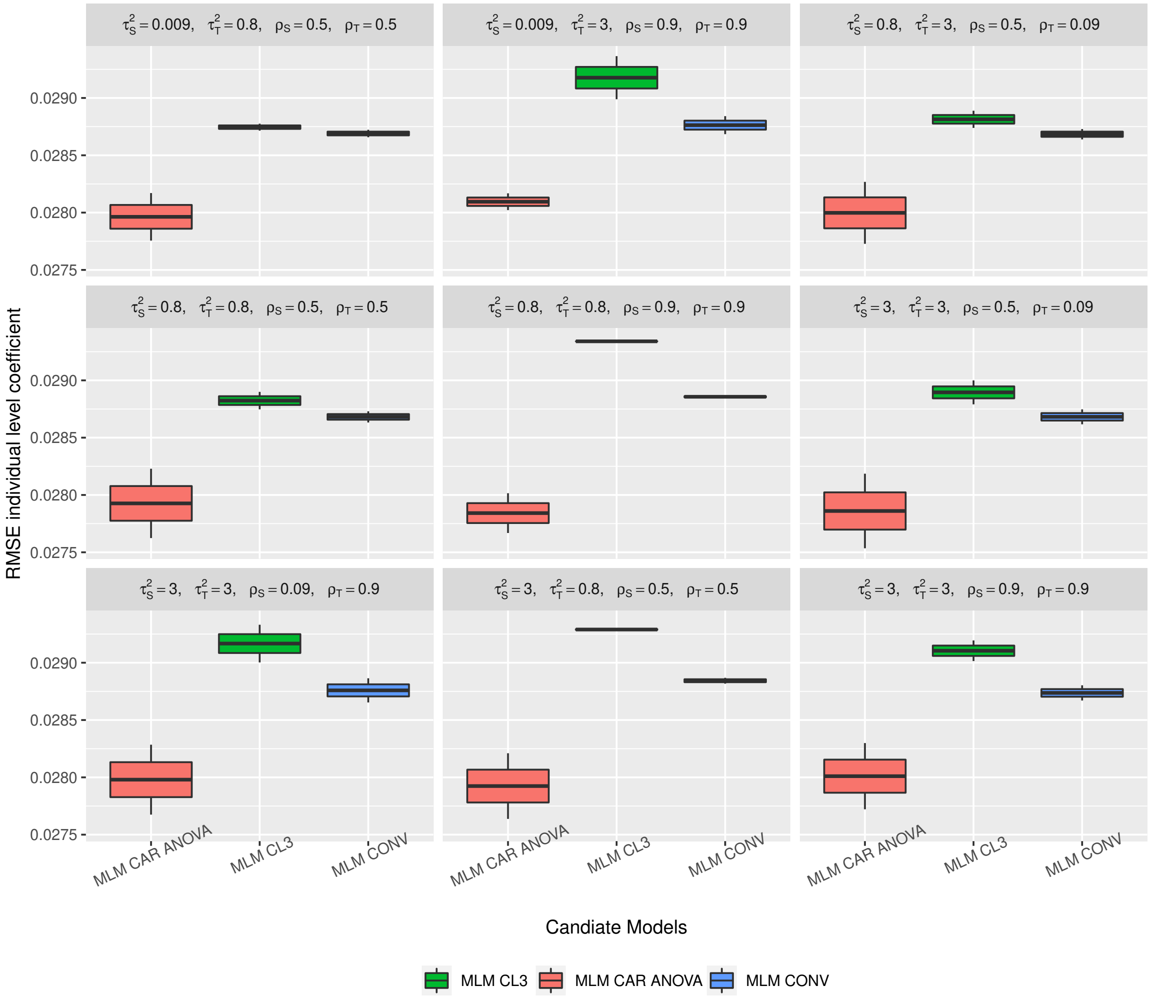

FIGURE A.3: Comparison of the Root Mean Square Error (RMSE) for the individual level variable coefficient, for the set of selected scenarios of the simulated spatial effect. The true value for the individual level coefficient is −1.72. τ_S^2 and ρ_S, τ_T^2 and ρ_T are overall variance and autocorrelation parameters from equation (A.1), for space and time respectively.

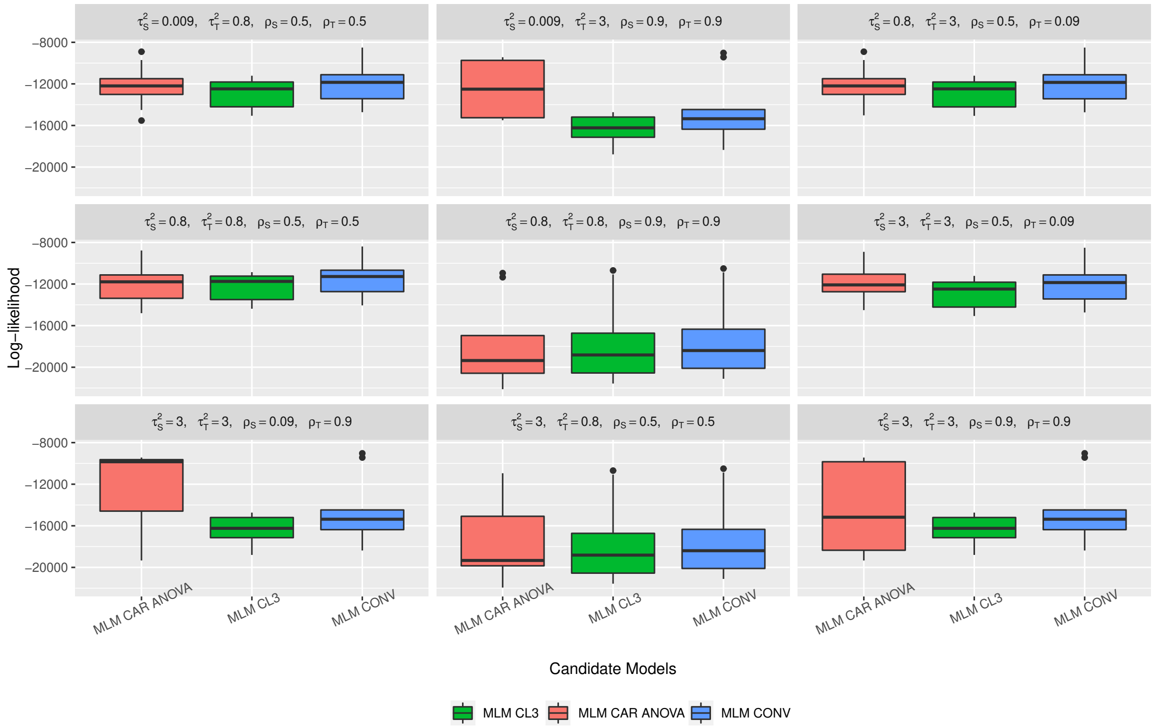

FIGURE A.4: Comparison of the posterior log-likelihoods, for a set of selected scenarios of the simulated spatio-temporal effect, longitudinal. τ^2_S and ρ_S, τ^2_T and ρ_T are overall variance and autocorrelation parameters from equation (A.1), for space and time respectively.

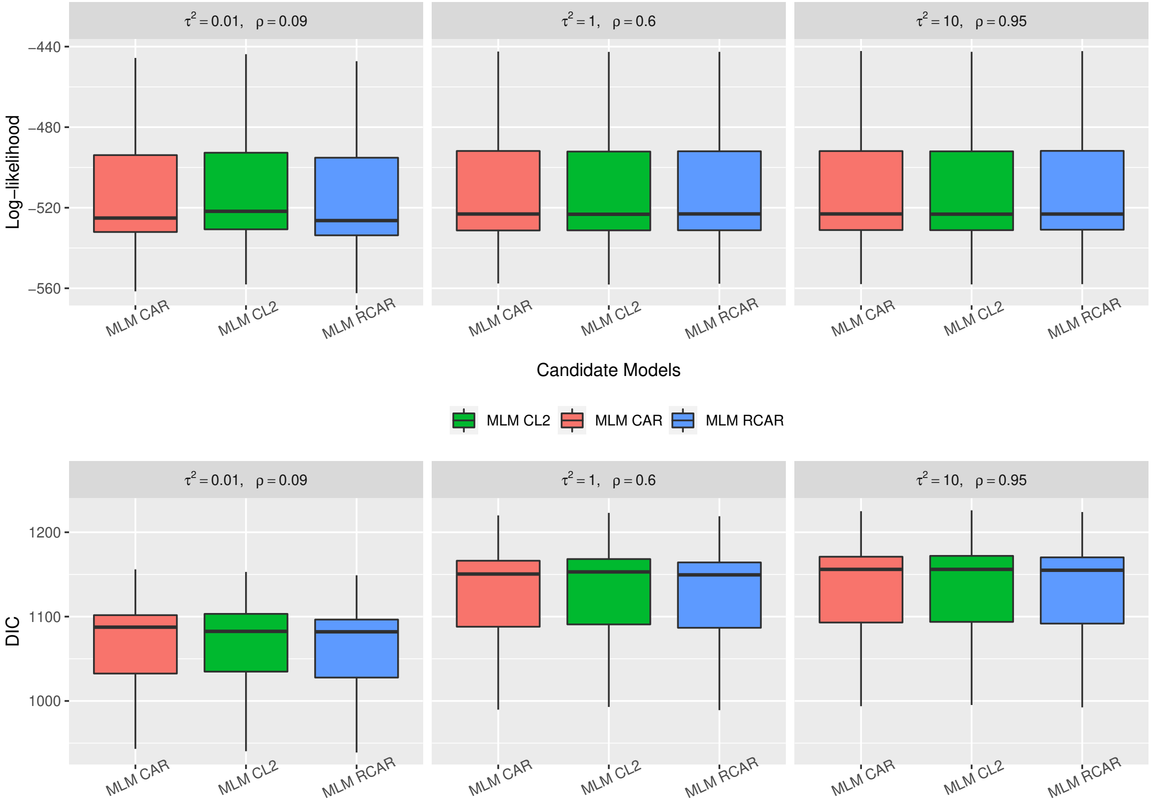

FIGURE A.5: Comparison of the goodness of fit, for a set of selected scenarios of the simulated spatial effect, cross-sectional. τ^2 and ρ are the overall spatial variance and autocorrelation parameters from equation (A.2), respectively.

