UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Review of Bayesian hierarchical areal wombling techniques with application to COVID-19

Edmond Sacla Aide

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :

Prof. dr. Christel FAES

2022
2023

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Review of Bayesian hierarchical areal wombling techniques with application to COVID-19

Edmond Sacla Aide

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :

Prof. dr. Christel FAES

Abstract

Background: Many research areas and especially epidemiologists and geostatistical researchers have shown increasing interest in boundary analysis known as spatial wombling. Spatial models for areal data are used to get a smooth risk surface map by accounting for the variability in the areas. It is also interesting to identify the differences in adjacent areas and highlight those boundaries that have a high difference amongst neighboring areas like the difference boundaries. Several methods have been proposed in the literature to formally identify those boundaries.

Objectives: This study aimed to conduct a review of different wombling methods for areal data and to use the methods to investigate difference boundaries in the COVID-19 incidence at different time points in Belgium (Wave 1, Wave 2, Wave 3).

Methodology: We searched 3 English language databases (PubMed, JSTOR, and Google Scholar) for studies published between 1951 and April 15, 2023. Eligible studies were spatial boundary analysis for areal data with application to disease or health-related outcomes. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist was used to conduct the systematic review. A spatial univariate areal wombling of COVID-19 incidence in Belgium was conducted throughout the three different waves of the pandemic applying both algorithm-based and model-based wombling. For the model-based wombling, we used both globally and locally smooth Conditional Autoregressive Priors (CAR) models. A residual-based wombling was conducted as well.

Results: After screening, a total of 24 papers were included in the review. Two main wombling techniques exist: Crisp wombling and Fuzzy wombling and they differ from each other by the way of calculating the boundary membership value (BMV). The wombling can be either algorithm-based or model-based using Bayesian hierarchical models.

Univariate spatial wombling of COVID-19 incidence in Belgium has identified boundaries in the incidence map during the three different waves. Wave 2 presented a more remarkable difference splitting the country into two regions: the North marked by a medium incidence and the South marked by a strong incidence. Algorithm-based wombling has generally identified more boundaries compared to model-based wombling. The residual wombling demonstrated that the identification of the boundaries may be correlated with some spatially oriented covariates.

Conclusion: Wombling of COVID-19 incidence has identified boundaries in the map across the different waves of the pandemic. The type of wombling and the Bayesian hierarchical model affected the number of identified boundaries.

Key Words: Wombling; Areal data; Bayesian hierarchical model; COVID-19; Belgium

Acknowledgments

I am highly thankful to the Almighty God for bestowing me with strength and good health during my entire time in my education and completing my essay project. It would not have been possible without you, God.

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Christel FAES for their continuous massive guidance, mentorship, academic support, and time to scrutinize my work that led to the timely completion of my study.

Much appreciation to VLIR-OUS for the scholarship, without financial sponsorship, I could not have enrolled in the program. Thanks to Hasselt University fraternity for the enabling environment, they created for studies that helped me to fully concentrate on my studies.

I also acknowledge my friends for their support and encouragement throughout my study.

My deepest gratitude goes to my beloved family members and relatives for always believing in me. Thank you, Mum, for your prayers and support.

May God always bless you all!

Contents

List of Figures v
List of Tables vi

1 Introduction **1**

2 Part I: Spatial, Spatio-temporal, and Multivariate areal Wombling of diseases: a systematic review **4**
2.1 Methodology 4
2.1.1 Data sources and search strategy 4
2.1.2 Eligibility criteria 4
2.1.3 Study selection 4
2.1.4 Data collection process 5
2.1.5 Ethics Statement 5
2.2 Results 5
2.2.1 Search results 5
2.2.2 Characteristics of the studies included 5
2.2.3 Spatial areal Modeling with Conditional Autoregressive Priors . . . 6
2.2.4 Evolution of spatial areal wombling analysis over time 11

3 Part II: Wombling of Spatial COVID-19 incidence **16**
3.1 Data description 16
3.2 Methodology 16
3.2.1 Exploratory data analysis 16
3.2.2 Traditional or algorithm-based wombling 17
3.2.3 Spatial areal modeling of COVID-19 incidence with conditional autoregressive priors 17
3.2.4 Bayesian hierarchical model based wombling 18
3.2.5 Model-based wombling using dissimilarity metric 18
3.2.6 Implementation of wombling techniques 18
3.3 Results 19
3.3.1 Exploratory data analysis 19
3.3.2 Traditional or Algorithmic-based wombling 20
3.3.3 Spatial areal modeling with conditional autoregressive priors 21
3.3.4 Bayesian hierarchical model based wombling 25
3.3.5 Residual-based wombling 31
3.3.6 Model-based wombling using dissimilarity metric 33

4 Discussion **37**

5 Ethical thinking, societal relevance, and stakeholder awareness **39**

6 Conclusion and future work **40**

References **41**

Appendices **44**

List of Figures

1 Flow diagram of study selection 6
2 Exploratory data analysis 20
3 Global smooth model-based maps 23
4 Locally smooth model-based maps 25
5 Comparison of wombling for Wave 1 27
6 Comparison of wombling for Wave 2 29
7 Comparison of wombling for Wave 3 31
8 Mean-based vs Residual-based wombling 33
9 Map displaying the estimated incidence rate and the locations of the boundaries 35
A.2.1 Convergence of the Markov chains 1 (a: wave1; b: wave2; c: wave3) 55
A.2.2 Model-based wombling for wave 1 (municipalities in boundaries 56
A.2.3 Model-based wombling for wave 2 (municipalities in boundaries 57
A.2.4 Model-based wombling for wave 3 (municipalities in boundaries 58

List of Tables

1 Data description 16
2 Algorithm-based wombling 21
3 Global CAR model selection 22
4 Summary of global CAR models 23
5 Locally CAR model selection 24
6 Summary of locally smooth models 25
7 model-based wombling wave 1 26
8 model-based wombling wave 2 28
9 model-based wombling wave 3 30
10 Comparison of mean-based and residual-based wombling 32
11 Boundary detection using dissimilarity metric 34
12 Comparison of different wombling approach 36
13 Sensitivity analysis (average deprivation score in Globally smooth CAR model) 36

1 Introduction

Spatial data analysis of health-related outcomes has received increasing attention in the spatial statistics literature. Indeed, as Geographical Information Systems (GIS) become more widely available, researchers and administrators in public health are increasingly experiencing areal statistics aggregated as case counts or rates across areal units or regions (states, counties, census tracts, or ZIP codes) (Li et al., 2015).

In public health applications, spatial data analysis frequently begins with statistical models for areal data, which include regional aggregates of health outcomes across delineated administrative units (Gao et al., 2022). By smoothing across and borrowing information from its geographical neighbors, statistical models for areal data can account for known sources of variability in the data as well as sparsely sampled regions (Banerjee and Gelfand, 2006; Li et al., 2015). Spatial models for areal data are used to get a smooth risk surface map by accounting for the variability in the areas.

An especially pressing issue is determining statistically significant differences between surrounding locations and, as a result, defining the geographical barriers or difference boundaries that separate them. The fundamental causes of these borders or barriers are usually of scientific and administrative importance (Li et al., 2015). Spatial analysts, on the other hand, have recently shown an increasing interest in finding zones or boundaries that suggest significant changes in the values of spatially oriented variables (Lu and Carlin, 2005). In other words, interest can also be in identifying the differences in adjacent areas, and highlighting those boundaries that have a high difference amongst neighboring areas, i.e. the difference boundaries.

The purpose of spatial boundary analysis, which is the determination of boundaries on a map that separates areas with higher and lower values, is to uncover significant barriers and the underlying influences responsible for these barriers (Lu and Carlin, 2005). This boundary detection problem is sometimes referred to as wombling, after Womble's seminal study (Womble, 1951). Since then, wombling has gained popularity as a method for evaluating geographical linkages among many other disciplines, including genetics, demography, linguistics, ecology, and environmental science (Li et al., 2015). In disciplines like landscape topography, systematic biology, sociology, ecology, and public health, the process we generally refer to as "wombling" is also called barrier analysis or edge identification (Li et al., 2015). The method has been first developed for point-referenced data and extended later to areal data.

Boundary analysis has significantly evolved since Womble's groundbreaking paper in 1951, moving from straightforward algorithmic or deterministic methods to today's highly so-

phisticated methods using Bayesian hierarchical models. This evolution has been made feasible by the development of increasingly complex Bayesian statistical techniques. The inadequacies of earlier methods are considered when creating new approaches.

On December 31, 2019, the World Health Organization (WHO) regional office was notified about a cluster of pneumonia cases of unknown origin related to a market in Wuhan, China (Zhu et al., 2020). SARS-COV-2, a new coronavirus, was identified as the cause of the infections and has subsequently expanded globally (Zhu et al., 2020). Several methodologies have been explored to predict the pandemic's outbreak, including compartmental models and statistical models. Many epidemiological studies have been conducted to investigate the regional spread of COVID-19 (Fatima et al., 2021). Epidemiological analysis of the outbreak has been used to estimate epidemiologically relevant parameters (Read et al., 2021; Li et al., 2020; Yang et al., 2020; Guan et al., 2020; Backer et al., 2020), and available mathematical models have been used to track and predict the spread of the epidemics (Gilbert et al., 2020). COVID-19 pandemic severity has been greatly influenced by crowding, as indicated by increased prevalence in large cities compared to smaller cities and rural areas (Read et al., 2021). Every country has been affected differently and has shown a distinct pattern of incidence and death as a result of a variety of underlying factors (Fatima et al., 2021). Studying the distribution of the disease and how it spreads across time and space is fundamental to both health geography and spatial epidemiology (Glass, 2000). Understanding the geographical distribution of infection and its relationship with the population and environment is critical, especially in the early phases of an outbreak (Kang et al., 2020). Because transmission rates are higher when people are close to each other, the concept of spatial and spatial-temporal proximity is profoundly associated with infectious disease transmission (Pfeiffer et al., 2008). Disease spatial patterns frequently indicate linkages between disease and potential risk factors in a geographic area (Waller, 2006).

Many studies have been conducted to investigate the spatial and spatiotemporal trends in COVID-19 incidence in various countries worldwide. By accounting for the variability in the areas, these algorithms produce a smooth risk surface map. Yet, there may be an interest in discovering discrepancies in adjacent areas and highlighting those boundaries that have a substantial difference between neighboring areas (the difference boundaries).

Boundary analysis assists in identifying regional variances across shared boundaries in order to find homogeneous zones or key barriers (Lu and Carlin, 2005). In the sphere of public health, wombling is particularly effective for improving disease preventive and control decision-making by finding zones of significantly differing incidence or death (Lu et al., 2007). This could assist decision-makers in allocating greater resources to the pandemic's

worst-impacted regions.

The remainder of this thesis is organized as follows. Section 2 presents a systematic review of areal wombling methods. Section 3 conducts wombling of spatial COVID-19 incidence in Belgium at different time points in Belgium (Wave 1, Wave 2, Wave 3). Section 4 discusses our findings. Section 5 presents the ethical thinking, societal relevance, and stakeholder awareness and Section 6 presents the conclusion, and limitations and suggests directions for future research.

2 Part I: Spatial, Spatio-temporal, and Multivariate areal Wombling of diseases: a systematic review

2.1 Methodology

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist was used to conduct the systematic review (Page et al., 2021).

2.1.1 Data sources and search strategy

To reduce the potential for bias, a comprehensive search technique involving various electronic literature databases was used. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) criteria (Page et al., 2021), we searched three English literature databases from 1951 (after the foundational work of Womble, 1951) through 30 March 2023 (PubMed, JSTOR, and Google Scholar). These databases were chosen because English-speaking researchers frequently utilize them.
The following phrases were combined with the boolean expression 'OR' within groups or 'AND' between groups to form the search syntax: (1) Wombling – Boundary analysis – Detection of zone – Difference boundaries; (2) Areal data – Lattice data – Polygon data – Aggregate data – Spatially homogenized data; and (3) Spatial – Spatiotemporal – Spatial dynamics.