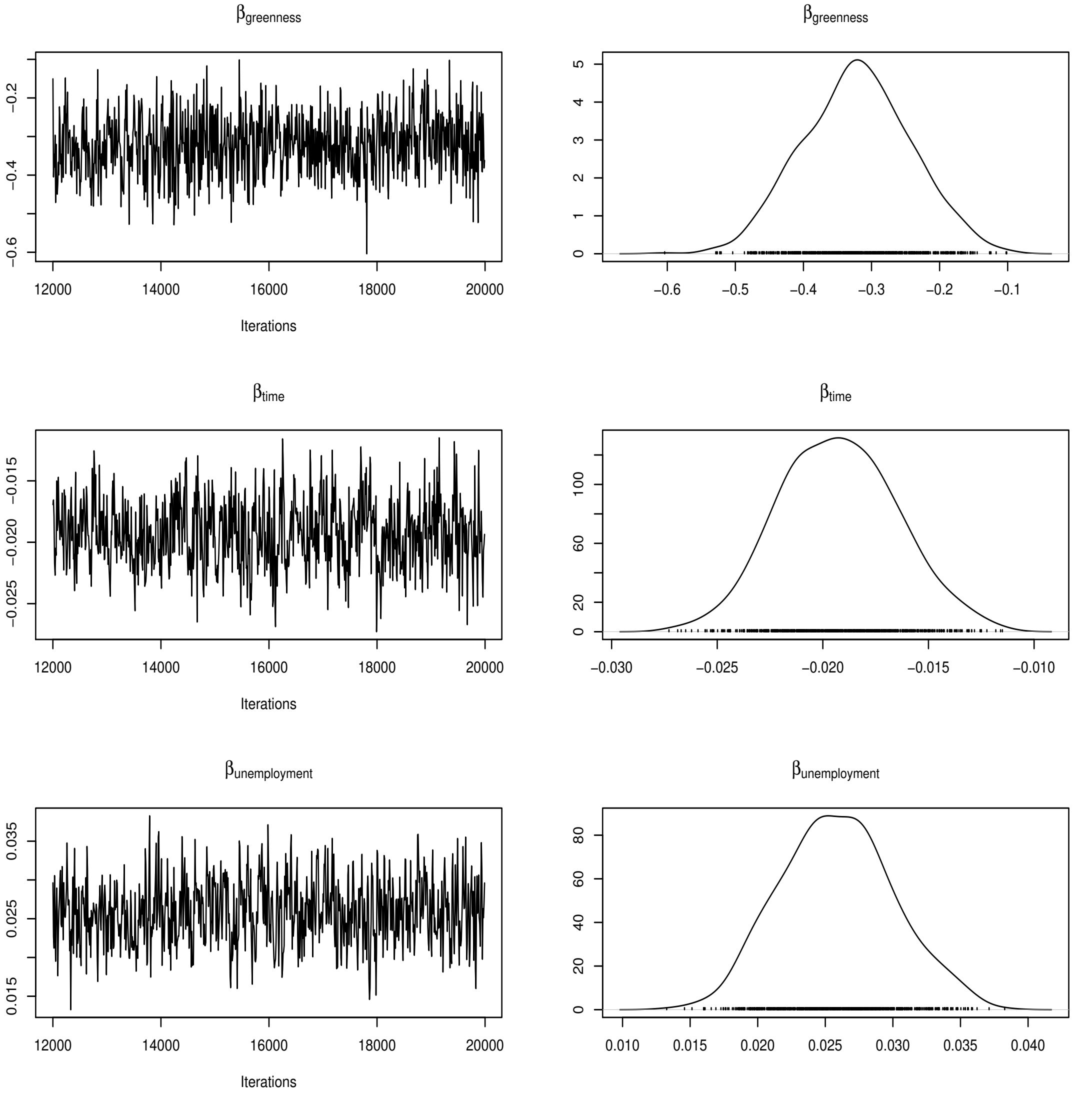

FIGURE A.6: Post diagnostic plots (trace and density plots) of the fixed effect parameters of interest.

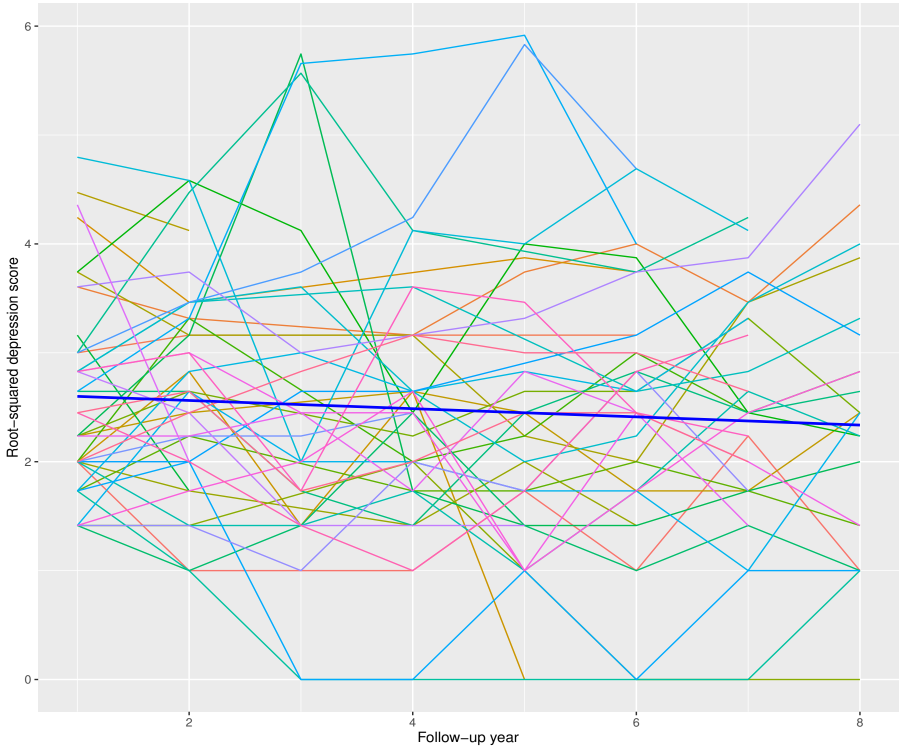

FIGURE A.7: Exploratory plot of the individual trajectories and mean trajectory for randomly selected 50 participants, to identify average individual trends within subjects. The dense linear blue line represents the means trajectory. It is indicative of a linear decreasing trend.

A.2 Simulation of the spatio-temporal effect

We use equation (A.1), step by step, to generate the random effect ψ and then the dependent variable y.

$$\begin{cases} \Phi_{vec} &= (\Phi_1, \Phi_2, \dots, \Phi_N)^T \;\text{ with }\; \Phi_t \sim \mathcal{N}_K(0_K, \tau_S^2 Q^{-1}), \;\; t = 1, \dots, N, \\ \Delta_{vec} &= \Delta \otimes (1, \dots, K)^T \;\text{ with }\; \Delta \sim \mathcal{N}_N(0_N, \tau_T^2 Q_D^{-1}) \\ \Psi_{vec} &= \Phi_{vec} + \Delta_{vec}, \\ \psi_{tj} &= (\Psi_{vec})_{(t-1)\cdot K + j}, \\ y_{tij} &= X_{tij}^T \beta + \psi_{tj} + r_{0ij} + r_{1ij} t + e_{tij}, \; t = 1, \dots, N, \;\; i = 1, \dots n, \; j = 1, \dots, K, \end{cases} \tag{A.1}$$

where

$Q = \rho_S R + (1-\rho_S)I,\ \ R_{jk} = -w_{jk},\ j \neq k,\ R_{jj} = \sum_{k \neq j} w_{jk},\ \ j,k \in \{1,\ldots,K\}$, $Q_D = \rho_T R^D + (1-\rho_T)I^D,\ \ R^D_{tl} = -d_{tl}$ for $t \neq l$, $R^D_{tt} = \sum_{l \neq t} d_{tl},\ \ t,l \in \{1,\ldots,N\}$, $X^T_{tij}\beta = \beta_0 + \beta_1 t + \beta_2 x_{tij} + \beta_3 h_{ij} + \beta_4 z_{tj}$. I and I^D are K and N dimensional unit matrices, respectively. $\otimes$ denotes the Kronecker Product.

A.3 Simulation of the spatial effect

We use equation (A.2) step by step, to generate the random effect ψ and then the dependent variable y.

$$\begin{cases} (\psi_1, \psi_2, \ldots, \psi_K)^T \sim \mathcal{N}_K(0, \tau^2 Q^{-1}), \\ y_{ij} = X^T_{ij}\beta + \psi_j + e_{ij}, i = 1,\ldots,N,\ \ j = 1,\ldots,K, \end{cases} \tag{A.2}$$

with $X^T_{ij}\beta = \beta_0 + \beta_1 x_{1ij} + \beta_2 z_{1j}$. I denotes a $K \times K$ unit matrix. τ^2 is a variance parameter that controls the strength of the overall spatial structure.

A.4 Full conditionals for the parameters of interest, MLM tCARs

Interest might centre on the global regression coefficients β, the random effects being introduced merely to permit the assumption of conditional independence. We provide the general path to find full conditionals for all parameters for the Gibbs sampler without expressing a definitive value for a single parameter.

Once more, we consider equation 5.1 of Chapter 5, with the example of the convolution model for the spatio-temporal random effect. As prior forβ, $\beta \sim N(0, \Sigma_\beta)$. Let us write equation 5.1 in a more general and matrix form:

$E(Y) = \eta = X\Lambda = X_0\Lambda_0 + X_1\Lambda_1 + X_2\Lambda_2 + X_3\Lambda_3 + X_4\Lambda_4$, where $X = (X_0, X_1, X_2, X_3, X_4)$ is the design matrix. We suppose that there are n_i participants in area i, $i = 1,\ldots,K$. $\sum_{i=1}^K n_i = n$ is the number of participants.

$$\begin{aligned} Y \;=\; & (y_{t_1 11}, y_{t_2 11}, \ldots, y_{T11}, y_{t_1 21}, y_{t_2 21}, \ldots, y_{T21}, \ldots, y_{t_1 n_1 1}, y_{t_2 n_1 1}, \ldots, y_{T n_1 1}, \\ & y_{t_1 (n_1+1)2}, y_{t_2 (n_1+1)2}, \ldots, y_{T(n_1+1)2}, \ldots, y_{t_1 (n)K}, y_{t_2 (n_1+1)2}, \ldots, y_{T(n)K})' \end{aligned}$$

X_0 is a $N \times p$ matrix

$$X_0 = \begin{pmatrix}
1 & g(t_1) & x_{t_1 11} & \cdots & h_{11} & \cdots & z_{t_1 1} \\
1 & g(t_2) & x_{t_2 11} & \cdots & h_{11} & \cdots & z_{t_2 1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & g(T) & x_{T11} & \cdots & h_{11} & \cdots & z_{T1} \\
1 & g(t_1) & x_{t_1 21} & \cdots & h_{21} & \cdots & z_{t_1 1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & g(T) & x_{T21} & \cdots & h_{21} & \cdots & z_{T1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & g(t_1) & x_{t_1 n_1 1} & \cdots & h_{n_1 1} & \cdots & z_{t_1 1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & g(T) & x_{T n_1 1} & \cdots & h_{n_1 1} & \cdots & z_{T1} \\
1 & g(t_1) & x_{t_1 (n_1+1)2} & \cdots & h_{(n_1+1)2} & \cdots & z_{t_1 2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & g(T) & x_{T(n_1+1)2} & \cdots & h_{(n_1+1)2} & \vdots & z_{T2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & g(t_1) & x_{t_1 nK} & \cdots & h_{nK} & \cdots & z_{t_1 K} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & g(T) & x_{TnK} & \cdots & h_{nK} & \cdots & z_{TK}
\end{pmatrix}$$

$\Lambda_0 = (\beta_0, \beta_1, \ldots, \beta_{p-1})$,

X_1 is a $N \times n\dot{K}$ matrix. $\Lambda_1 = (r_{011}, r_{021}, \ldots, r_{0n_1 1}, r_{0(n_1+1)2}, r_{0(n_1+2)2}, \ldots, r_{0nK})'$,

$$X_1 = \begin{pmatrix}
\mathbf{1}_T & \mathbf{0}_T & \cdots & \cdots & \cdots & \cdots & \mathbf{0}_T \\
\mathbf{0}_T & \mathbf{1}_T & \mathbf{0}_T & \cdots & \cdots & \cdots & \mathbf{0}_T \\
\vdots & \mathbf{0}_T & \mathbf{1}_T & \mathbf{0}_T & \cdots & \cdots & \mathbf{0}_T \\
\vdots & \vdots & \mathbf{0}_T & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{0}_T & \cdots & \cdots & \cdots & \cdots & \mathbf{1}_T & \mathbf{0}_T \\
\mathbf{0}_T & \cdots & \cdots & \cdots & \cdots & \mathbf{0}_T & \mathbf{1}_T
\end{pmatrix}$$

$\mathbf{1}_T$ is the column vector $(T \times 1)$ of 1 while $\mathbf{0}_T$ is the column vector $(T \times 1)$ of 0.

$$\Lambda_2 = (r_{111}, r_{121}, \ldots, r_{1n_1 1}, r_{1(n_1+1)2}, r_{1(n_1+2)2}, \ldots, r_{1nK})',$$

$$X_2 = \begin{pmatrix} G_T & \mathbf{0}_T & \ldots & \ldots & \ldots & \ldots & \mathbf{0}_T \\ \mathbf{0}_T & G_T & \mathbf{0}_T & \ldots & \ldots & \ldots & \mathbf{0}_T \\ \vdots & \mathbf{0}_T & G_T & \mathbf{0}_T & \ldots & \ldots & \mathbf{0}_T \\ \vdots & \vdots & \mathbf{0}_T & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_T & \ldots & \ldots & \ldots & \ldots & G_T & \mathbf{0}_T \\ \mathbf{0}_T & \ldots & \ldots & \ldots & \ldots & \mathbf{0}_T & G_T \end{pmatrix},$$

where G_T is the $(T \times 1)$ vector $G_T = \begin{pmatrix} g(t_1) \\ g(t_2) \\ \vdots \\ g(T) \end{pmatrix}$

$$\Lambda_3 = (\phi_{11}, \phi_{21}, \ldots, \phi_{T1}, \phi_{12}, \phi_{22}, \ldots, \phi_{T2}, \ldots, \phi_{1K}, \phi_{2K}, \ldots, \phi_{TK})'.$$

$$X_3 = \begin{pmatrix}
I_T & \mathbf{0}_T & \dots & \dots & \dots & \dots & \mathbf{0}_T \\
I_T & \mathbf{0}_T & \dots & \dots & \dots & \dots & \mathbf{0}_T \\
\vdots & \mathbf{0}_T & \dots & \dots & \dots & \dots & \mathbf{0}_T \\
I_T & \mathbf{0}_T & \dots & \dots & \dots & \dots & \mathbf{0}_T \\
\mathbf{0}_T & I_T & \mathbf{0}_T & \dots & \dots & \dots & \mathbf{0}_T \\
\mathbf{0}_T & I_T & \mathbf{0}_T & \dots & \dots & \dots & \mathbf{0}_T \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{0}_T & I_T & \mathbf{0}_T & \dots & \dots & \dots & \mathbf{0}_T \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{0}_T & \dots & \dots & \dots & \dots & I_T & \mathbf{0}_T \\
\mathbf{0}_T & \dots & \dots & \dots & \dots & I_T & \mathbf{0}_T \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{0}_T & \dots & \dots & \dots & \dots & I_T & \mathbf{0}_T \\
\mathbf{0}_T & \dots & \dots & \dots & \dots & \mathbf{0}_T & I_T \\
\mathbf{0}_T & \dots & \dots & \dots & \dots & \mathbf{0}_T & I_T \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{0}_T & \dots & \dots & \dots & \dots & \mathbf{0}_T & I_T
\end{pmatrix},$$

$\Lambda_4 = (\omega_{11}, \omega_{21}, \dots, \omega_{T1}, \omega_{12}, \omega_{22}, \dots, \omega_{T2}, \dots, \omega_{1K}, \omega_{2K}, \dots, \omega_{TK})'$.

$X_4 = X_3$.

Λ_0, Λ_1, Λ_2, Λ_3, Λ_4 are assumed to be normally distributed and overall, the prior specification could be defined as:

$$f(\Lambda|\sigma) \propto \exp(-\frac{1}{2}\sum_{i=0}^{4} \Lambda_i' V_{\sigma_i^2} \Lambda_i) = \exp(-\frac{1}{2}\Lambda' V_{\sigma^2} \Lambda), \tag{A.3}$$

where V_{σ^2} is block diagonal with $V_{\sigma_i^2}$ for block i.

$V_{\sigma_i^2}$ is the covariance matrix for the parameter vector Λ_i.
Following Lindley and Smith (1972), the posterior full conditionals is given by:

$$\Lambda|Y \sim N((X'X + V_{\sigma^2})^{-1}X'Y, (X'X + V_{\sigma^2})^{-1}), \tag{A.4}$$

provided the inverse $(X'X + V_{\sigma^2})^{-1}$ exists.

A.5 Full conditionals for the parameters of interest, MLM CARs

The full conditionals for the cross-sectional analysis from equation (4.5) is a special case of Appendix A.4. We disregard the temporal components and, and the corresponding covariate matrices X_1 and X_2. We consider the **Besag-York-Mollié** model, i.e. $\psi = \phi + \omega$.

Now, we can write equation 4.5 in a more general and matrix form:

$E(Y) = \eta = X\Lambda = X_0\Lambda_0 + X_3\Lambda_3 + X_4\Lambda_4$, where $X = (X_0, X_3, X_4)$ is the design matrix. We suppose that there are n_i participants in area i, $i = 1, \ldots, K$. $\sum_{i=1}^{K} n_i = n$ is the number of participants.

$$Y = (y_{11}, y_{21}, \ldots, y_{n_1 1}, y_{(n_1+1)2}, \ldots, y_{(n)K})'$$

X_0 is a $n \times p$ matrix

$$X_0 = \begin{pmatrix} 1 & x_{11} & \ldots & h_{11} & \ldots & z_1 \\ 1 & x_{21} & \ldots & h_{21} & \ldots & z_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_1 1} & \ldots & h_{n_1 1} & \ldots & z_1 \\ 1 & x_{(n_1+1)2} & \ldots & h_{(n_1+1)2} & \ldots & z_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{nK} & \ldots & h_{nK} & \ldots & z_K \end{pmatrix}$$

$\Lambda_0 = (\beta_0, \beta_1, \ldots, \beta_{p-1})$,

$\Lambda_3 = (\phi_1, \phi_2, \ldots, \phi_K)'$.

$$X_3 = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ \vdots & 0 & \dots & \dots & \dots & \dots & 0 \\ 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & \dots & 1 & 0 \\ 0 & \dots & \dots & \dots & \dots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & \dots & 1 & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix},$$

with $\Lambda_4 = (\omega_1, \omega_2, \dots, \omega_K)'$.

and $X_4 = X_3$.

Λ_0, Λ_3, Λ_4 are assumed to be normally distributed and overall, the prior specification could be defined as:

$$f(\Lambda|\sigma) \propto \exp(-\frac{1}{2}\sum_{i=0}^{4} \Lambda_i' V_{\sigma_i^2} \Lambda_i) = \exp(-\frac{1}{2}\Lambda' V_{\sigma^2} \Lambda), \tag{A.5}$$

where V_{σ^2} is block diagonal with $V_{\sigma_i^2}$ for block i.

$V_{\sigma_i^2}$ is the covariance matrix for the parameter vector Λ_i.
Following Lindley and Smith (1972), the posterior full conditionals is given by:

$$\Lambda|Y \sim N((X'X + V_{\sigma^2})^{-1}X'Y, (X'X + V_{\sigma^2})^{-1}), \tag{A.6}$$

provided the inverse $(X'X + V_{\sigma^2})^{-1}$ exists.

Appendix B

Software implementations

Here, we display only essential parts of the R codes and SAS codes developed and used in this thesis.