2.1.2 Eligibility criteria

The following eligibility criteria were defined: (1) English papers published from 1951 to April 15, 2023. (2) Spatial boundary analysis or Detection of difference boundaries. (3) Research scales at the municipality level, county level, and state level. (4) Application to disease, public health, or health-related outcome. (5) Published in a peer-reviewed journal. The exclusion criteria were as follows: (1) review articles; editorials or published letters (2) books; (3) other fields of application: agriculture, ecology, population genetics, vegetation sciences, demography, and criminology.

2.1.3 Study selection

After deduplication, the papers were screened using the title, abstracts, and keywords, as well as the entire text of the publications when more information was required for eligibility identification.

2.1.4 Data collection process

We extracted information about wombling techniques as well as the Bayesian hierarchical models used as background in the case of model-based wombling. Descriptive details obtained included: Journal – Title – Authors – Country - Year of publication – Disease – Research scale (counties, zip codes, municipalities) – Type of wombling – Type of data (Spatial, Spatio-temporal) – Number of outcomes (univariate, Multivariate) – Software package if available.

2.1.5 Ethics Statement

Because this was a systematic review, no ethical approval was required.

2.2 Results

2.2.1 Search results

A preliminary systematic literature search yields 134 results. After deleting duplicates, we kept 123 records for the title and abstract screening. Because they did not match the review eligibility requirements, 44 records were excluded. 55 records are unrelated to disease from the 79 potentially relevant studies reviewed in full text. Population genetics, ecology, vegetative sciences, spatial demography, and criminology were among the application fields we excluded from this review. Other papers used point-referenced or point-process data. The screening method results in 24 eligible studies. The screening process is presented in the diagram of Figure 1.

2.2.2 Characteristics of the studies included

We collected contextual elements from the papers included in the review after a short scan of the text. 19 (79.17 %) of the 24 research focused on disease spatial wombling, 2 (8.33%) papers on spatiotemporal wombling, and 3 (12.5%) papers on multivariate wombling. For the model's applicability, all of the studies employed county-level data. The literature studied two types of wobbling: Crisp wobbling and Fuzzy wobbling. Wombling can also be algorithmic, or model-based. Model-based wombling employs Bayesian hierarchical models as well as non-parametric Bayesian hierarchical models with adjacency modeling.
There are three options for model-based wombling: mean-based wombling, residual-based (random effect-based) wombling, and variance-based wombling.
16 of the 19 studies dealt with univariate wombling (one outcome) while 3 dealt with multivariate wobbling.

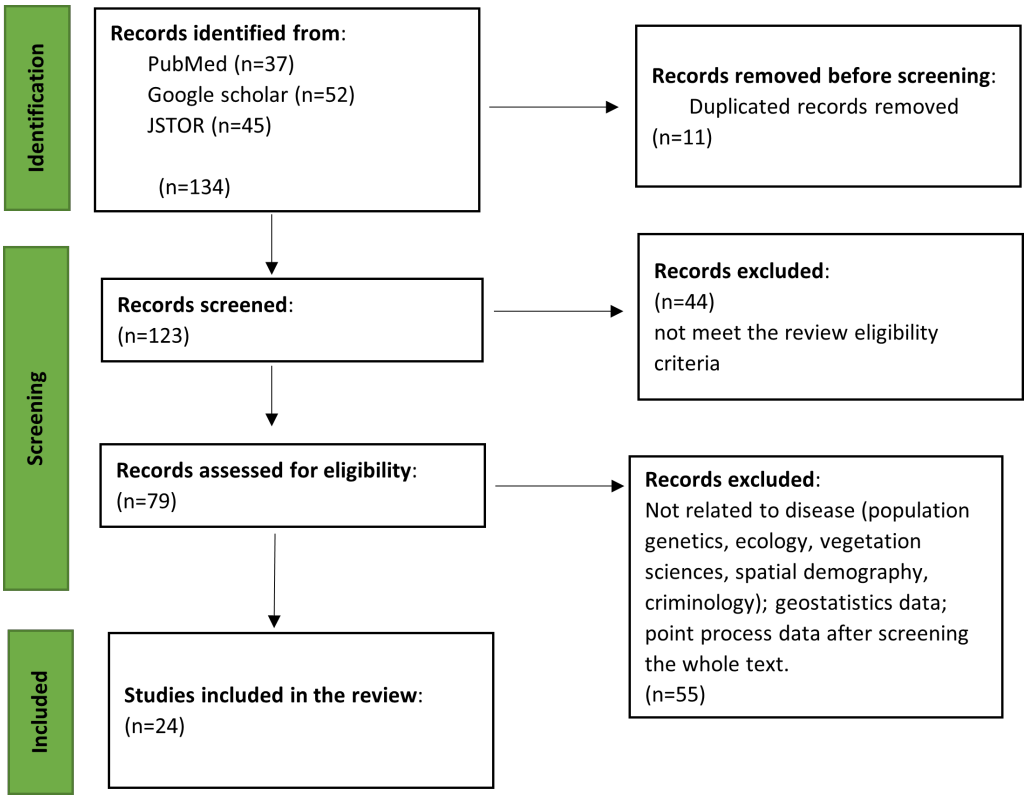

Figure 1: Flow diagram of study selection

Furthermore, the diseases investigated in the papers include influenza, pneumonia, breast cancer, colorectal cancer, pancreatic cancer, lung cancer, cervix cancer, respiratory and circulatory disorders, and others.

2.2.3 Spatial areal Modeling with Conditional Autoregressive Priors

Areal data show spatial autocorrelation, with observations from nearby areal units likely to have comparable values. A portion of this spatial autocorrelation may be modeled by known covariate risk factors in a regression model, although the spatial structure is often retained in the residuals after these covariate effects are accounted for. Unmeasured confounding, neighborhood effects, and grouping effects can all cause residual spatial autocorrelation. As part of a Bayesian hierarchical model, the most frequent solution for residual autocorrelation is to augment the linear predictor with a collection of spatially autocorrelated random effects. These random effects are commonly represented with a conditional autoregressive prior (Besag et al., 1991), which creates spatial autocorrelation via the areal unit adjacency

structure. Several CAR priors have been presented in the literature, including the intrinsic and Besag-York-Mollié (BYM) models (Besag et al., 1991), as well as the (Leroux et al., 2000) alternative.

These CAR priors, on the other hand, cause random effects to reflect a single global level of spatial autocorrelation ranging from independence to remarkable spatial smoothness. A uniform degree of spatial autocorrelation for the entire region may be implausible for real-world data, which may instead show sub-regions of spatial autocorrelation separated by discontinuities. Several techniques, notably (Lee and Mitchell, 2012) and (Lee and Sarran, 2015) have been presented for extending the class of CAR priors to deal with localized spatial smoothing among random effects.

In this section, we present all the details about the globally and locally spatial smoothing CAR models.

The study region S is partitioned into K non-overlapping areal units $S = S_1, ..., S_K$, which are linked to a corresponding set of responses $Y = (Y_1, ..., Y_K)$, and a vector of known offsets $O = 0_1, ..., 0_K$. The spatial variation in the response is modeled by a matrix of covariates $X = (x_1, ..., x_K)$ and a spatial structure component $\Psi = (\Psi_1, ..., \Psi_2)$, the latter of which is included to model any spatial autocorrelation that remains in the data after the covariate effects have been accounted for (Lee, 2017). For a count outcome variable, the spatial generalized linear mixed model is given by:

$Y_k \sim Poisson(\mu_k)$

and $ln(\mu_k) = x_k^T \beta + O_k + \Psi_k$

Globally smooth CAR models

The globally smooth model is one that employs priors that require random effects to reflect a single global level of spatial autocorrelation, ranging from independence to strong spatial smoothness. Independence, Intrinsic CAR model (Besag model), Besag-York-Mollié (BYM) model, and Leroux model are the globally smooth CAR model. A model selection technique based on DIC (Deviance Information Criterion) can be used to find the best globally smooth model. The use of AIC and BIC does not seem sensible here as the theory that supports them does not extend to the random effects setting.

a-Besag-York-Mollie (BYM) CAR model

The convolution or Besag-York-Mollie (BYM) CAR model outlined in (Besag et al., 1991) contains both spatially autocorrelated and independent random effects and is given by:

$$
\begin{aligned}
\psi_k &= \phi_k + \theta_k \\
\phi_k \mid \phi_{-k}, \mathbf{W}, \tau^2 &\sim \mathrm{N}\left(\frac{\sum_{i=1}^{K} w_{ki}\phi_i}{\sum_{i=1}^{K} w_{ki}}, \frac{\tau^2}{\sum_{i=1}^{K} w_{ki}}\right) \\
\theta_k &\sim \mathrm{N}\left(0, \sigma^2\right) \\
\tau^2, \sigma^2 &\sim \text{Inverse-Gamma}(a, b).
\end{aligned}
$$

Here $\theta = (\theta_1, ..., \theta_K)$ are independent random effects with zero mean and constant variance, while spatial autocorrelation is modeled via random effects $\phi = (\phi_1, ..., \phi_K)$. The conditional expectation for the latter is the average of the random effects in nearby areas, while the conditional variance is inversely proportional to the number of neighbors. This is appropriate because if the random effects are significantly spatially autocorrelated, then the more neighbors a region has, the more information there is about the value of its random effect from its neighbors, and so the uncertainty decreases.

This model contains two random effects for each data point, and as only their sum is identifiable from the data only $\Phi_k = \phi_k + \theta_k$ is returned to the user.

b- Leroux model

Leroux et al. (2000) developed an alternative CAR prior to modeling different levels of spatial autocorrelation with a single set of random effects.

$$
\begin{aligned}
&\psi_k = \phi_k \\
&\phi_k \mid \phi_{-k}, \mathbf{W}, \tau^2, \rho \sim \mathrm{N}\left(\frac{\rho \sum_{i=1}^{K} w_{ki}\phi_i}{\rho \sum_{i=1}^{K} w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{i=1}^{K} w_{ki} + 1 - \rho}\right) \\
&\tau^2 \sim \text{Inverse-Gamma}(a, b) \\
&\rho \sim \text{Uniform}(0, 1)
\end{aligned}
$$

Here ρ is a spatial dependence parameter taking values in the unit interval and can be fixed. Specifically, $\rho = 1$ corresponds to the intrinsic CAR model (defined for ϕ in the BYM model above), while $\rho = 0$ corresponds to independence ($\phi_k \sim N(0, \tau^2)$).

Locally smooth CAR models

The globally smooth model makes use of a uniform level of spatial autocorrelation for the entire region, and this may be unrealistic for real data, which instead may exhibit subregions of spatial autocorrelation separated by discontinuities. One of the approaches to overcome this issue is to use a locally smooth model. We used here two CAR priors to deal with localized spatial smoothing amongst the random effects: (Lee and Mitchell, 2012) and (Lee and Sarran, 2015). Again, the DIC criterion can be used to select the best localized

smooth model that fits well the data.

For the set of random effects, the CAR priors provided above impose a single global level of spatial smoothing, which for Leroux model is controlled by ρ. This is illustrated by the partial autocorrelation structure implied by that model, which for (ϕ_k, ϕ_j) is given by

$$\text{COR}\left(\phi_k, \phi_j \mid \phi_{-kj}, \mathbf{W}, \rho\right) = \frac{\rho w_{kj}}{\sqrt{\left(\rho \sum_{i=1}^{K} w_{ki} + 1 - \rho\right)\left(\rho \sum_{i=1}^{K} w_{ji} + 1 - \rho\right)}}$$

The idea is that for non-neighboring area units ($w_{kj} = 0$) the random effects are conditionally independent, while for neighboring area units ($w_{kj} = 1$) their partial autocorrelation is controlled by ρ.

a-Lee and Mitchell (2012)

Lee and Mitchell (2012) proposed a method to capture localized spatial autocorrelation and identify boundaries in the random effects. If areal neighbors are adjacent ($k_j = 1$), (ϕ_k, ϕ_j) are spatially autocorrelated and smoothed over in the modeling process. (ϕ is the random effect of a given areal unit). If adjacent neighbors are not adjacent ($w_{kj} = 0$), no smoothing between (ϕ_k, ϕ_j), and they are modeled as a conditional independent.

The model makes use of the Leroux model and fixed ρ at 0.99 and this ensures that the random effects exhibit strong spatial smoothing globally, which can be altered locally by estimating $w_{kj}|k \sim j$.

Each adjacency matrix W_{kj} is modeled as a function of dissimilarity between areal units (S_k, S_j).

The dissimilarity matrix can be estimated based on social or physical factors (e.g.: rate of smoking, presence of river, railway line, ... etc.). Based on the dissimilarity, two approaches are proposed for the estimation of $w_{kj}|k \sim j$.

Binary model

$$w_{kj}(\boldsymbol{\alpha}) = \begin{cases} 1 & \text{if } \exp\left(-\sum_{i=1}^{q} z_{kji}\alpha_i\right) \geq 0.5 \text{ and } k \sim j \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha_i \sim \text{ Uniform } (0, M_i) \qquad \text{for } i = 1, \ldots, q$$

M_i is the upper limits for the priors for α_i and depends on the distribution of Z_{kj} and are chosen weakly informative.

Non-binary model

$$w_{kj}(\boldsymbol{\alpha}) = \exp\left(-\sum_{i=1}^{q} z_{kji}\alpha_i\right)$$
$$\alpha_i \sim \text{Uniform}(0, 50) \quad \text{for } i = 1, \ldots, q.$$

The q regression parameters $\alpha = (\alpha_1, ..., \alpha_q)$ represent the effect of dissimilarity Z_{kj} metrics on $w_{kj}|k \sim j$.
For the binary model, if $\alpha_i < -ln(0.5)/max(Z_{kji})$, then the ith dissimilarity metric has not solely identified any boundaries because exp(-$\alpha_i Z_{kji}$) > 0.5 for all $k \sim j$.
Finally, W_{kj} contains 3 values: NA (non-adjacent) ; 0 (no boundary); 1 (boundary).

b-Lee and Sarran (2015)

An option to the above is to add a piecewise constant intercept or cluster model to the set of spatially smooth random effects, allowing for huge jumps in the mean surface between neighboring areal units in different clusters. The idea here is that in addition to the global random effect (ϕ) in the Leroux model, a set of spatially smooth random intercepts are incorporated into the model. The constant random intercepts are defined for each cluster of areas in the study areas.
Lee and Sarran (2015) proposed a model to partition the region under study into G clusters each with its own intercept term $(\lambda_1, ..., \lambda_G)$.

$$\psi_k = \phi_k + \lambda_{Z_k}$$
$$\phi_k \mid \phi_{-k}, \mathbf{W}, \tau^2 \sim \text{N}\left(\frac{\sum_{i=1}^{K} w_{ki}\phi_i}{\sum_{i=1}^{K} w_{ki}}, \frac{\tau^2}{\sum_{i=1}^{K} w_{ki}}\right)$$
$$\tau^2 \sim \text{Inverse-Gamma}(a, b)$$
$$\lambda_i \sim \text{Uniform}\left(\lambda_{i-1}, \lambda_{i+1}\right) \quad \text{for } i = 1, \ldots, G$$
$$f\left(Z_k\right) = \frac{\exp\left(-\delta\left(Z_k - G^*\right)^2\right)}{\sum_{r=1}^{G} \exp\left(-\delta\left(r - G^*\right)^2\right)}$$
$$\delta \sim \text{Uniform}(1, M).$$

A weakly informative uniform prior is specified for the penalty parameter $\delta \sim Uniform(1, M)$ (by default M = 10), so that the data play the dominant role in estimating its value.
An area k is assigned to one of the G intercepts by $Z_k \epsilon\, 1, ..., G$. Z_k is penalized towards the middle intercept value so that extreme intercept classes (1 or G) may be empty.
Note that in our analysis, we used G=3 corresponding to the 3 language communities in

Belgium.

2.2.4 Evolution of spatial areal wombling analysis over time

Since Womble's foundational work in 1951, boundary analysis has evolved significantly over time, progressing from simple algorithmic methods to highly sophisticated methods used today. The development of increasingly advanced Bayesian statistical techniques has enabled this evolution. We present a brief overview of the development of wombling approaches in areal data. There are two types of wombling techniques: Crisp wombling and Fuzzy wombling.

a- Crisp areal wombling

The interest in areal wombling is the difference boundaries, which are those borders that separate two adjacent counties having dramatically different observed response values.

Algorithm-based wombling

The algorithm-based wombling consists of assigning a boundary likelihood value (BLV) to each area border based on a gradient distance metric between neighboring observations.

$$BLV_{ij} = \Delta_{ij} = \|Y_i - Y_j\| \tag{1}$$

with $\|.\|$ a distance metric

The locations with similar BLV value (lower or higher) are more likely to belong to the same difference boundary as the outcome change rapidly here. In Crisp wombling, when the BLV exceeds a predefined threshold (let's say $c \geq 0$), belongs to a boundary membership value (BMV) of 1 and 0 otherwise.

The algorithm-based wombling directly uses the outcome variable (Y_i) to compute the boundary likelihood value. This approach does not take into consideration the uncertainty in the outcome but also, difficult to access how concentrated is the distribution of the BLV.

Model-based wombling

To overcome the issue with algorithm-based Crisp wombling, Hodges et al. (2003) examined areal data using a linear model (Gaussian). However, a linear model is not applicable for the most prevalent type of areal data, count data. Instead of using the raw or brute value of the outcome in the computation of BLV, Lu and Carlin (2005) proposed a hierarchical approach to estimate the value of the outcome at each location. The idea is to first model the outcome variable as a function of measured covariates and spatial random effects. The posterior distribution of the BLVs is then used to calculate the BMV. The model selection procedure is used to select the hierarchical model that fits well the data. The traditional or algorithm-based BMV should be superior to the hierarchical model-based BMV as the

latter will properly account for uncertainty in the estimate values throughout the process rather than averaging this uncertainty out before the BLV is completed (Lu and Carlin, 2005). But how does the method work?

As count data is most commonly encountered in areal settings, a Poisson log-linear form is used:

$$Y_i \sim Poisson(\mu_i)$$

where $log(\mu_i) = log(E_i) + X_i'\beta + \phi_i$

This model allows a vector of region-specific covariates X_i (if available), and a random effect vector $\phi = (\phi_1, ..., \phi_N)'$ that is given a conditionally autoregressive (CAR) specification (Besag, 1974). A common form of the distribution (often called intrinsic CAR, or IAR model) (Besag et al., 1991) has improper joint distribution, but intuitive conditional distributions of the form:

$$\phi_i|\phi_{i\neq j} \sim N(\bar{\phi_i}, 1/(\tau m_i))$$

where N denotes the normal distribution, $\bar{\phi_i}$ is the average of the random effect of the regions that are adjacent to ϕ_i and m_i is the number of these adjacencies; this distribution is usually abbreviated as $CAR(\tau)$ with τ a typically set and equal to some fixed value, or assigned a distribution itself (usually a relatively vague γ distribution).

Markov chain Monte Carlo (MCMC) samples $\mu_i^{(g)}$, $g = 1, ..., G$ from the marginal posterior distribution $p(\mu_i|y)$ can be obtained for each i (Banerjee and Gelfand, 2006). For example, the model-based standardized mortality rate (η_i) is equal to:

$$\eta_i = \frac{\mu_i}{E_i}, i = 1, ..., N \tag{2}$$

The BLV for boundary (i, j) as:

$$\Delta_{ij} = \|\eta_i - \eta_j\| \tag{3}$$

for all i adjacent to j, Crisp or Fuzzy wombling boundaries are then based on the posterior distribution of the BLVs. In the case of Crisp wombling, we might define ij to be part of the boundary if and only if $E(\Delta_{ij}|y) > c$ for some constant $c > 0$, or if and only if $P(\Delta_{ij}|y > c) > c^*$ for some constant $0 < c^* < 1$.

As $\Delta_{ij}^{(g)} = \|\eta_i^{(g)} - \eta_j^{(g)}\|$, and the boundaries are based on their empirical distribution. The posterior means is estimated as:

$$\hat{E}(\Delta_{ij|y}) = \frac{1}{G}\Sigma_{g=1}^{G}\Delta_{ij}^{(g)} = \frac{1}{G}\Sigma_{g=1}^{G}\|\eta i^{(g)} - \eta_j^{(g)}\| \tag{4}$$

b- Fuzzy areal wombling

Algorithm-based wombling

Crisp wombling boundaries are straightforward and easy to understand, but assessing the certainty or magnitude of the distribution of the boundary likelihood value is complex. Fuzzy wombling boundaries are preferable because they do not rely on binary BMV.
Instead of defining a threshold, the BMV can also be defined as follows:

$$BMV_{ij} = \frac{\|Y_i - Y_j\|}{max(\|Y_i - Y_j\|)} \tag{5}$$

and indicate partial membership in the boundary.
The BMV in this case fluctuates between zero and one, indicating partial participation in the border.
BMVs with Fuzzy wombling can take values between zero and one (0,1). This makes it possible for some places to be more relevant in deciding the boundary. The typical Fuzzy technique avoids using a 0-1 choice to include a segment in the boundary, but if BMV is between zero and one, it may not be read as a probability of being part of a boundary because no stochastic model is linked with it.
A hierarchical Bayesian technique provides a suitable and practical option.