B.1 Besag Newell

```
library(openxlsx)  # To read  xlsx-Data
library(rgdal)     # To readn Shapefile-Data (readOGR)
library(spdep)     # To produce a binary adjacent matrix
library(ape)       # Moran’s I
library(boot)      # Besag and Newell with Opgam
library(DCluster)  # Besag und Newell
library(dplyr)     # Graphics
library(ggmap)
library(tmap)
library(INLA)      # to fit the convolution model for the spatial model



# Besag und Newell: year 0

sids <- data.frame(Observed = dataM1$Depr_sum)
sids <- cbind(sids, Expected = dataM1$Erwart_bas)
sids <- cbind(sids, x = dataM1$X_Coord.x, y = dataM1$Y_Coord.x)

bnresults7 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                    iscluster = bn.iscluster, set.idxorder = TRUE,
                    k = 7, model = "poisson",
                    R = 100, mle = calculate.mle(sids))
bnresults16 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                     iscluster = bn.iscluster, set.idxorder = TRUE,
                     k = 16, model = "poisson",
                     R = 100, mle = calculate.mle(sids))
bnresults20 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                     iscluster = bn.iscluster, set.idxorder = TRUE,
                     k = 20, model = "poisson",
                     R = 100, mle = calculate.mle(sids))
bnresults25 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                     iscluster = bn.iscluster, set.idxorder = TRUE,
                     k = 25, model = "poisson",
                     R = 100, mle = calculate.mle(sids))

# Significant  districts:

sig <- function(bn){
  si <- c()
  for(i in 1: length(bn$x)){
    si[i] <- which(dataM1$X_Coord.x == bn$x[i] & dataM1$Y_Coord.x == bn$y[i])
  }
  return(si)
}

l7 <- dataM1[sig(bnresults7), "Stteil"]
l16 <- dataM1[sig(bnresults16), "Stteil"]
l20 <- dataM1[sig(bnresults20), "Stteil"]
l25 <- dataM1[sig(bnresults25), "Stteil"]

ST <- dataM1$Stadt.y
ST[which(ST == "Essen")] <- 1
ST[which(ST == "Bochum")] <- 2
ST[which(ST == "Muelheim")] <- 0

par(mfrow = c(3, 1), mar = .1 + c(0.1, 0.3, 0.1, 0.3), mai= c(0.1, 0, 0.2, 0.1))
plot(dataM1$X_Coord.x, dataM1$Y_Coord.x, pch = as.numeric(ST),
     xaxt="n", yaxt = "n", bty = "n", xlab = "", ylab = "",
     col = "darkgrey", cex.main = 1,
     ylim = c(min(dataM1$Y_Coord.x), max(dataM1$Y_Coord.x) + 1000) , main= "Baseline")
points(x= bnresults7$x, y = bnresults7$y, col = "darkgreen",
       pch = 1, cex = 4, lwd = 2)
points(x= bnresults16$x, y = bnresults16$y, col = "blue",
       pch = 1, cex = 5, lwd = 2)
points(x= bnresults20$x, y = bnresults20$y, col = "green",
       pch = 1, cex = 6, lwd = 2)
```

```
points(x= bnresults25$x, y = bnresults25$y, col = "red",
       pch = 1, cex = 7, lwd = 2)
legend("bottomright",
       legend = c("k=7","k=16", "k=20", "k=25", "Mülheim", "Essen", "Bochum"),
       pch = c(19, 19, 19, 19, 0, 1, 2),
       col = c("darkgreen","blue", "green", "red", "darkgrey", "darkgrey",
               "darkgrey"), cex = 1.4)
text(bnresults7$x, bnresults7$y, labels = l7,  pos = c(2, 2, 4), cex = 1.3)
text(bnresults16$x, bnresults16$y, labels = l16,  pos = 4, cex = 1.3)
text(bnresults20$x, bnresults20$y, labels = l20,
     pos = 4, cex = 1.3)
text(bnresults25$x[2], bnresults25$y[2], labels = " ",
     pos = 3, cex = 1.3)

box()


# Besag and Newell: year5

sids <- data.frame(Observed = dataM5$Depr_sum)
sids <- cbind(sids, Expected = dataM5$Erwart_bas)
sids <- cbind(sids, x = dataM5$X_Coord.x, y = dataM5$Y_Coord.x)

bnresults7 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                    iscluster = bn.iscluster, set.idxorder = TRUE,
                    k = 7, model = "poisson",
                    R = 100, mle = calculate.mle(sids))
bnresults16 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                     iscluster = bn.iscluster, set.idxorder = TRUE,
                     k = 16, model = "poisson",
                     R = 100, mle = calculate.mle(sids))
bnresults20 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                     iscluster = bn.iscluster, set.idxorder = TRUE,
                     k = 20, model = "poisson",
                     R = 100, mle = calculate.mle(sids))
bnresults25 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                     iscluster = bn.iscluster, set.idxorder = TRUE,
                     k = 25, model = "poisson",
                     R = 100, mle = calculate.mle(sids))

# Significant districts:
```

```
sig <- function(bn){
  si <- c()
  for(i in 1: length(bn$x)){
    si[i] <- which(dataM5$X_Coord.x == bn$x[i] & dataM5$Y_Coord.x == bn$y[i])
  }
  return(si)
}

l7 <- dataM5[sig(bnresults7), "Stteil"]
l16 <- dataM5[sig(bnresults16), "Stteil"]
l20 <- dataM5[sig(bnresults20), "Stteil"]
l25 <- dataM5[sig(bnresults25), "Stteil"]

ST <- dataM1$Stadt.y
ST[which(ST == "Essen")] <- 1
ST[which(ST == "Bochum")] <- 2
ST[which(ST == "Muelheim")] <- 0

#par(mfrow = c(1, 1), mar = .1 + c(0, 0.1, 0, 0.3))
plot(dataM5$X_Coord.x, dataM5$Y_Coord.x, pch = as.numeric(ST),
     xaxt="n", yaxt = "n", bty = "n", xlab = "", ylab = "",
     col = "darkgrey", cex.main = 1,
     ylim = c(min(dataM5$Y_Coord.x), max(dataM5$Y_Coord.x) + 1000),
      main= "First follow-up")
points(x= bnresults7$x, y = bnresults7$y, col = "darkgreen",
       pch = 1, cex = 4, lwd = 2)
points(x= bnresults16$x, y = bnresults16$y, col = "blue",
       pch = 1, cex = 5, lwd = 2)
points(x= bnresults20$x, y = bnresults20$y, col = "green",
       pch = 1, cex = 6, lwd = 2)
points(x= bnresults25$x, y = bnresults25$y, col = "red",
       pch = 1, cex = 7, lwd = 2)
legend("bottomright",
       legend = c("k=7","k=16", "k=20", "k=25", "Mülheim",
       "Essen", "Bochum"),
       pch = c(19, 19, 19, 19, 0, 1, 2),
       col = c("darkgreen","blue", "green", "red", "darkgrey", "darkgrey",
               "darkgrey"), cex = 1.4)
text(bnresults7$x, bnresults7$y, labels = l7,  pos = c(2, 2, 4), cex = 1.3)
text(bnresults16$x, bnresults16$y, labels = l16,  pos = 4, cex = 1.3)
text(bnresults20$x, bnresults20$y, labels = l20,
     pos = 4, cex = 1.3)
```

```
text(bnresults25$x[2], bnresults25$y[2], labels = "Hordel",
     pos = 3, cex = 1.3)

box()

# Besag and Newell: year 10

sids <- data.frame(Observed = dataM10$Depr_sum)
sids <- cbind(sids, Expected = dataM10$Erwart_bas)
sids <- cbind(sids, x = dataM10$X_Coord.x, y = dataM10$Y_Coord.x)

bnresults7 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                    iscluster = bn.iscluster, set.idxorder = TRUE,
                    k = 7, model = "poisson",
                    R = 100, mle = calculate.mle(sids))
bnresults16 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                     iscluster = bn.iscluster, set.idxorder = TRUE,
                     k = 16, model = "poisson",
                     R = 100, mle = calculate.mle(sids))
bnresults20 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                     iscluster = bn.iscluster, set.idxorder = TRUE,
                     k = 20, model = "poisson",
                     R = 100, mle = calculate.mle(sids))
bnresults25 <- opgam(sids, thegrid = sids[,c("x","y")], alpha = .05,
                     iscluster = bn.iscluster, set.idxorder = TRUE,
                     k = 25, model = "poisson",
                     R = 100, mle = calculate.mle(sids))

# Significant districts:

sig <- function(bn){
  si <- c()
for(i in 1: length(bn$x)){
  si[i] <- which(dataM10$X_Coord.x == bn$x[i] & dataM10$Y_Coord.x == bn$y[i])
}
  return(si)
}

l7 <- dataM10[sig(bnresults7), "Stteil"]
l16 <- dataM10[sig(bnresults16), "Stteil"]
l20 <- dataM10[sig(bnresults20), "Stteil"]
l25 <- dataM10[sig(bnresults25), "Stteil"]
```

```
ST <- dataM1$Stadt.y
ST[which(ST == "Essen")] <- 1
ST[which(ST == "Bochum")] <- 2
ST[which(ST == "Muelheim")] <- 0

#par(mfrow = c(1, 1), mar = .1 + c(0, 0.1, 0, 0.3))
plot(dataM10$X_Coord.x, dataM10$Y_Coord.x, pch = as.numeric(ST),
     xaxt="n", yaxt = "n", bty = "n", xlab = "", ylab = "",
     col = "darkgrey", cex.main = 1,
     ylim = c(min(dataM10$Y_Coord.x), max(dataM10$Y_Coord.x) + 1000),
      main= "Second follow-up")
points(x= bnresults7$x, y = bnresults7$y, col = "darkgreen",
       pch = 1, cex = 4, lwd = 2)
points(x= bnresults16$x, y = bnresults16$y, col = "blue",
       pch = 1, cex = 5, lwd = 2)
points(x= bnresults20$x, y = bnresults20$y, col = "green",
       pch = 1, cex = 6, lwd = 2)
points(x= bnresults25$x, y = bnresults25$y, col = "red",
       pch = 1, cex = 7, lwd = 2)
legend("bottomright",
       legend = c("k=7","k=16", "k=20", "k=25", "Mülheim",
        "Essen", "Bochum"),
       pch = c(19, 19, 19, 19, 0, 1, 2),
       col = c("darkgreen","blue", "green", "red", "darkgrey",
       "darkgrey",
               "darkgrey"), cex = 1.4)
text(bnresults7$x, bnresults7$y, labels = l7,
pos = c(2, 2, 4), cex = 1.3)
text(bnresults16$x, bnresults16$y, labels = l16,
 pos = 4, cex = 1.3)
text(bnresults20$x, bnresults20$y, labels = l20,
     pos = 4, cex = 1.3)
text(bnresults25$x[2], bnresults25$y[2], labels = " ",
     pos = 3, cex = 1.3)

box()
```