Model-based wombling

A hierarchical Bayesian model considers the process's stochasticity as well as a method to directly analyze the uncertainty in our Fuzzy BMV (availability of posterior distribution). To tackle the issues faced by the algorithm-based Fuzzy wombling, Lu and Carlin (2005) introduced a hierarchical Bayesian model. Fuzzy wombling is based on the posterior distribution of the BLVs after estimating the model-based value of the outcome. The model selection procedure is used to select the hierarchical model that fits well the data. The hierarchical Bayesian approach offers a direct and convenient solution. Suppose we select a cutoff c such that, were we certain a particular BLV exceeds c, we would also be certain the corresponding segment was part of the boundary. Estimates of Δ_{ij} are obtained using a Markov chain Monte Carlo (MCMC) algorithm to draw G samples of the modeled response η_i^g, $g = 1, ..., G$ from the posterior distribution $p(\eta_i|y)$ (where y represents observations of the response variable) for each areal unit i and each MCMC iteration g to obtain.
$\Delta_{ij}^{(g)} = |\eta_i^{(g)} - \eta_j^{(g)}|$
As we have the full distribution of every Δ_{ij}, we can compute $P(\Delta_{ij} > c|y)$, and take this probability as our Fuzzy BMV for the segment ij. Indeed, the availability of the posterior distribution provides another benefit: a way to directly assess the uncertainty in our Fuzzy

BMVs. The Monte Carlo estimate of $P(\Delta_{ij} > c|y)$ is derived as:

$$\hat{p_{ij}} = P(\Delta_{ij} > c|y) = \frac{\#\Delta_{ij}^{(g)} > c}{G} \tag{6}$$

This is nothing but a binomial proportion, where its components are independent.
The Gibbs samples Δ_{ij} are not independent in general, as they arise from a Markov chain, but it is possible to make them approximately so simply by subsampling retaining only every M^{th} sample. Note that this subsampling does not remove the spatial dependence among the Δ_{ij}. This approach makes use of the CAR model. However, the CAR model smooths across all geographical neighbors and can lead to over-smoothing and subsequent underestimation of several BLV. Lu et al. (2007); Ma et al. (2010) proposed adjacency matrix within a hierarchy. Rather than thresholding BLVs, Lu et al. (2007) assume the given areal boundaries in the Markov random fields (MRF) are random, Bernoulli variables, modeled using logistic regression in order to implement the wombling.
The hierarchical model-based approach addressed the estimate of the adjacency matrix inside a hierarchical framework utilizing priors on the edges (Lu and Carlin, 2005; Lu et al., 2007; Ma et al., 2010). Inference from these models, on the other hand, is typically highly sensitive to priors' specifications on certain parameters. Li et al. (2015) proposed a class of more flexible and robust nonparametric Bayesian hierarchical models to address this issue. The specification of the adjacent matrix W, which governs spatial smoothing, varies between these models. They require complex MCMC model composition that is computationally costly, particularly for big maps.
Li et al. (2015) investigated another approach: Bayesian hypothesis testing and adjusting multiple tests using forms discovery rate (FDR). But the model is still computationally intensive and requires benchmark.
Another approach in the hierarchical framework is the use of prior on the adjacency relationship (Ma et al., 2010). The issue with this is the prior information. Continuous priors for the ϕ_i do not work as they render $p(\phi_i = \phi_j|i \sim j) = 0$. The Dirichlet process (Ferguson, 1973) comes as a choice to model the spatial effect on discrete realization: nonparametric Bayesian. Areal information is incorporated into the stick-breaking weights (Areally-Referenced Stick-Breaking process (ARSB)) and a copula-type formulation in the technique.

Model-based wombling using dissimilarity metric

The Crisp and Fuzzy wombling introduces some levels of subjectivity. In fact, by defining a threshold in the Crisp wombling or a cutting point in the Fuzzy wombling, we are controlling somehow the number of boundary segments to be detected. These approaches have been

also criticized by Jacquez et al. (2000) who think that by specifying a threshold or cutting point, the investigator is choosing the number of boundaries that are identified even though this is unknown and the goal of the analysis. A wombling technique is incorporated in Lee and Mitchell (2012). The approach includes a dissimilarity metric in the Leroux model. Details about these models are presented above in the locally smooth CAR model of (Lee and Mitchell, 2012).

Residual based-wombling

Although BLVs based on expected values η_i are one method of exploring boundary probabilities, calculating BLVs using spatial random effects ϕ_i may be more illuminating. The ϕ_i can be thought of as spatial residuals (Fitzpatrick et al., 2010). High probability residual-based boundaries designate regions that differ in their unmodeled heterogeneity, highlighting boundaries that are not explained by the covariates.

In contrast, if a map of residual-based boundaries contains few barriers, the covariates explain (or are connected with factors that explain) the identified boundaries. Close analysis of boundary probabilities based on spatial residuals could be particularly beneficial in ecological and epidemiological research aimed at elucidating the factors determining range edges and how these vary across space (Fitzpatrick et al., 2010).

3 Part II: Wombling of Spatial COVID-19 incidence

3.1 Data description

Different datasets have been used in this project: the confirmed cases of COVID-19 by date and municipality; Belgium population data at the municipality level in 2020; the Shapefile of Belgium at the municipality level; and deprivation scores at the municipality level in 2011. The confirmed case of COVID-19 data was downloaded from the Belgian Institute for Health (Sciensano) whereas the three remaining data were downloaded from the STATBEL website.

Table 1 provides a general description of the datasets, including the sample size, the number and type of major variables, the existence of missing values, and the data format. Let us emphasize that the datasets contain no missing values. However, when the confirmed number of cases is less than 5, a limit of detection was reported. These observations were replaced with a random value between 0 and 4. We retrieved the total number of confirmed cases of COVID-19 in each municipality in Belgium for the different waves of the pandemic from these datasets. The first wave extends from March 1, 2020, to June 1, 2020; the second wave extends from September 1, 2020, to December 31, 2020; and the third wave extends from September 1, 2021, to December 31, 2021.

To correct for the differing populations of the municipalities, we used the Incidence Rate per Thousand (IR1000), which is calculated as $IR1000 = \frac{1000*y}{n}$, where y is the number of confirmed cases in the municipality and n is the population of the municipality as a whole.

Table 1: Data description

Data	Samples size	Number	Type of variables	Missing values	Format
Confirmed cases by date and municipality	419413	7	Categorical (NIS5: Code of the municipality, TX_DESCR_NL: Name in Netherlands, TX_ADM_DSTR_DESCR_NL: Arrondissement, PROVINCE, REGION) Numeric (CASES: number of confirmed cases) , Date	Yes ("<5")	csv
Administered vaccines by week, municipality, age and dose	1048575	5	Categorical (YEAR_WEEK: week of the year, NIS5: Code of the municipality, AGEGROUP: age class, DOSE: Type of vaccin) Numeric (CUMUL: cummulative number of vaccin administrated)	Yes ("<10")	csv
Belgium's Population data at municipality level	581	4	Categorical (CD_REFNIS: Code of the municipality, TX_DESCR_NL : Name in Netherlands,7 TX_PROV_DESCR_NL: Province) Numeric (pop: total population per municipality)	no	txt
Shapefile of Belgium at municipality level	-	-	-	-	shp
Deprivation scores	590	7	deprivation domains: Income; Employment; Education ; Housing ; Health ; crime and the total deprivation; overall deprivation	no	csv

3.2 Methodology

3.2.1 Exploratory data analysis

To examine the relevance of spatial autocorrelation, we compute Moran's I statistic (Moran 1950) and perform a permutation test. The permutation test, which uses the moran.mc()

function from the *spdep* package (Bivand et al., 2015), has a null hypothesis of no spatial autocorrelation and an alternative hypothesis of positive spatial autocorrelation. Moran's I statistic was used as an explanatory measure to test for spatial autocorrelation.

3.2.2 Traditional or algorithm-based wombling

The incidence rate per 1000 was used to define a boundary likelihood value (BLV) and the boundary membership value (BMV) for two neighboring municipalities (i, j). For Crisp wombling, three threshold values have been used to compute the BMV and correspond to the 1st, 2nd, and 3rd quantiles of the raw incidence rate per thousand.
Three different cutting points (50%, 75%, and 90%) were defined based on boundary membership value to assign a segment to a boundary membership. The number of segments and municipalities in the boundary for each wombling technique and each threshold/cutting point is summarized in a table. We made use of maps to display the raw incidence rate per 1000 as well as maps showing the segments and municipalities at the boundary. The municipalities at the boundary were identified using the approach proposed by (Legewie, 2018). In fact, the boundary likelihood value or boundary value refers to a pair of adjacent regions represented by the borderline segment between the two municipalities. This issue was addressed by defining the boundary value for a municipality as the maximum boundary value between the focal municipality and its neighbors.

3.2.3 Spatial areal modeling of COVID-19 incidence with conditional autoregressive priors

Before going to the model-based wombling, we first investigated which globally and locally smooth models fitted well the incidence rate per 1000 for the different waves. Details about these models are presented above. Only the Independence model, Besag-York-Mollié (BYM) model (Besag et al., 1991), and Leroux model (Leroux et al., 2000) were considered in the analysis regarding the globally smooth models.
All models were fitted in a Bayesian setting using Markov Chain Monte Carlo (MCMC) simulation in the package CARBayes (Lee, 2017). A non-informative prior was used for the random effect. These choices are designed to be vague enough to allow the data to dominate the determination of the posterior ($G(0.1, 0.1)$ prior for τ). Three different chains were used to draw inferences about the parameters. For each chain, we used 300,000 iterations with the initial 100,000 iterations discarded (i.e., burn-in part). A thinning of 20 was specified. The level of thinning was applied to the MCMC samples to reduce their

temporal autocorrelation. The inference of the model was based then on 10,000 post-burn-in and thinned MCMC samples. To assess the convergence of our chains, we performed a visual inspection of the trace plots and the autocorrelation plots and a statistical test such as the Gelman-Rubin diagnostic test (Gelman et al., 2014).
The fitted values were extracted from the selected model and the model-based incidence rate per 1000 is calculated. A model-based map was used to display the model-based IR1000.

3.2.4 Bayesian hierarchical model based wombling

The model-based wombling for each wave was based on the Bayesian real model identified in the previous section. Details about the procedure are presented above. The same threshold values and cutting points respectively for the Crisp and Fuzzy wombling were considered as in the case of algorithm-based wombling.

3.2.5 Model-based wombling using dissimilarity metric

To define the dissimilarity metric, we use the average deprivation score as a covariate. This index incorporates data from six deprivation categories, measuring several sorts of deprivation aspects such as income, employment, education, housing, health, and crime. These domains' building blocks are indicators, which are either rates or proportions of the population in a certain statistical sector experiencing some form of deprivation. For example, the rate of drug-related crime, the standardized suicide rate, or the proportion of working-age people who are unemployed.
The wombling using dissimilarity metric was conducted in the package CARBayes (Lee, 2017) using the function S.CARdissimilarity().
Note that, a residual wombling was also conducted by extraction from the fitted models, the residual values, and applying a Fuzzy wombling on residual incidence rate per 1000.

3.2.6 Implementation of wombling techniques

Very few statistical packages exist for the implementation of wombling techniques. The first software implemented was BoundarySeer which is commercial software. BoundarySeer only implemented algorithmic wombling techniques. However, to recognize the inherent variability and spatial association in the data, (Lu et al., 2007) was the first to compute the boundary likelihood value (BLV) R language. They later expanded the approach to Bayesian context specifically using Bayesian hierarchical in the estimation of the fitted value of the outcome of interest. The Bayesian modeling process was performed in WinBugs. Many authors conducted later the Bayesian models in another statistical package

like Openbugs and more recently a package dedicated to CAR prior modeling (CARBayes). Some authors even expand the models to the Bayesian nonparametric model. But the main problem is that all these authors wrote their own function in R language to perform a wombling. There is no R or other language package for wombling. Our investigation with these authors allowed us to understand that such packages will soon be available in R.
We have used in this project the R functions that these authors have already written as a basis for the implementation of our analysis. It is Joscha Legewie (Algorithmic areal wombling) from Harvard University and Sudipto Banerjee (Bayesian hierarchical areal wombling) from the University of California, Los Angeles.
Our contribution consisted first in extending the code of Lewigi which only allowed to identify of segments in boundaries to the identification of municipalities in boundaries. Second, to replace the nonparametric hierarchical Bayesian implemented by Sudipto Banerjee in his code, by the Bayesian hierarchical model with CAR prior in the CARBayes package (Lee, 2017).
All the analyses have been implemented in R statistical software (R Core Team, 2022).

3.3 Results

3.3.1 Exploratory data analysis

Figure 2 presents the exploration of the spatial trend in the COVID-19 incidence for the three different waves. The histograms of the distribution of Moran's I value under the hypothesis of independence (Figure 2.a) show that we do have enough evidence again the hypothesis of spatial independence. So, there seems to be a spatial correlation in the IR1000 across municipalities in Belgium and for all waves. The intensity of this spatial correlation is stronger during waves 2 and 3 compared to wave 1 (Figure 2.b).

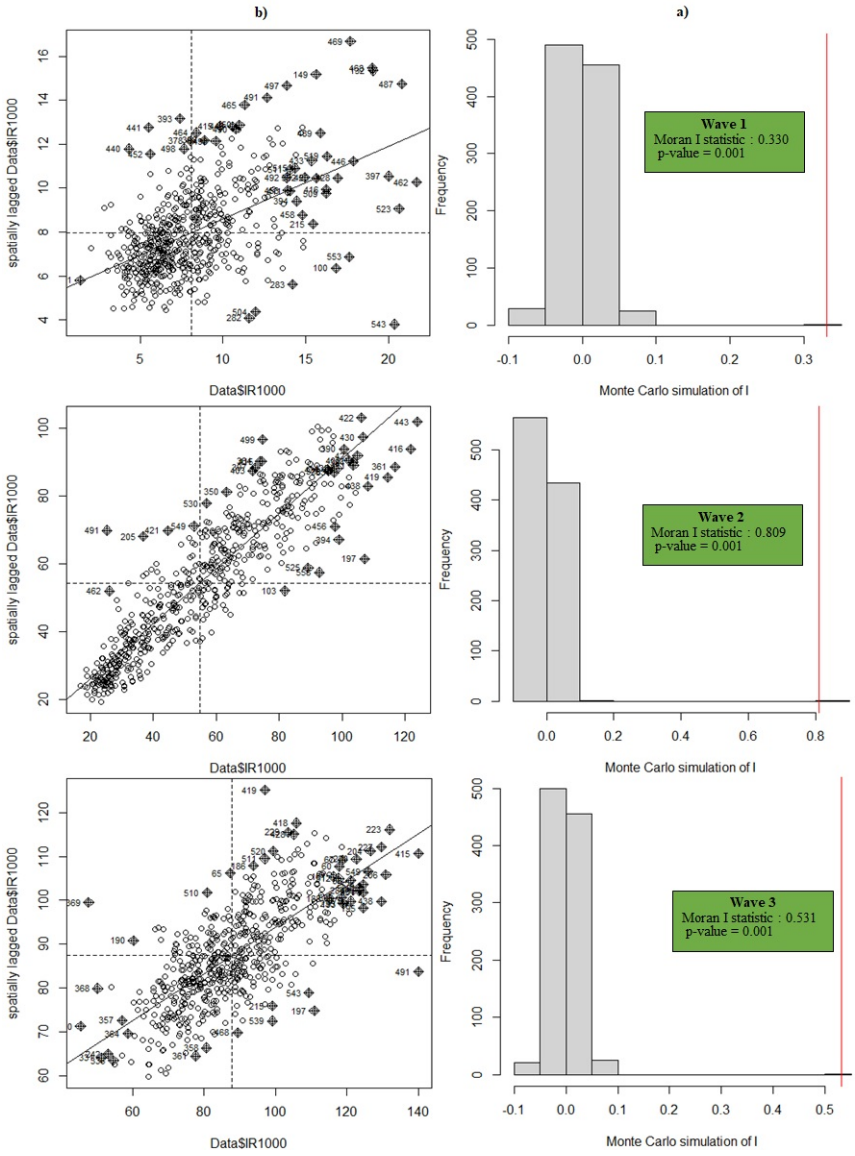

Figure 2: Exploratory data analysis

3.3.2 Traditional or Algorithmic-based wombling

Table 2 presents the number and the percentage of boundaries identified in the Belgian COVID-19 data across the three waves of the pandemic using algorithm-based wombling techniques (Crisp and Fuzzy wombling). From this table, the number of boundaries (segments or areas) decreases with the increasing value of the threshold or the cutting point. Considering the median value of the incidence rate per 1000 as the threshold, we identified 55 (3.38%), 12(0.74%), and 0(0%) boundaries respectively for the 1^{st}, 2^{nd}, and 3^{rd} wave

for the Crips wombling. During wave 3, the incidence rate was high everywhere, so the difference in incidence rate between two neighboring municipalities is very small compared to the threshold we have defined to identify boundaries. The Crisp wombling has therefore not identified any boundaries.
Likewise, by considering a cutting point of 50% in the partial boundary membership for the Fuzzy wombling, we identified respectively 41 (2.52%), 42 (2.77%), and 48 (2.95%) for the 1^{st}, 2^{nd}, and 3^{rd} wave.

Table 2: Algorithm-based wombling

Crisp wombling			Fuzzy wombling		
Threshold	Lines	Area	Threshold	Lines	Area
Wave 1					
6	138 (8.49%)	171 (29.43%)	50%	41 (2.52%)	56 (9.64%)
8	55 (3.38%)	72 (12.39%)	75%	6 (0.37%)	9 (1.55%)
10	31 (1.91%)	43 (7.40%)	90%	2 (0.12%)	4 (0.69%)
Wave 2					
34	51 (3.14%)	63 (10.84%)	50%	45 (2.77%)	57 (9.81%)
55	12 (0.74%)	14 (2.41%)	75%	14 (0.86%)	17 (2.92%)
72	2 (0.12%)	4 (0.69%)	90%	6 (0.37%)	10 (1.72%)
Wave 3					
77	0 (0%)	0 (0%)	50%	48 (2.95%)	75 (12.91%)
87	0 (0%)	0 (0%)	75%	6 (0.36%)	9 (1.55%)
97	0 (0%)	0 (0%)	90%	2 (0.12%)	4 (0.69%)

3.3.3 Spatial areal modeling with conditional autoregressive priors

Globally smooth model

Table 3 provides the global CAR model selection for the three waves. It helps to identify which conditional autoregressive model fits the data well. From this table, it appears that the Leroux model fits better the data for waves 1 and 3 whereas the BYM model is more appropriate for wave 2 data.

Table 3: Global CAR model selection

Model	DIC	p.d	WAIC	p.w	LMPL	loglikelihood
			Wave 1			
Independence	4848.714	525.836	4736.41	298.5741	-2665.78	-1898.521
Leroux	4822.331	501.7464	4721.913	291.1461	-2593.51	-1909.419
BYM	4830.063	503.8506	4732.752	294.8034	-2613.73	-1911.181
			Wave 2			
Independence	6275.981	725.2565	7464.45	1099.49	-8066.72	-2412.734
Leroux	5964.863	556.2432	5840.371	307.3963	-3674.03	-2426.188
BYM	5955.859	552.8552	5808.829	292.7127	-3192.99	-2425.074
			Wave 3			
Independence	6378.041	604.9958	6330.171	382.4983	-3906.65	-2584.025
Leroux	6281.526	544.0917	6161.516	304.5838	-3474.86	-2596.671
BYM	6281.096	546.1334	6154.821	301.8702	-3404.99	-2594.415

DIC: Deviance Information Criterion - $p.d$: number of effective parameters - $WAIC$: Watanabe-Akaike Information Criterion - $p.w$: number of effective parameters - $LMPL$: Log Marginal Predictive Likelihood

Table 4 provides the summary of the best global CAR models for the three different waves. The results show a significant spatial dependence (ρ) in the IR1000 for all the waves. This can also be seen from the model-based maps (Figure 3) where there is no regular distribution of the incidence across the country. Some municipalities are more affected by the pandemic than others. Specifically, during the first wave of the pandemic, the Eastern part of the country (province of Liege and Limburg) was more affected by the disease compared to other provinces. During the second wave, the smoothed incidence rate split the country into two clusters: The South is marked by a high incidence of the disease (shaded orange and red) compared to the North (shaded yellow). However, during the third wave, the disease is highly prevalent throughout almost the whole country. Only part of the western zone of the country seems somewhat spared (province of Hainaut).

Table 4: Summary of global CAR models

Coef	Mean	2.5% CI	97.5% CI	n.effective	Geweke.diag
			Wave 1		
Intercept	0.070	0.058	0.082	9528.7	0.8
τ^2	0.302	0.243	0.367	8678.4	-0.5
ρ	0.676	0.495	0.861	8418.6	-0.4
			Wave 2		
Intercept	-0.004	-0.011	0.002	931.4	-0.6
τ^2	0.168	0.149	0.189	9555	0.4
ρ	0.993	0.978	0.999	9476.1	-0.6
			Wave 3		
Intercept	0.047	0.042	0.053	1439	0.1
τ^2	0.065	0.057	0.074	10000	0.3
ρ	0.926	0.827	0.990	9828.2	0.6

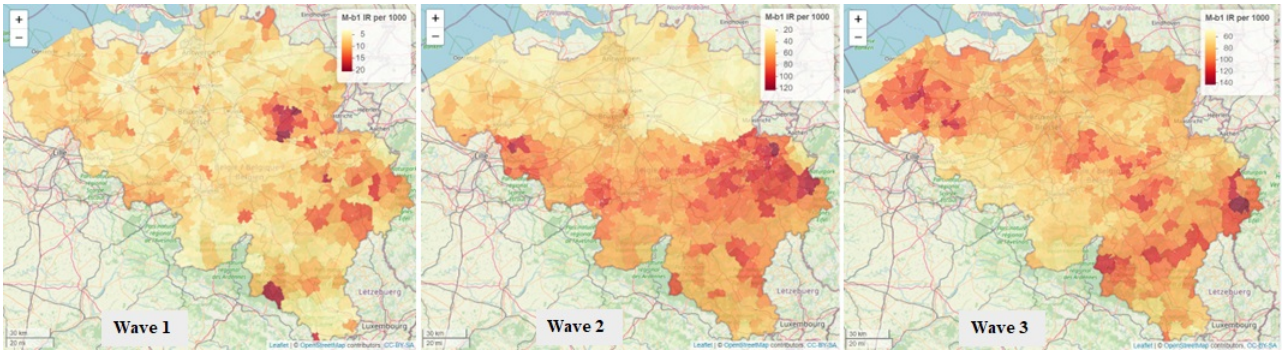

Figure 3: Global smooth model-based maps

Locally smooth model

Table 5 provides the locally smooth CAR model selection for the three waves. This comparison is made to select the best localized conditional autoregressive model that fits the data well for the different waves. From this table, it appears that the localized smooth CAR model of (Lee and Mitchell, 2012) fits better the data for the first wave whereas the localized smooth CAR model of (Lee and Sarran, 2015) turns out to be the best for waves 2 and 3.

Table 5: Locally CAR model selection

Model	DIC	p.d	WAIC	p.w	LMPL	loglikelihood
			Wave 1			
Lee and Mitchell (2012)	4632.861	386.838	4541.188	219.330	-2407.862	-1929.592
Lee and Sarran (2015)	4815.360	489.886	4746.471	302.207	-2964.221	-1917.794
			Wave 2			
Lee and Mitchell (2012)	5933.900	533.637	5804.262	291.294	-3208.829	-2433.313
Lee and Sarran (2015)	5565.292	154.628	5833.291	303.436	-3271.056	-2628.018
			Wave 3			
Lee and Mitchell (2012)	6172.192	478.889	6049.583	259.809	-3236.584	-2607.207
Lee and Sarran (2015)	5402.653	-334.710	6175.117	312.121	-3537.865	-3036.037

DIC: Deviance Information Criterion - $p.d$: number of effective parameters - $WAIC$: Watanabe-Akaike Information Criterion - $p.w$: number of effective parameters - $LMPL$: Log Marginal Predictive Likelihood

Figure A.2.1.B (in the appendix) presents the convergence of the Markov chains for the best-localized CAR model for each wave. The plot of the samples for the regression parameter for each Wave is shown in Figures a, b, and c respectively, and shows good mixing between and convergence of the chains, as they all have very similar means. Also, most of the values of the potential scale reduction (Gelman et al., 2014), are all below 1.1 suggestive of convergence.