B.2 Spatial model for disease mapping

```
# Convolution model:
```

```
# does an  extra-Poisson-Variation exist?

mean(dataM1$Depr_sum)
var(dataM1$Depr_sum)
mean(dataM5$Depr_sum)
var(dataM5$Depr_sum)

mean(dataM10$Depr_sum)
var(dataM10$Depr_sum)

# Read the adjacent matrix:

source("daten_adjac.R")
set.seed(1066)


formula21 <- Depr_sum ~ (f(ID, model = "bym", graph = ruhr,
                          param = c(0.5, 0.0005)) +
                          dataM1$green_s +
                           dataM1$mean_max_lden+dataM1$mean_unemplBL)

mod2.ruhr1 <- inla(formula21, family = "poisson", E = dataM1$Erwart_bas,
                  data = dataM1, control.compute = list(config = TRUE))

n <- 100000 # sample size

Temp21 <- inla.posterior.sample(n = n, result = mod2.ruhr1,
                               use.improved.mean = TRUE)
Temp21 <- lapply(Temp21, function(x){x$latent[1 : nrow(dataM1)]})
Temp21 <- matrix(unlist(Temp21), byrow = TRUE, nrow = n, ncol = nrow(dataM1))
Temp21 <- exp(Temp21)
RRa21 <- Temp21
RR21 <- apply(Temp21, MARGIN = 2, FUN = mean) # Mittelwerte als Schätzer

#save(file="INLA_results_RUHR21.Rdata", list = ls(all = TRUE))




# BYM with covariables year 5:

set.seed(1066)
```

```
formula25 <- Depr_sum ~ (f(ID, model = "bym", graph = ruhr,
                           param = c(0.5, 0.0005)) +
                           dataM5$green_s + dataM5$mean_max_lden+dataM5$mean_unemplBL)

mod2.ruhr5 <- inla(formula25, family = "poisson", E = dataM5$Erwart_bas,
                   data = dataM5, control.compute = list(config = TRUE))

n <- 100000 # Sample size

Temp25 <- inla.posterior.sample(n = n, result = mod2.ruhr5,
                                use.improved.mean = TRUE)
Temp25 <- lapply(Temp25, function(x){x$latent[1 : nrow(dataM5)]})
Temp25 <- matrix(unlist(Temp25), byrow = TRUE, nrow = n, ncol = nrow(dataM5))
Temp25 <- exp(Temp25)
RRa25 <- Temp25
RR25 <- apply(Temp25, MARGIN = 2, FUN = mean) # Mittelwerte als Schätzer

#save(file="INLA_results_RUHR25.Rdata", list = ls(all = TRUE))


# BYM with covariables year 10:

set.seed(1066)
formula210 <- Depr_sum ~ (f(ID, model = "bym", graph = ruhr,
                           param = c(0.5, 0.0005)) +
                           dataM10$green_s + dataM10$mean_max_lden+dataM10$mean_unemplBL)

mod2.ruhr10 <- inla(formula210, family = "poisson", E = dataM10$Erwart_bas,
                   data = dataM10, control.compute = list(config = TRUE))

n <- 100000 # Sample size

Temp210 <- inla.posterior.sample(n = n, result = mod2.ruhr10,
                                use.improved.mean = TRUE)
Temp210 <- lapply(Temp210, function(x){x$latent[1 : nrow(dataM10)]})
Temp210 <- matrix(unlist(Temp210), byrow = TRUE, nrow = n, ncol = nrow(dataM10))
Temp210 <- exp(Temp210)
RRa210 <- Temp210
RR210 <- apply(Temp210, MARGIN = 2, FUN = mean) # mean value as estimate

#save(file="INLA_results_RUHR210.Rdata", list = ls(all = TRUE))
```

```
# display of the estimates from the convolution model with covariates:
data1$RRfalt21 <- RR21
lnd_mydata1@data <- left_join(lnd_mydata1@data, data1, by = "Stteil")

tm41 <- qtm(lnd_mydata1, fill = "RRfalt21", borders = "black",
           fill.style = "equal", title = "Risik estimate (BYM)
           with covariates",
           title.cex = 1.3, bg.color = "white", fill.palette = "-RdYlGn",
           fill.labels = c("0.79 - 0.90", "0.90 - 1.01", "1.01 - 1.12",
                           "1.12 - 1.23", "1.23 - 1.34")) +
  tm_layout(outer.margins=c(0, 0, 0, 0),
            inner.margins=c(0.15, 0.01, 0.01, 0.01), asp = NA,
            bg.color = "white", frame = TRUE,
            legend.position = c("right", "bottom"),
            title.position =  c("right", "bottom"), legend.text.size = 1.1)
tm41

# Standard deviation:
sd(RR21)
# display of the estimates from convolution model with covariates year 5:

data5$RRfalt25 <- RR25
lnd_mydata5@data <- left_join(lnd_mydata5@data, data5, by = "Stteil")

tm45 <- qtm(lnd_mydata5, fill = "RRfalt25", borders = "black",
            fill.style = "equal", title = "Risk estimate (BYM)
            with covariates",
            title.cex = 1.3, bg.color = "white", fill.palette = "-RdYlGn",
            fill.labels = c("0.79 - 0.90", "0.90 - 1.01", "1.01 - 1.12",
                            "1.12 - 1.23", "1.23 - 1.34")) +
  tm_layout(outer.margins=c(0, 0, 0, 0),
            inner.margins=c(0.15, 0.01, 0.01, 0.01), asp = NA,
            bg.color = "white", frame = TRUE,
            legend.position = c("right", "bottom"),
            title.position =  c("right", "bottom"), legend.text.size = 1.1)
tm45

# Standard deviation:
```

```
sd(RR25)

# display of the estimates from convolution model with covariates year 10:

data10$RRfalt210 <- RR210
lnd_mydata10@data <- left_join(lnd_mydata10@data, data10, by = "Stteil")

tm410 <- qtm(lnd_mydata10, fill = "RRfalt210", borders = "black",
            fill.style = "equal", title = "Risk estimate (BYM)
           with covariables",
            title.cex = 1.3, bg.color = "white", fill.palette = "-RdYlGn",
            fill.labels = c("0.79 - 0.90", "0.90 - 1.01", "1.01 - 1.12",
                            "1.12 - 1.23", "1.23 - 1.34")) +
  tm_layout(outer.margins=c(0, 0, 0, 0),
            inner.margins=c(0.15, 0.01, 0.01, 0.01), asp = NA,
            bg.color = "white", frame = TRUE,
            legend.position = c("right", "bottom"),
            title.position =  c("right", "bottom"), legend.text.size = 1.1)
tm410

# Standard deviation:
sd(RR210)


## Merge data for mapping
merge(emp,dept,by="DEPTNO")[,c("ENAME","DNAME")]


datarr1$SMR1 <- datarr1$SMR
datarr5$SMR5 <- datarr5$SMR
datarr10$SMR10 <- datarr10$SMR
datarr1 <- datarr1[, c("Stteil", "SMR1", "SMRgew1")]
datarr5 <- datarr5[, c("Stteil",  "SMR5","SMRgew5")]
datarr10 <- datarr10[, c("Stteil",  "SMR10", "SMRgew10")]
data1 = data1[, c("Stteil", "RRfalt1",  "RRfalt21")]
data5 = data5[, c("Stteil", "RRfalt5" ,  "RRfalt25")]
data10 = data10[, c("Stteil", "RRfalt10",  "RRfalt210")]

library(plyr)
Risiko_Mapping = join_all(list(datarr1,datarr5,datarr10, data1, data5, data10),
 by='Stteil', type='left')
```

```
# convert to excell

library(xlsx)
Risiko_Mapping = write.xlsx(Risiko_Mapping, "Risiko_Mapping.xlsx")
#Extreme estimates year 0:
dataf21 <- lnd_mydata1@data
attach(dataf21)
1 <- cbind(dataf21[, c("Stteilnu.x", "RRfalt21", "Ris_pop",
                           "mean_age", "NDVIMean06")])
dataf21 <- dataf21[order(-dataf21$RRfalt21), ]
dataf21

# Extreme estimates year 5:
dataf25 <- lnd_mydata5@data
1 <- cbind(dataf25[, c("Stadt.x", "Stteil", "RRfalt25", "Ris_pop",
                       "mean_age", "NDVIMean06")])
dataf25 <- dataf25[order(-dataf25$RRfalt25), ]
dataf25
# Extreme estimates year 10:
dataf210 <- lnd_mydata10@data
1 <- cbind(dataf210[, c("Stadt.x", "Stteil", "RRfalt210", "Ris_pop",
                       "mean_age", "NDVIMean09")])
dataf210 <- dataf210[order(-dataf210$RRfalt210), ]
dataf210
```