Table 6 provides the summary of the best local CAR models for the different waves. The results show a significant spatial dependence in the IR1000 for all the waves. This can also be seen from the model-based maps (Figure 4) where there is no regular distribution of the incidence rate across the country. Some municipalities are more affected than others. During the first wave of the pandemic, the Eastern part of the country (province of Liege and Limburg) was more affected compared to other provinces. During the second wave, the smoothed incidence rate split the country into two clusters: The south is marked by a high incidence of the disease (shaded orange and red) compared to the north (shaded yellow). However, during the third wave, the disease is highly prevalent throughout almost the whole country. Only part of the western zone of the country seems somewhat spared (province of Hainaut). The smoothed incidence rate from the localized smooth CAR models is quite similar to the one from the globally smooth CAR models even though the globally smoothed incidence seems a little bit bigger.

Table 6: Summary of locally smooth models

Coef	Mean	2.5% CI	97.5% CI	n.effective	Geweke.diag	alpha.min
Wave 1						
Intercept	0.067	0.054	0.080	3944.1	1	-
τ^2	0.057	0.049	0.067	9058.8	-1.2	-
Z.Deprivation	1.909	1.870	1.935	5205.1	1.1	0.384
Wave 2						
λ_1	-0.556	-0.565	-0.548	4740.000	0.4	-
λ_2	0.113	0.098	0.128	55.800	1	-
λ_3	0.489	0.473	0.504	61.400	1.2	-
τ^2	0.148	0.130	0.169	1575.900	-0.2	-
δ	1.011	1.000	1.040	9361.900	0.3	-
Wave3						
λ_1	-0.169	-0.190	-0.146	107.100	1.1	-
λ_2	0.045	0.035	0.055	120.800	-1.2	-
λ_3	0.258	0.232	0.281	119.600	-1.1	-
τ^2	0.037	0.031	0.043	212.900	-0.8	-
δ	1.602	1.284	1.895	35.200	-1.2	-

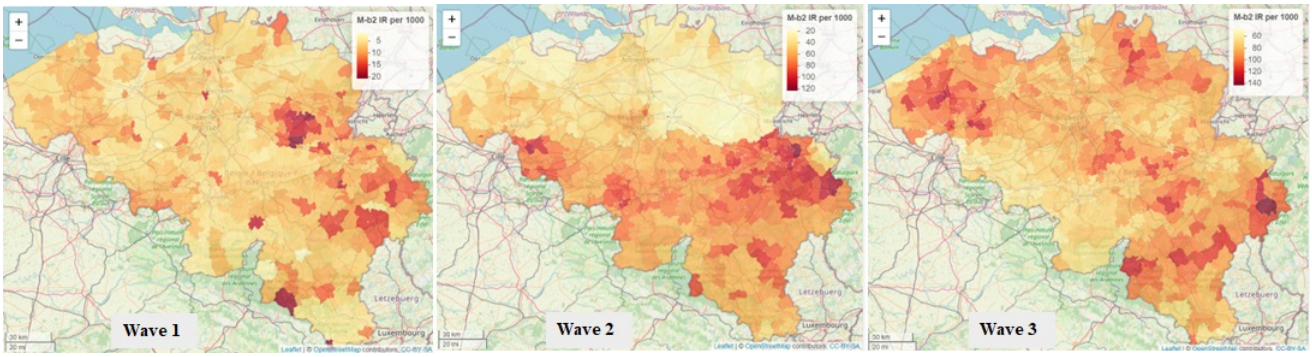

Figure 4: Locally smooth model-based maps

3.3.4 Bayesian hierarchical model based wombling

Wave 1

Table 7 presents a summary of the number of boundaries identified in the map based on the globally and locally smooth models for wave 1. From this table, we can notice that the number of boundaries decreases when we increase the value of the threshold for the case of Crisp wombling and the cutting point for the case of Fuzzy wombling. For the globally smooth model-based wombling, we identify 30 (1.85%) boundaries with the Crisp wombling using the median value of the IR1000 as threshold whereas the locally smooth model-based

identified 49 (3.01%). For the Fuzzy wombling and a cutting point of 50%, we identified 48 (2.95%) boundaries with the globally smooth model and 47 (2.89%) with the locally smooth model. The Fuzzy wombling yields approximately the same number of boundaries for both modeling approaches whereas the Crisp wombling revealed a big difference in the number of identified boundaries for the different approaches.

Figure 5 presents the locations of these boundaries on the map. The figure also makes a comparison of the different wombling approaches: algorithm-based, and model-based (global and locally smooth models). The algorithm-based wombling has identified more boundaries than the model-based wombling for the Crisp wombling. The Fuzzy wombling presents approximately the same number of boundaries for approaches (algorithm-based model-based wombling).

The majority of the identified boundaries visually correspond to sizeable changes in the incidence rate, suggesting that the models have the power to distinguish between boundaries and non-boundaries. The notable boundaries are the demarcation between the low incidence rate (shaded yellow) municipalities in the Eastern part of Belgium (Liege and Limburg) and their neighboring municipalities with high incidence rates on both sides (shaded orange and red). The boundaries shown in these maps are not too close and this suggests that the spatial pattern in incidence rate is more complex than being partitioned into groups of non-overlapping areas of incidence rate of covid-19.

Table 7: model-based wombling wave 1

Crisp wombling			Fuzzy wombling		
Threshold	Lines	Area	Cutting point	Lines	Area
Globally smooth					
6	87 (5.35%)	112 (19.28%)	50%	48 (2.95%)	66 (11.36%)
8	30 (1.85%)	42 (7.23%)	75%	9 (0.55%)	15 (2.58%)
13	10 (0.62%)	19 (3.27%)	90%	3 (0.18%)	5 (0.86%)
Locally smooth					
6	135 (8.30%)	167 (28.74%)	50%	47 (2.89%)	63 (10.84%)
8	49 (3.01%)	67 (11.53%)	75%	8 (0.49%)	13 (2.24%)
13	24 (1.47%)	33 (5.68%)	90%	3 (0.18%)	5 (0.86%)

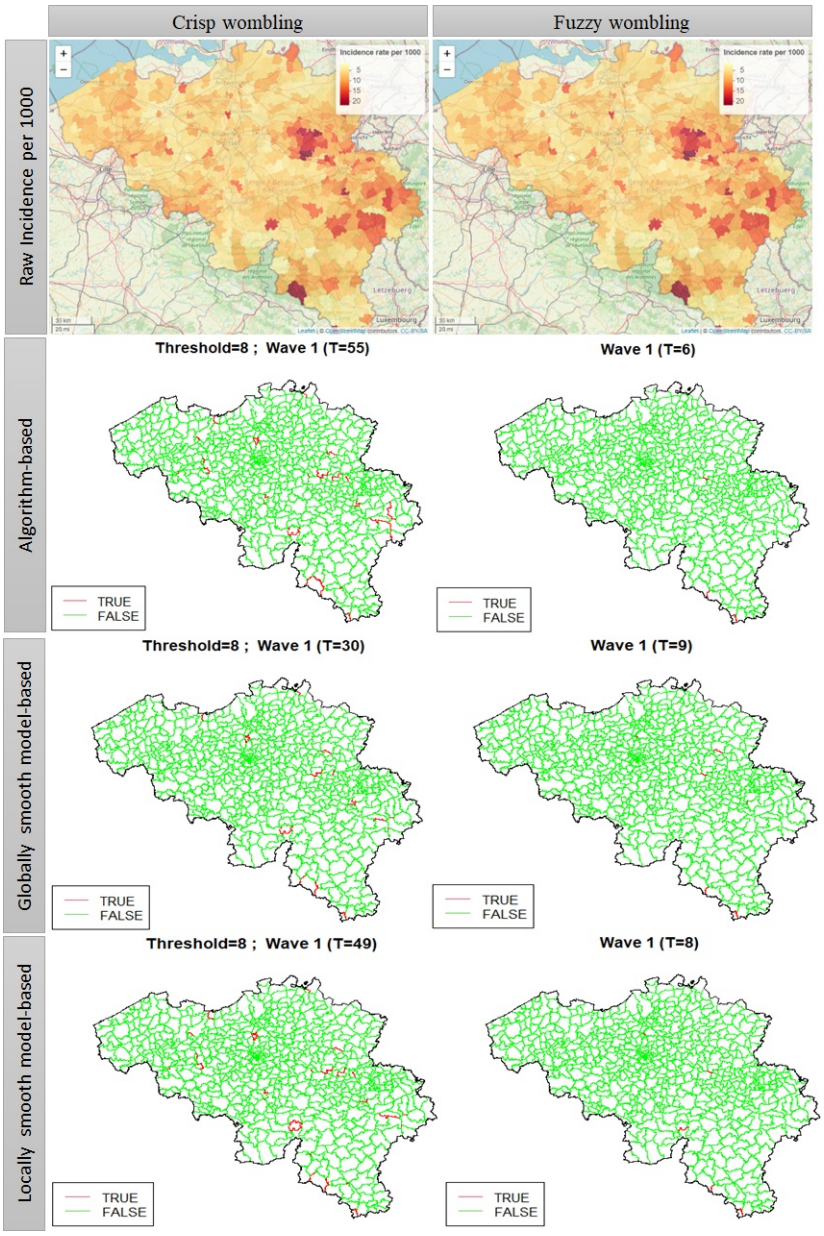

Figure 5: Comparison of wombling for Wave 1

Wave 2

Table 8 presents a summary of the number of boundaries identified in the map based on the globally and locally smooth models. From this table, we can notice that the number of boundaries decreases when we increase the value of the threshold for the case of Crisp wombling and the cutting point for the case of Fuzzy wombling. For the globally smooth model-based wombling, we identified 7 (0.43%) with the Crisp wombling using the median

value of the IR1000 as threshold whereas the locally smooth model-based wombling identified 8 (0.49%). For the Fuzzy wombling and a cutting point of 50%, we identified 9 (0.55%) boundaries with the globally smooth model and 10 (0.62%) for the locally smooth model. Crisp and Fuzzy wombling yield approximately the same number of boundaries for both modeling approaches (globally smooth vs locally smooth).

Figure 6 presents the locations of these boundaries on the maps. The figure also makes a comparison of the different wombling approaches: algorithm-based, and model-based (global and locally smooth models). The algorithm-based wombling has identified more boundaries than the model-based wombling for both Crisp and Fuzzy wombling. The identified boundaries visually correspond to sizeable changes in the incidence rate, suggesting that the model has the power to distinguish between boundaries and non-boundaries. The notable boundaries are the demarcation between the low incidence rate (shaded yellow) municipalities represented by the north part and municipalities with high incidence rates in the south (shaded orange and red). The boundaries shown in these maps are too closed and this suggests that the spatial pattern in incidence rate is not complex and can be partitioned into groups of non-overlapping areas of incidence rate.