B.3 Spatio-temporal Poisson model

```
rm(list=ls())
library("CARBayesdata")
library("sp")
library(rgdal)
lnd_mydata = readOGR(dsn = "StudiengebietStadtteileGauss",
layer = "StudiengebietStadtteileGauss",  use_iconv = TRUE, encoding = "UTF-8")
lnd_mydata@data <- lnd_mydata@data[order(lnd_mydata@data$Stteil),]
# The data set
#lnd_mydata <- spTransform(lnd_mydata, CRS("+proj=longlat +datum=WGS84 +no_defs"))
library(xlsx)
library(ggmap)
library(tmap)
```

```
dep_score = read.xlsx("combine_SMR_NDVI_jahr_neu_dep.xlsx",
sheetName="combine_SMR_NDVI_jahr_neu_dep", use_iconv = TRUE, encoding = "UTF-8")
#data <- read.xlsx("combine_SMR_NDVI.xlsx", sheetName="combine_SMR_NDVI",
 use_iconv = TRUE, encoding =" UTF-8")
#dataM <- data
```

```
 library("spdep")
 W.nb <- poly2nb(lnd_mydata, row.names = SMR.av$Stteil)
 W.list <- nb2listw(W.nb, style = "B")
 W <- nb2mat(W.nb, style = "B")

 postscript(file=" A.eps", onefile=FALSE, horizontal=FALSE) ## produce eps file

 #Assessing the presence of spatial autocorrelation

 # first computing the residuals from a simple
 # overdispersed Poisson log-linear model that incorporates the covariate effects



 formula1 <- Depr_sum_i ~ offset(log(Erwart_bas)) + green_s
 formula1b <- Depr_sum_i ~ green_s

 formula2 <- Depr_sum_i ~ offset(log(Erwart_bas)) + green_s  +mean_age
 formula2b <- Depr_sum_i ~ green_s  +mean_age

 formula3 <- Depr_sum_i ~ offset(log(Erwart_bas)) + green_s  +mean_age +
  mean_oecdnet_s
 formula3b <- Depr_sum_i ~ green_s  +mean_age + mean_oecdnet_s

 formula4 <- Depr_sum_i ~ offset(log(Erwart_bas)) + green_s  +mean_age +
  mean_oecdnet_s + umgezogen_perc
 formula4b <- Depr_sum_i ~ green_s  +mean_age + mean_oecdnet_s + umgezogen_perc
 formula5 <- Depr_sum_i ~ offset(log(Erwart_bas)) + green_s  +mean_age +
  mean_oecdnet_s +
 umgezogen_perc + sum_comorbi
 formula5b <- Depr_sum_i ~ green_s  +mean_age + mean_oecdnet_s + umgezogen_perc +
 sum_comorbi



 formula6 <- Depr_sum_i ~ offset(log(Erwart_bas)) + green_s  +mean_age + mean_oecdnet_s
 + umgezogen_perc + sum_comorbi+ mean_bmi
 formula6b <- Depr_sum_i ~ green_s  +mean_age + mean_oecdnet_s + umgezogen_perc +
  sum_comorbi + mean_bmi



 formula7 <- Depr_sum_i ~ offset(log(Erwart_bas)) + green_s  +mean_age + mean_oecdnet_s
 + umgezogen_perc + sum_comorbi+ mean_bmi + median_max_lden
 formula7b <- Depr_sum_i ~ green_s  +mean_age + mean_oecdnet_s + umgezogen_perc +
  sum_comorbi +mean_bmi + median_max_lden




 formula8 <- Depr_sum_i ~ offset(log(Erwart_bas)) + green_s  +mean_age + mean_oecdnet_s
 + umgezogen_perc + sum_comorbi+ mean_bmi + median_max_lden + mean_unemplBL
 formula8b <- Depr_sum_i ~ green_s  +mean_age + mean_oecdnet_s + umgezogen_perc +
 sum_comorbi +mean_bmi + median_max_lden + mean_unemplBL




 formula9 <- Depr_sum_i ~ offset(log(Erwart_bas)) + green_s   + umgezogen_perc +
 sum_comorbi + median_max_lden
 formula9b <- Depr_sum_i ~ green_s    + umgezogen_perc + sum_comorbi + median_max_lden

 formula10 <- Depr_sum_i ~   offset(log(Erwart_bas)) + green_s  +mean_age + mean_oecdnet_s
 + mean_unemplBL
 formula10b <- Depr_sum_i ~   green_s  +mean_age + mean_oecdnet_s + mean_unemplBL



model1 <- glm(formula = formula9, family = "quasipoisson",
                  data = M_12)
resid.glm <- residuals(model1)
summary(model1)#$coefficients




#Overdispersion?
summary(model1)$dispersion
#To quantify the presence of spatial autocorrelation in
#the residuals from this model we can compute Moran’s I statistic
#(Moran 1950) and conduct
#a permutation test for each year of data separately.
# The permutation test has the null hypothesis
#of no spatial autocorrelation and an alternative hypothesis of spatial
# autocorrelation
#(either positive or negative), and is conducted using the moran.mc()
# function from the spdep
#package.

moran.mc(x = resid.glm, listw = W.list, nsim = 10000)

## Spatio-temporal modelling with CARBayesST

# all data vectors (response, offset and covariates) have to be ordered
#so that the first K
# data points relate to all spatial units at time 1, the next K data points
#to all spatial units at
# time 2 and so on

library("CARBayesST")
model2 <- ST.CARar(formula = formula3, family = "poisson",
                   data = dep_scoreM, W = W,
                   burnin = 20000, n.sample = 22000,
                   thin = 10)




dep_scoreM$depprop <- dep_scoreM$Depr_sum / dep_scoreM$Ris_pop
boxplot(dep_scoreM$depprop ~ dep_scoreM$jahr, range = 0, xlab = "Year",
            ylab = "dep rate",
            col = "darkseagreen", border = "navy")
```

```
library("dplyr")
depprop.av <- summarise(group_by(dep_scoreM, Stteil),
                              depprop.mean = mean(depprop))
lnd_mydata@data$dep <- depprop.av$depprop.mean


rate.est <- matrix(model2$fitted.values / dep_scoreM$Erwart_bas,
                    nrow = nrow(W), byrow = FALSE)



rate.est <- as.data.frame(rate.est)
 colnames(rate.est) <- c("Risk_year0", "Risk_year5", "Risk_year7", "Risk_year8",
 "Risk_year9", "Risk_year10", "Risk_year11", "Risk_year12")


 rate.obs <- matrix(model2$fitted.values ,
                    nrow = nrow(W), byrow = FALSE)



 rate.obs <- as.data.frame(rate.obs)
 colnames(rate.obs) <- c("obs_year0", "obs_year5", "obs_year7", "obs_year8",
               "obs_year9",  "obs_year10", "obs_year11", "obs_year12")



 lnd_mydata@data <- data.frame(lnd_mydata@data, rate.est, rate.obs)
  breakpoints <- c(0, quantile(depprop.av$depprop.mean, seq(0.1, 0.9, 0.1)),
                      0.1)



  library(xlsx)
  write.xlsx(lnd_mydata@data, "lnd_mydata_risk.xlsx")


   spplot(lnd_mydata, c("Risk_year0", "Risk_year5", "Risk_year7", "Risk_year8",
    "Risk_year9", "Risk_year10", "Risk_year11", "Risk_year12"),
            names.attr = c("Risk_year0", "Risk_year5", "Risk_year7", "Risk_year8",
           "Risk_year9", "Risk_year10", "Risk_year11", "Risk_year12"),
         sp.layout = list(l1, l2, l3, l4),
           xlab = "Easting", ylab = "Northing", scales = list(draw = T),
           at = breakpoints, col.regions = terrain.colors(n = length(breakpoints - 1)),
           par.settings=list(fontsize=list(text=20)))
```

```
# Kartierung der SMR:

tm11 <- qtm(lnd_mydata1, fill = "SMR", borders = "black", fill.style =
              "fixed", fill.breaks = c(0, 0.1, 0.9, 1.2 ,3.61),
            title = "SMR", title.cex = 1.5, fill.palette = "-RdYlGn",
            fill.labels = c("0", "0.1 - 0.9", "0.9 - 1.2", "1.2 - 3.6")) +
  tm_layout(inner.margins = c(0.1, 0.01, 0.01, 0.01),
            outer.margins = 0, asp = NA,
            bg.color = "white", frame = TRUE,
            legend.position = c("right", "bottom"),
            title.position =  c("right", "bottom"),
            legend.text.size = 1.4)
tm11
```

```
model2b <- ST.CARar(formula = formula3b, family = "binomial",
                   data = dep_scoreM, W = W, trial = dep_scoreM$Ris_pop,
                     prior.tau2=NULL,
                   burnin = 20000, n.sample = 22000,
                   thin = 10)
```

```
model2_0 <- ST.CARar(formula = formula3, family = "poisson",
                   data = dep_scoreM, W = W, fix.rho.S=TRUE, rho.S=0,
                   burnin = 20000, n.sample = 22000,
                   thin = 10)
```

```
 model2b_0 <- ST.CARar(formula = formula3b, family = "binomial",
                     data = dep_scoreM, W = W,  fix.rho.S=TRUE,
                     trial = dep_scoreM$Ris_pop, rho.S=0,   prior.tau2=NULL,
                     burnin = 20000, n.sample = 22000,
                     thin = 10)




 model2_1 <- ST.CARar(formula = formula3, family = "poisson",
                      data = dep_scoreM, W = W, fix.rho.S=TRUE, rho.S=1,
                      burnin = 20000, n.sample = 22000,
                      thin = 10)




 model2_1 <- ST.CARar(formula = formula3, family = "poisson",
                      data = dep_scoreM, W = W, fix.rho.S=TRUE, rho.S=1,  prior.tau2=NULL,
                      burnin = 2000, n.sample = 5000,
                      thin = 10)


 model2_1_0tau <- ST.CARar(formula = formula3, family = "poisson",
                      data = dep_scoreM, W = W, rho.S=1, prior.tau2=c(1,1),
                      burnin = 2000, n.sample = 5000,
                      thin = 10)


 # return output to the terminal
sink("model3_13.txt")  #writ all output to file tp.txt

 print(model2)
 print(model2b)

 print(model2_0)
 print(model2b_0)

 print(model2_1)
 print(model2b_1)
```

```
## Temporal analysis
model2_1_0tau <- ST.CARar(formula = formula3, family = "poisson",
                          data = dep_scoreM, W = W,
                          burnin = 40000, n.sample = 70000,
                          thin = 10)

#print(model10)


 library("CARBayes")
 parameter.summary <- summarise.samples(exp(model2$samples$beta),
                                        quantiles = c(0.5, 0.025, 0.975))
 round(parameter.summary$quantiles, 4)

 parameter.summary <- summarise.samples(exp(model2b$samples$beta),
                                        quantiles = c(0.5, 0.025, 0.975))
 round(parameter.summary$quantiles, 4)

 parameter.summary <- summarise.samples(exp(model2_0$samples$beta),
                                        quantiles = c(0.5, 0.025, 0.975))
 round(parameter.summary$quantiles, 4)

 parameter.summary <- summarise.samples(exp(model2b_0$samples$beta),
                                        quantiles = c(0.5, 0.025, 0.975))
 round(parameter.summary$quantiles, 4)

 parameter.summary <- summarise.samples(exp(model2b_1$samples$beta),
                                        quantiles = c(0.5, 0.025, 0.975))
 round(parameter.summary$quantiles, 4)

 parameter.summary <- summarise.samples(exp(model2b_1$samples$beta),
                                        quantiles = c(0.5, 0.025, 0.975))
 round(parameter.summary$quantiles, 4)

 sink()
```

B.4 Multilevel conditional autoregressive models, MLM tCARs

We focus essentially on the developed MLM tCARs.

```
## Prepare the data sets as well as initial values for the Bayesian setting

rm(list=ls())
library(R2WinBUGS) # For Bayesian analysis
library(MASS) #  normal function in vectorized form
library(rgdal)
library(spdep)
library(batchtools)
library(coda)
library(tidyr)
library(simstudy)

trad_multl_spatialg_model <- function(){
  for (i in 1:N){


    #Data Model
    yij[i]~dnorm(mu[i],tau.sf12)

    #Process Model
    mu[i]<-alpha[id_num[i]]+Z[i]+X[i]+D[i]

    #Individual Effect
    X[i]<-beta_1*green_c[i]+beta_2*age_period_c[i]

    #Temporal Effect
    D[i]<-delta*period[i]+period[i]*u1[id_num[i]]
    #Spatial Effect
    Z[i]<-gamma*unempl_c[i] #+omega[Stt_num[i],period[i]]
  }

  # Individual Unobserved Effect Prior
  for (j in 1:NI){
    alpha[j]~dnorm(omega[Stt_num[j],period[j]],tau.alpha) #dnorm(theta, tau.alpha)#
    u1[j]~dnorm(0, zalpha.lin)
  }

  #Spatial Unobserved Effect Prior
  for (t in 1:TT){

    for (r in 1:K){
      omega[r,t]~dnorm(theta,tau.omega)

    }


  }

  # tau.omega<-1/(sigma.omega*sigma.omega)
  # sigma.omega~dunif(0,100)

  tau.omega~dgamma(2,100)

  beta_1~dnorm(0,0.00001)
  beta_2~dnorm(0,0.00001)
  gamma~dnorm(0,0.00001)
  delta~dnorm(0,0.00001)


  theta~dflat()

  # tau.sf12<-1/(sigma.sf12*sigma.sf12)
  #  sigma.sf12~dunif(0,100)

  tau.sf12~dgamma(1,0.01)

  #tau.alpha<-1/(sigma.alpha*sigma.alpha)
  #sigma.alpha~dunif(0,100)

  tau.alpha~dgamma(1,0.01)
  zalpha.lin~dgamma(1,0.01)


}
```

```
# Modell als R-Funktion (aber WinBUGS Syntax)
Car_anova.modelg <- function(){
  for (i in 1:N){


    #Data Model
    yij[i]~dnorm(mu[i],tau.sf12)

    #Process Model
    mu[i]<-alpha[id_num[i]]+Z[i]+X[i]+D[i]

    #Individual Effect
    X[i]<-beta_1*green_c[i]+beta_2*age_period_c[i]

    #Temporal Effect
    D[i]<-delta*period[i]+period[i]*u1[id_num[i]]
    #Spatial Effect
    #Spatial Effect
    Z[i]<-gamma*unempl_c[i] #+phi[period[i],Stt_num[i]]
  }

  # Individual Unobserved Effect Prior
  for (j in 1:NI){
    alpha[j]~dnorm(phi[period[j],Stt_num[j]], tau.alpha)
    #dnorm(omega[Stt_num[j],jahr_num[j]],tau.alpha)
    #dnorm(theta, tau.alpha)
    u1[j]~dnorm(0, zalpha.lin)
  }

  ## define delta_lin
  for(t in 1:TT){
    delta.lin[t]~dnorm(Sd.lin[t],taud.lin[t])

    Sd.lin[t] <- (rhod.lin/(1-rhod.lin+rhod.lin*num_t[t]))*sum(Ds.delta[C_t[t]+1:C_t[t+1]])

    taud.lin[t] <- inv.deltad*(1-rhod.lin+rhod.lin*num_t[t])
  }

  # sum weighted errors over neighbors
  for (i in 1:sumNumNeigh_t) { Ds.delta[i] <- delta.lin[adj_t[i]] }
```

```
## define lin_phi
for(k in 1:K){
  lin.phi[k]~dnorm(S.lin[k],tau.lin[k])

  S.lin[k] <- (rho.lin/(1-rho.lin+rho.lin*num[k]))*sum(Ws.phi[C[k]+1:C[k+1]])

  tau.lin[k] <- inv.delta*(1-rho.lin+rho.lin*num[k])

}

# sum weighted errors over neighbors
for (i in 1:sumNumNeigh) { Ws.phi[i] <- lin.phi[adj[i]] }

for(t in 1:TT){ for (k in 1:K){
  phi[t,k] <-   delta.lin[t] + lin.phi[k]+ gamma.lin[k,t]
}}

for (t in 1:TT){
  for (r in 1:K){
    gamma.lin[r,t]~dnorm(theta,tau.gamma)

  }
}

#Priors
#Priors

rho.lin~dunif(0,1)
rhod.lin~dunif(0,1)

beta_1~dnorm(0,0.00001)
beta_2~dnorm(0,0.00001)

gamma~dnorm(0,0.00001)


delta~dnorm(0,0.00001)

inv.deltad~dgamma(2,100)

inv.delta~dgamma(2,100)
```

```
  tau.gamma~dgamma(1,0.01)

  tau.sf12~dgamma(1,0.01)

  theta~dflat()
  tau.alpha~dgamma(1,0.01)
  zalpha.lin~dgamma(1,0.01)
}

# Modell als R-Funktion (aber WinBUGS Syntax)
convolution.modelg <- function(){
  for (i in 1:N){


    #Data Model
    yij[i]~dnorm(mu[i],tau.sf12)

    #Process Model
    mu[i]<- alpha[id_num[i]]+Z[i]+X[i]+D[i]

    #Individual Effect
    X[i]<- beta_1*green_c[i]+beta_2*age_period_c[i]

    #Temporal Effect
    D[i]<-delta*period[i]+period[i]*u1[id_num[i]]

    #Spatial Effect
    Z[i]<-gamma*unempl_c[i]
    # +phi[period[i],Stt_num[i]]+omega[Stt_num[i],period[i]]
  }

  # Individual Unobserved Effect Prior
  for (j in 1:NI){
    psi[j]<-phi[period[j],Stt_num[j]]+omega[Stt_num[j],period[j]]
    alpha[j]~dnorm (psi[j], tau.alpha)#(theta,tau.alpha)
    u1[j]~dnorm(0, zalpha.lin)
  }
  for(i in 1:sumNumNeigh) {
    weights[i] <- 1
  }

  #Spatial Unobserved Effect Prior
  for (t in 1:TT){
    phi[t,1:K]~car.normal(adj[], weights[],  num[], tau.phi[t])
    tau.phi[t]~dgamma(2,100)


    for (r in 1:K){
      omega[r,t]~dnorm(theta,tau.omega[t])

    }

    tau.omega[t]~dgamma(2,100)

  }

  #Priors
 beta_1~dnorm(0,0.00001)
  beta_2~dnorm(0,0.00001)
  gamma~dnorm(0,0.00001)
  delta~dnorm(0,0.00001)
 theta~dflat()
tau.sf12~dgamma(1,0.01)
  tau.alpha~dgamma(1,0.01)
  zalpha.lin~dgamma(1,0.01)
}


Mylist = list(MyData, TT, W, K, N, adj,  NumCells, sumNumNeigh, num, R, I, C, num_t,
       NI, sumNumNeigh_t, adj_t,
              C_t, Lt, L, n.all, dat_unempl, LA1, theta0_y, beta_y, delta_v, delta_y,
              parameters.trad_multl_spatialg, parameters.Car_anova.modelg, inits2g,
               inits2g_anov, inits2g2, inits2g_anov2, inits2g_conv, inits2g_conv2,
                parameters.convolutiong)

#run simulation
d = NULL #start with an empty dataset
funct_data <- function(data, job, tau_2_S, tau_2_T, rho_S, rho_T){
  #tau_2 <- tau_2_rho[index, "tau_2"]
  #rho <- tau_2_rho[index, "rho"]
  MyData=data[[1]]
  TT=data[[2]]
  W=data[[3]]
  K=data[[4]]
 N=data[[5]]
 adj = data[[6]]
 NumCells=data[[7]]
 sumNumNeigh=data[[8]]
 num=data[[9]]
 R=data[[10]]
 I=data[[11]]
 C=data[[12]]
 num_t = data[[13]]
 NI=data[[14]]
 sumNumNeigh_t = data[[15]]
 adj_t=data[[16]]
 C_t=data[[17]]
 Lt=data[[18]]
 L=data[[19]]
 n.all = data[[20]]
 dat_unempl = data[[21]]
 LA1 = data[[22]]
 theta0_y = data[[23]]
 beta_y = data[[24]]
 delta_v = data[[25]]
 delta_y = data[[26]]
 parameters.trad_multl_spatialg = data[[27]]
 parameters.Car_anova.modelg = data[[28]]
 inits2g = data[[29]]
 inits2g_anov=data[[30]]
 inits2g2 = data[[31]]
 inits2g_anov2=data[[32]]
 inits2g_conv=data[[33]]
 inits2g_conv2=data[[34]]
 parameters.convolutiong=data[[35]]

 ### Generate data with CARanova

 distance <- as.matrix(dist(1:TT))
 D <-array(0, c(TT,TT))
 D[distance==1] <-1

 #From W the precision matrix can be computed for the multivariate Gaussian representation
 #of the spatial random effects ?? from (6) as follows:
Q.W <- rho_S * (diag(apply(W, 2, sum)) - W) + (1-rho_S) * diag(rep(1,K))
 Q.W.inv <- solve(Q.W)
```

```
  library("MASS")
  phi <- mvrnorm(n = 1, mu = rep(0, K), Sigma = (tau_2_S * Q.W.inv))
  phi <- phi - mean(phi)
  phi.long <- rep(phi, TT)
  Q.D <- rho_T*(diag(apply(D, 2, sum)) - D) + (1-rho_T) * diag(rep(1, TT))
  Q.D.inv <- solve(Q.D)
  Delta <- mvrnorm(n = 1, mu = rep(0, TT), Sigma = (tau_2_T * Q.D.inv))
  Delta <- Delta - mean(Delta)
  delta.long <- kronecker(Delta, rep(1, K))

  x <- rnorm(n = n.all, mean = 0, sd = 1)
  gamma <- rnorm(n = n.all, mean = 0, sd = sqrt(0.01))

  carerror =  phi.long + delta.long + gamma
  #carerror_cl= phi.long_cl

  ## Add Stteil to the generated data set at the district level for future merge
dat_unempl = data.frame(dat_unempl, LA1@data[,-c(9,10)])
  dat_unempl_1 = dat_unempl
  dat_unempl_1$period = rep(1,108)
  dat_unempl_2 = dat_unempl
  dat_unempl_2$period = rep(2,108)
  dat_unempl_3 = dat_unempl
  dat_unempl_3$period = rep(3,108)
  dat_unempl_4 = dat_unempl
  dat_unempl_4$period = rep(4,108)
 dat_unempl_5 = dat_unempl
  dat_unempl_5$period = rep(5,108)
  dat_unempl_long = rbind(dat_unempl_1, dat_unempl_2, dat_unempl_3,
  dat_unempl_4, dat_unempl_5)
  dat_unempl_carerror = data.frame(dat_unempl_long, carerror)

  ### merge with data at geographycal level
  data_long = merge(MyData, dat_unempl_carerror[,c("period","Stteil_num","Stteil",
   "Stadt", "carerror", "unempl")], by= c("Stteil", "period"), all.x=T)
  #data_long = merge (data_time, dat_unempl_carerror, by = "Stteil")
  MyData1 = data_long
  MyData1$unempl_bar <- mean(MyData1$unempl)
  MyData1$unempl_c <- MyData1$unempl - MyData1$unempl_bar
  MyData1$carerror = as.vector(MyData1$carerror)
  MyData1$carerror_cl = as.vector( MyData1$carerror_cl)
  MyData1$period_c=MyData1$period -1

  MyData1$yij <- theta0_y + beta_y[1]*MyData1$green_c + beta_y[2]*MyData1$age_period_c +
    beta_y[3]*MyData1$unempl_c+
    delta_y*MyData1$period +MyData1$e +MyData1$r0ij  +MyData1$r1ij*MyData1$period +
    MyData1$carerror

  mydata_W = list(yij=MyData1$yij, Stt_num = MyData1$Stt_num, id_num= MyData1$id_num,
  period = MyData1$period, period_c = MyData1$period_c,
                  green_c = MyData1$green_c, unempl_c = MyData1$unempl_c,
                  age_c = MyData1$age_c, age_period_c=MyData1$age_period_c, adj = adj,
                  NumCells=NumCells,
                  sumNumNeigh=sumNumNeigh, num=num, R=R, I=I, C=C, num_t = num_t, N=N,
                  TT=TT, K=K, NI=NI, sumNumNeigh_t = sumNumNeigh_t, adj_t=adj_t, C_t=C_t,
                   Lt=Lt, L=L)


  ## Transform data in a form usable by  WinBUGS
  library(R2WinBUGS)
   data_WB <- mydata_W
  return(data_WB)
}

### Creating Jobs

# parameters for the problem
tau_2_S = c(0.009, 0.009, 0.009, 0.009, 0.8, 0.8, 0.8, 0.8, 3, 3, 3, 3)
tau_2_T =  c(3, 0.8, 0.009, 3, 3, 0.8, 0.09, 0.8, 3, 3, 0.8, 3)
rho_S= c(0.9, 0.5, 0.09, 0.9, 0.5, 0.5, 0.5, 0.9, 0.5, 0.09, 0.5, 0.9)
rho_T= c(0.09, 0.5, 0.9, 0.9, 0.09, 0.5, 0.9, 0.9, 0.09, 0.9, 0.5, 0.9)
data_param=data.frame(tau_2_S, tau_2_T, rho_S, rho_T)
exp_par = list(car_long_prob = as.data.table(data_param))

# algorithm design: try combinations of kernel and epsilon exhaustively,
# Parameters for the function
funct_par = list(
  Car_conv.model_model= data.table(n.iter = 150000),
  Car_anova.model_model = data.table(n.iter = 150000),
  trad_multl_spatial_model =  data.table(n.iter = 150000)
)

ids = addExperiments(exp_par, funct_par, repls = 4)
```

```
## Before submitting the jobs

summarizeExperiments(by = c("problem", "algorithm", "tau_2_S", "tau_2_T", "rho_S",
"rho_T"))

### select some jobs only
id1 = head(findExperiments(algo.name = "trad_multl_spatial_model"), 1)
print(id1)
## Submit jobs
ids[ , chunk := 1] ### envois moi tous les jobs la en un bloc

batchtools::submitJobs(
  ids = ids, ### mapping entre le registre et la funcztiom
  resources = list(ntasks = 1,
                   account="lsickstadt",
                   ncpus = 1,
                 #  atonce = 18,
                   walltime = 600, ###
                   partition = "all", ###
                   memory = 30000L,
                   chunks.as.arrayjobs = TRUE),
  reg=reg
)
```