Table 8: model-based wombling wave 2

Crisp wombling			Fuzzy wombling		
Threshold	Lines	Area	Cutting point	Lines	Area
Globally smooth					
34	41 (2.52%)	54 (3.32%)	50%	41 (2.52%)	54 (9.29%)
55	7 (0.43%)	10 (1.72%)	75%	9 (0.55%)	12 (2.07%)
72	0 (0%)	0 (0%)	90%	5 (0.86%)	8 (1.38%)
Locally smooth					
34	45 (2.77%)	57 (9.81%)	50%	44 (2.71%)	56 (9.63%)
55	8 (0.49%)	11 (1.89%)	75%	10 (0.62%)	14 (2.41%)
72	0 (0 %)	0 (0 %)	90%	4 (0.25%)	6 (1.03%)

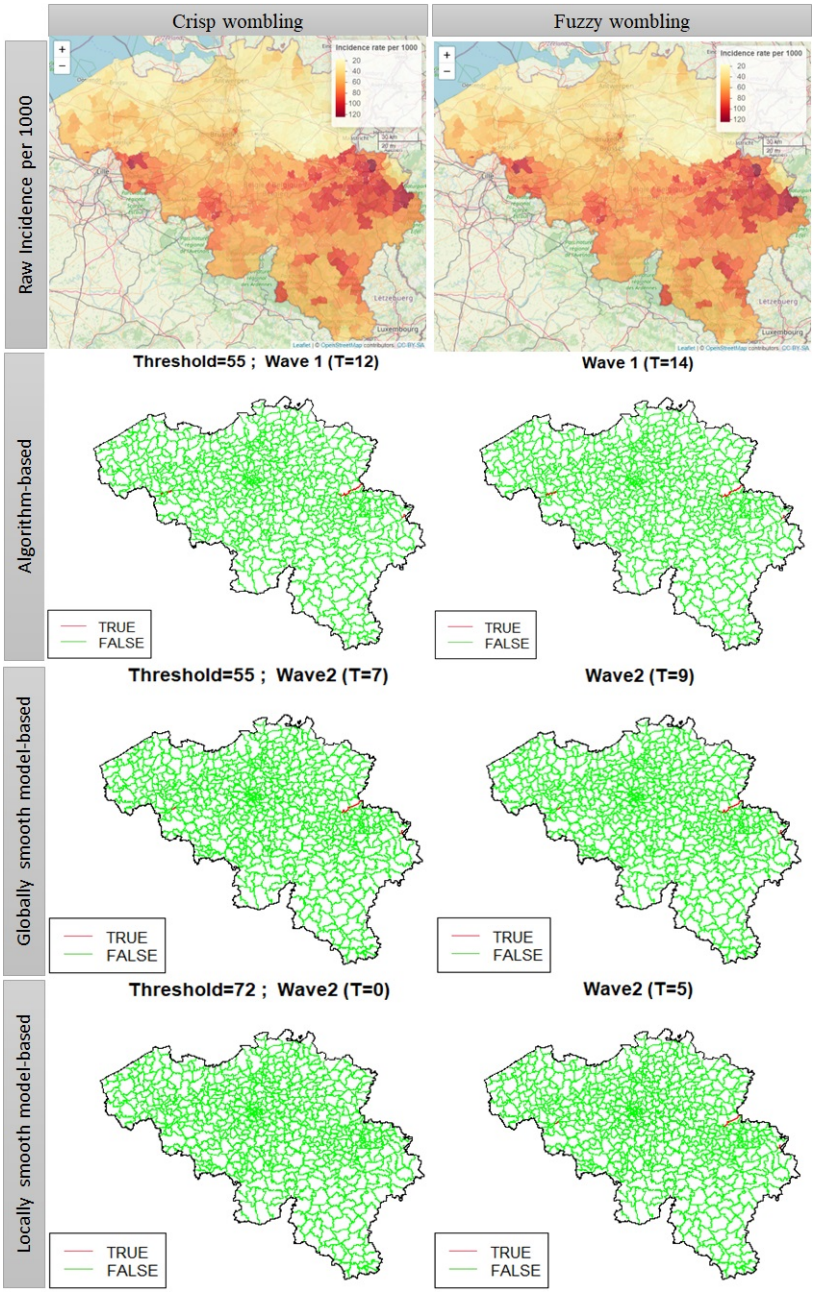

Figure 6: Comparison of wombling for Wave 2

Wave 3

Table 9 presents a summary of the number of boundaries identified in the map based on the globally and locally smooth models. From this table, we can notice that the number of boundaries decreases when we increase the value of the threshold for the case of Crisp

wombling and the cutting point for the case of Fuzzy wombling. For both globally and locally smooth model-based wombling, we have not identified any 0(0%) with the Crisp wombling using the median value of the IR1000 as the threshold. This is due to the BLV values. In fact, during wave 3, the incidence rate is high everywhere, so the difference in incidence rate between two neighboring municipalities is very small compared to the threshold we have defined to identify boundaries. The Crisp wombling has therefore not identified any boundaries.

For the Fuzzy wombling and a cutting point of 50%, we identified 85 (5.23%) boundaries with the global smooth model and 101 (6.21%) with the locally smooth model. Considering the Fuzzy wombling, the locally smooth model yields more boundaries compared to the global smooth model. Figure 7 presents the locations of these boundaries on the map. The figure also makes a comparison of the different wombling approaches: algorithm-based, and model-based (globally and locally smooth models). The Algorithm-based wombling has identified generally fewer boundaries compared to the model-based wombling for the Fuzzy wombling.

The majority of the identified boundaries visually correspond to sizeable changes in the incidence rate, suggesting that the model has the power to distinguish between boundaries and non-boundaries. The notable boundaries are the demarcation between the low incidence rate (shaded yellow) municipalities and their neighboring municipalities with high incidence rates on both sides (shaded orange and red).

Table 9: model-based wombling wave 3

Crisp wombling			Fuzzy wombling		
Threshold	Lines	Area	Threshold	Lines	Area
Global smooth					
77	0 (0%)	0 (0%)	50%	85 (5.23%)	122 (20.99%)
87	0 (0%)	0 (0%)	75%	15 (0.92%)	24 (4.13%)
97	0 (0%)	0 (0%)	90%	4 (0.25%)	6 (1.03%)
Local smooth					
77	0 (0%)	0 (0%)	50%	101 (6.21%)	137 (23.58%)
87	0 (0%)	0 (0%)	75%	17 (1.05%)	28 (4.81%)
97	0 (0%)	0 (0%)	90%	4 (0.25%)	6 (1.03%)

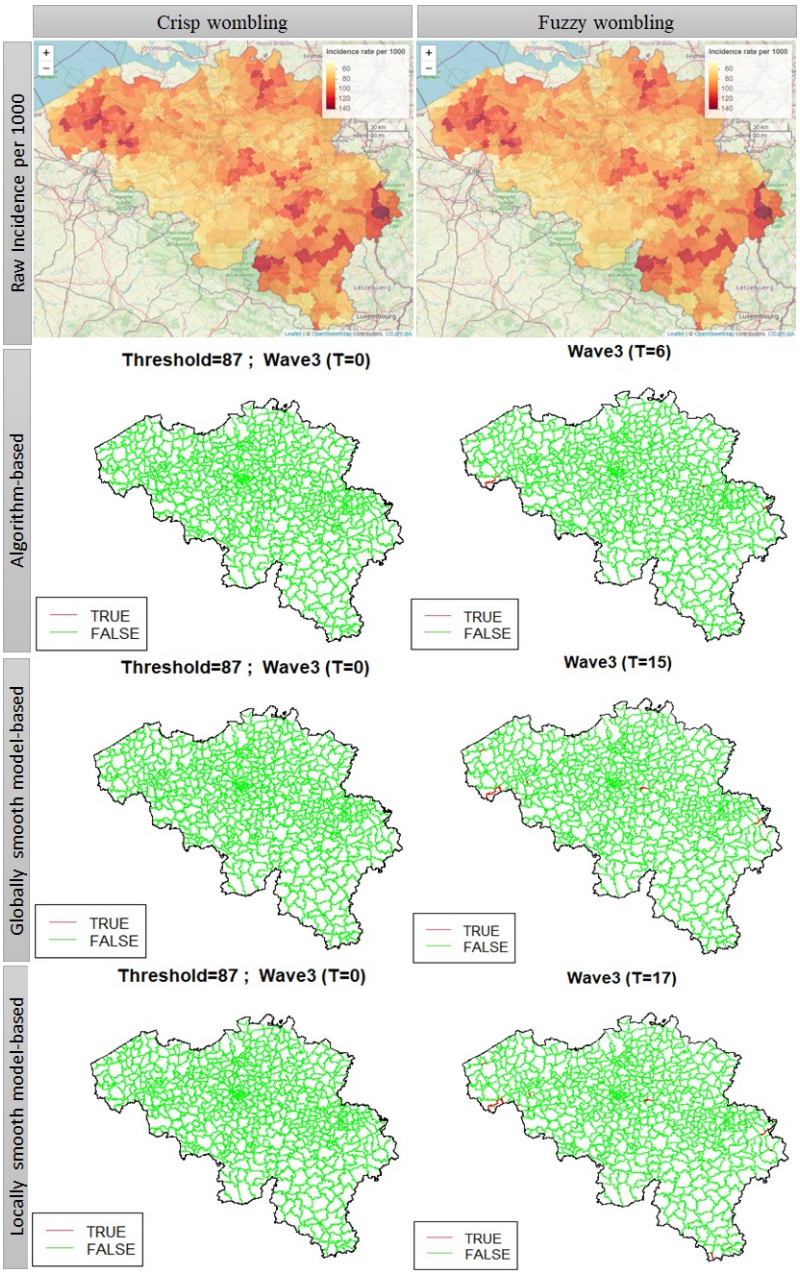

Figure 7: Comparison of wombling for Wave 3

3.3.5 Residual-based wombling

The following table (Table 10) presents a comparative study of the mean-based wombling (globally and locally smooth models) and the residual-based wombling. From the table, few boundaries are identified from the residual-based wombling compared to the mean-based

wombling. This applies to both globally and locally smooth models. This result suggests that some covariates may explain or correlate to the identified boundaries. The identified boundaries for both approaches are displayed in the maps in Figure 8. Note that only Fuzzy wombling has been used here.

Table 10: Comparison of mean-based and residual-based wombling

Cutting point	Globally smooth				Locally smooth			
	Mean based		Residual based		Mean based		Residual based	
	Segment	Area	Segment	Area	Segment	Area	Segment	Area
				Wave1				
50%	48 (2.95%)	66 (11.36%)	26 (1.60%)	32 (5.51%)	47 (2.89%)	63 (10.84%)	25 (1.54%)	31 (5.34%)
75%	9 (0.55%)	15 (2.58%)	4 (0.25%)	6 (1.03%)	8 (0.49%)	13 (2.24%)	5 (0.31%)	7 (1.20%)
90%	3 (0.18%)	5 (0.86%)	3 (0.18%)	4 (0.69%)	3 (0.18%)	5 (0.86%)	2 (0.12%)	3 (0.52%)
				Wave2				
50%	41 (2.52%)	54 (9.29%)	42 (2.58%)	51 (8.78%)	44 (2.71%)	56 (9.63%)	67 (4.12%)	82 (14.11%)
75%	9 (0.55%)	12 (2.07%)	8 (0.49%)	10 (1.72%)	10 (0.62%)	14 (2.41%)	7 (0.43%)	10 (1.72%)
90%	5 (0.86%)	8 (1.38%)	2 (0.12%)	3 (0.52 %)	4 (0.25%)	6 (1.03%)	2 (0.12%)	4 (0.69%)
				Wave3				
50%	85 (5.23%)	102 (17.56%)	43 (2.64%)	51 (8.78%)	101 (6.21%)	137 (23.58%)	14 (0.86%)	15 (2.58%)
75%	15 (0.92%)	24 (4.13%)	10 (0.62%)	12 (2.07%)	17 (1.05%)	28 (4.81%)	14 (0.86%)	15 (2.58%)
90%	4 (0.25%)	6 (1.03%)	3 (0.18%)	4 (0.69%)	4 (0.25%)	6 (1.03%)	4 (0.25%)	5 (0.86%)