B.5 Multilevel conditional autoregressive models, MLM CARs

Here, we consider essential parts of the MLM CAR as well as the MLM RCAR.

Model specification for MLM CAR in BUGS:

```
bugscar_model <- function(){
  for(i in 1:N){

    Y[i] ~ dnorm(mu[i],tau.sf12) #

    mu[i] <- beta_0 + beta_1*green0_c[i] + beta_2*age_c[i] +
      beta_3*unempl_c[i] + beta_4*sex[i] + u0[Stteil_num[i]]
  }
```

```
# Define the zero vector
for(i in 1:K) {
  zero0[i] <- 0
}

#Spatial Unobserved Effect Prior


# Define Q

# Q[1:K,1:K] <- rho.S*R[,] + (1-rho.S)*I[,]

for(l in 1:K){ for (j in 1:K){
  Q[l,j] <- rho.S*R[l,j] + (1-rho.S)*I[l,j]
}}

# prior.T[1:K, 1:K] <- tau.invs*Q[,]

for(l in 1:K){ for (j in 1:K){
  prior.T[l,j] <- (Q[l,j])/tau.invs
}}

phi[1:K]~dmnorm(zero0[], prior.T[,])

# mean.phi[1,1:K] <- rho.T*phi[1,]




for(i in 1:K){
  u0[i] <- phi[i]

}
beta_0 ~  dflat()
beta_1 ~  dnorm(0.0,0.001)
beta_2 ~  dnorm(0.0,0.001)
beta_3 ~ dnorm(0.0,0.001)
beta_4 ~ dnorm(0.0,0.001)
rho.S~dunif(0,1)
tau.sf12~dgamma(0.001,0.01)
```

```
  sigma.sf12<-1/(tau.sf12*tau.sf12)
  tau.invs ~ dunif(0,1)
}
```

Model specification for MLM RCAR in BUGS:

```
# C restricted CAR

## Define the L matrix

H <- as.matrix(dat_unempl$unempl)
Pcb = I -H%*%solve(t(H)%*%H)%*%t(H)
L <-t(eigen(Pcb)$vectors[,eigen(Pcb)$values > min(eigen(Pcb)$values)])
Lt <- t(L)


bugscar_modelR <- function(){
  for(i in 1:N){
    Y[i] ~ dnorm(mu[i],tau.sf12)

    mu[i] <- beta_0 + beta_1*green0_c[i] + beta_2*age_c[i] +
      beta_3*unempl_c[i] + beta_4*sex[i] + u0[Stteil_num[i]]
  }


  # Define the zero vector
  for(i in 1:K) {
    zero0[i] <- 0
  }

  # Define the zero vector
  for(i in 1:K-1) {
    zero00[i] <- 0
  }


  #Spatial Unobserved Effect Prior


  # Define Q
  for(l in 1:K){ for (j in 1:K){
    Q[l,j] <- rho.S*R[l,j] + (1-rho.S)*I[l,j]
  }}



  for (i in 1:K-1){
    for (j in 1:K){
    #  inprod_LQ[] <- L[i,] * Q[,j] sum(inprod_LQ[])#
      LQ[i,j] <- inprod(L[i,],Q[,j])
    }
  }


  for (i in 1:K-1){
    for (j in 1:K-1){
     # inpro_Q22[] <-LQ[i,] * Lt[,j] sum(inpro_Q22[]) #
      Qt22[i,j] <- inprod(LQ[i,],Lt[,j])
    }
  }


  gammaastast[1:K-1] ~ dmnorm(zero00[], Qt22[,])

  for(k in 1:K){
    Lgammaastast[k]<- inprod(Lt[k,],gammaastast[])/sqrt(tau.invs)
  }


  for(i in 1:K){
    u0[i] <- Lgammaastast[i]


  }
  beta_0 ~  dnorm(0.0,0.001)
  beta_1 ~  dnorm(0.0,0.001)
  beta_2 ~  dnorm(0.0,0.001)
  beta_3 ~ dnorm(0.0,0.001)
  beta_4 ~ dnorm(0.0,0.001)
  rho.S~dunif(0,1)
  tau.sf12~dgamma(0.001,0.01)
  sigma.sf12<-1/(tau.sf12*tau.sf12)
  tau.invs ~ dunif(0,1)
}
```

B.6 Implementation of the classical multilevel model in SAS

Here, we provide the main ideas of the implementation in SAS.

The unconditional model: model 1

```
PROC MIXED data=MyData covtest noclprint method = ML;
class Stt_num;
model yij=/solution ddfm = SATTERTHWAITE;
random intercept / sub=Stt_num type=vc;
run;
```

model 2

```
PROC MIXED data=MyData covtest noclprint method = ML;
class Stt_num;
model yij = ndvi100 sex1 age/solution ddfm = SATTERTHWAITE;
random intercept ndvi100 /sub=Stt_num type=vc;
run;
```

model 3

```
PROC MIXED data=MyData covtest noclprint method = ML;
class Stt_num;
model yij = ndvi100 sex1 age unemplBL/solution ddfm = SATTERTHWAITE;
random intercept ndvi100 /sub=Stt_num type=vc;
run;
```