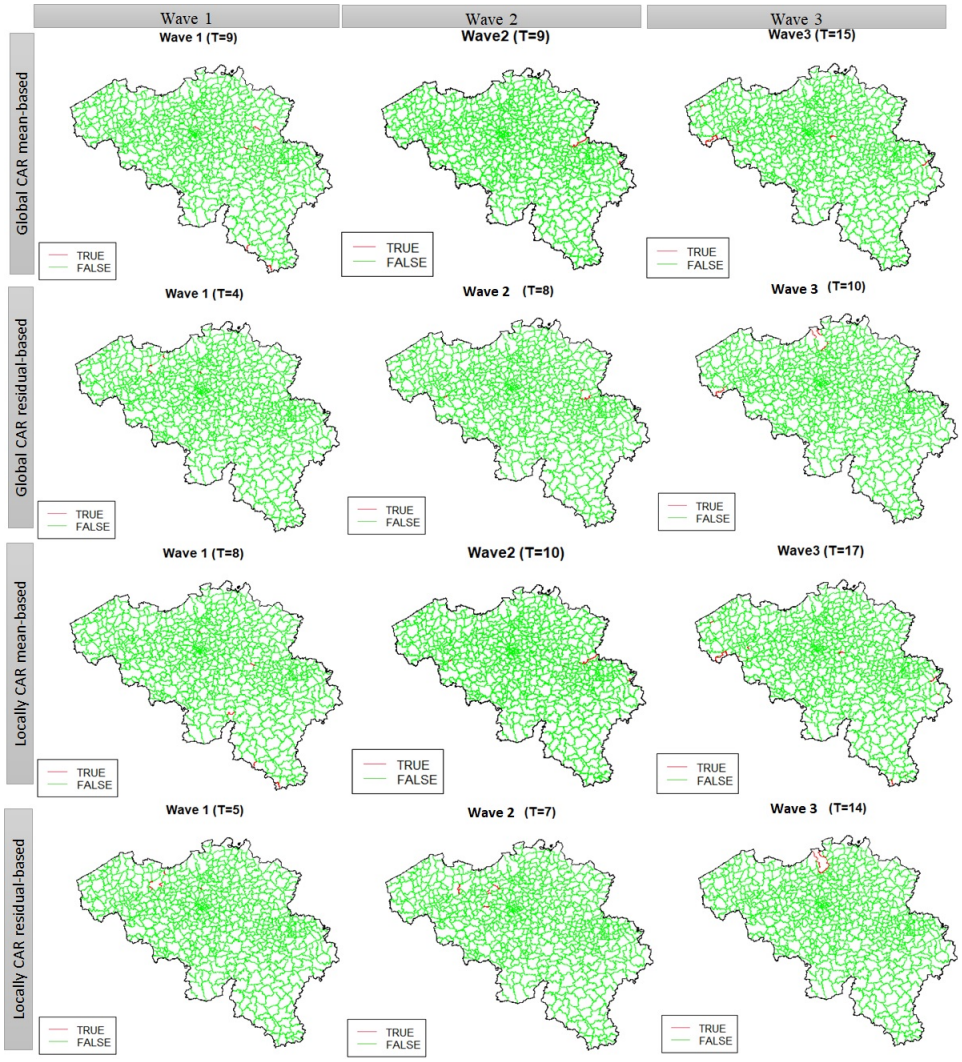

Figure 8: Mean-based vs Residual-based wombling

3.3.6 Model-based wombling using dissimilarity metric

Table 11 presents the summary of the localized spatial autocorrelation model proposed by (Lee and Mitchell, 2012) using the average deprivation score as a covariate to define the dissimilarity matrix. In this table, the value of alpha.min is the threshold value for the regression parameter α, below which the dissimilarity metric has no effect in identifying boundaries in the response (random effects) surface. For the first wave estimated alpha is 0.023, greater than the minimum alpha. So, the average deprivation score has an effect on the identification of boundaries during the first wave of the pandemic. The model has identified 38 (2.31%) boundaries in the map. For wave 2, the model presents a minimum

alpha greater than the estimated mean value of alpha (0.011). Therefore, the average deprivation score has no effect on the identification of boundaries during the second wave of the pandemic. There are no step changes identified in the random effect surface. Moreover, the model for wave 3 has identified step-changes identified in the random effect surface as the minimum alpha (0.0113) is below the estimated mean alpha (0.036). The maps in Figure 9 present the locations of these boundaries.

Table 11: Boundary detection using dissimilarity metric

Coef	Mean	2.5% CI	97.5% CI	n.effective	Geweke.diag	alpha.min
			Wave 1			
Intercept	0.071	0.056	0.086	6081.8	-0.7	-
τ^2	0.366	0.316	0.423	5376.4	1	-
Z.Deprivation	0.023	0.002	0.032	1227.6	-0.6	0.0123
no stepchange				1605		
stepchange				38		
			Wave 2			
Intercept	-0.009	-0.013	0.000	3.4	1.9	-
τ^2	0.167	0.147	0.186	100	1	-
Z.Deprivation	0.011	0.000	0.022	100	-0.3	0.0133
no stepchange				1643		
stepchange				0		
			Wave 3			
Intercept	0.004	0.004	0.004	0	-	-
τ^2	0.064	0.055	0.073	100	0	-
Z.Deprivation	0.036	0.034	0.041	37.6	-0.3	0.0113
no stepchange				1547		
stepchange				96		

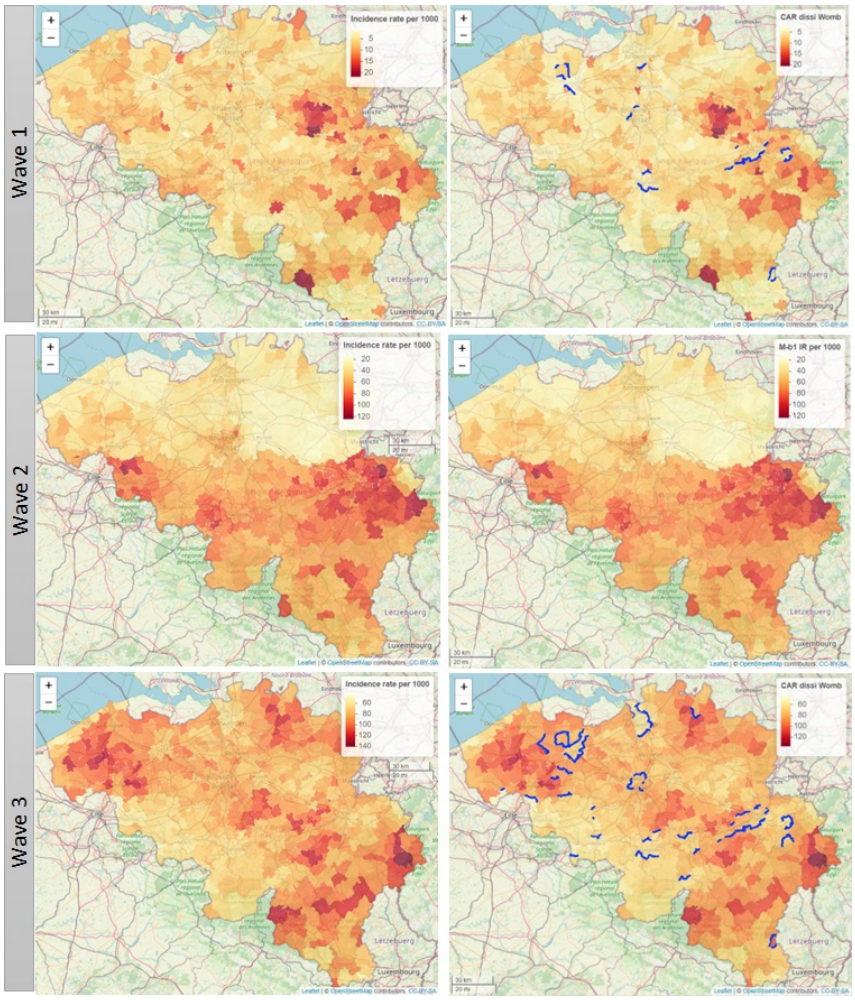

Figure 9: Map displaying the estimated incidence rate and the locations of the boundaries

Table 12 presents a comparison of the number of boundaries identified for different modeling approaches (mean-based, residual-based, and dissimilarity-based). It appears that the residual-based approach provides the lowest number of boundaries followed by the dissimilarity-based approach. The mean-based approach provides the highest number of boundaries generally. From this finding, we can say that some covariates are correlated to the identification of boundaries and the average deprivation score is not the perfect covariate to explain the spatial correlation in this boundary identification.

From the wombling approach using the dissimilarity metric (covariate = average deprivation score), the boundaries identified do not correspond exactly to the sizeable changes in the incidence rate. So, the model or the covariate used to define the dissimilarity metric has less power to distinguish between boundaries and non-boundaries. So, in order to test

Table 12: Comparison of different wombling approach

Wombling approach	Wave1	Wave2	Wave3
Globally CAR mean-based	48 (2.95%)	41 (2.52%)	85 (5.23%)
Globally CAR residual-based	26 (1.60%)	42 (2.58%)	43 (2.64%)
Locally CAR mean-based	47 (2.89%)	44 (2.71%)	101 (6.21%)
Locally CAR residual-based	25 (1.54%)	67 (4.12%)	14 (0.86%)
Dissimilarity-based	38 (2.34%)	0 (0 %)	96 (9.90%)

whether the covariate (average deprivation score) plays a role in the identification of the boundaries, we performed a sensitivity analysis by introducing the average deprivation score as a covariate in the models (globally smooth CAR models) and conducted mean-based and residual-based Fuzzy wombling. The result is presented in the next table (Table 13). Using deprivation score as a covariate in the model, the residual-based wombling yields a higher number of boundaries compared to the mean-based wombling. This indicates that boundaries designate regions that differ in their unmodeled heterogeneity, highlighting boundaries that are not explained by the covariate. So, the deprivation score is not really helpful in the identification of boundaries.

Table 13: Sensitivity analysis (average deprivation score in Globally smooth CAR model)

Cutting point	Mean based	Residual based
	Wave1	
50%	41 (2.52%)	226 (13.89%)
75%	6 (0.37)	97 (5.97%)
90%	2 (0.12%)	56 (3.44%)
	Wave2	
50%	45 (2.77%)	1021 (62.79%)
75%	14 (0.86)	804 (49.45%)
90%	6 (0.37%)	717 (44.09%)
	Wave3	
50%	17 (1.04%)	657 (40.41%)
75%	3 (0.18%)	376 (23.12%)
90%	2 (0.12%)	247 (15.22%)

4 Discussion

In this study, we reviewed the Bayesian hierarchical wombling techniques for areal data and apply some of the techniques to COVID-19 data in Belgium. We included 24 studies in the review: 19 in spatial areal wombling, 2 studies in spatiotemporal wombling, and studies in multivariate wombling. Mainly, 2 wombling techniques exist: Crisp wombling and Fuzzy wombling. The difference between these two approaches lies in the way of calculating the boundary membership value (BMV). The wombling can be algorithm-based or model-based. The Algorithm-based makes use of the raw health outcome in the computation of the boundary's likelihood value (BLV) while the model-based uses a Bayesian hierarchical model as background in the computation of the boundary's likelihood value. Different Bayesian hierarchical models exist to smooth the health outcome across the study areas: globally smooth CAR models and locally smooth CAR models. The globally smooth CAR models include independent, Besag, BYM, and Leroux models while locally smooth include (Lee and Mitchell, 2012) and (Lee and Sarran, 2015). The model-based wombling can use the residual (random effect) in the computation of the BLV. This approach helps to investigate the importance of the covariates in the boundaries detection. A more objective wombling approach is the use of dissimilarity metric to identify boundaries in a map. In this approach, instead of using a threshold or cutting point to identify boundaries, it uses a covariate to define a dissimilarity metric which is incorporated into the Leroux model and this covariate helps to identify boundaries in the map. After this synthesis of the literature relating to wombling techniques, we applied some of the techniques (spatial areal wombling) to covid-19 data for the three different waves of the pandemic in Belgium.

The result shows that the Crisp wombling for wave3 presented any boundaries compared to other waves. This could be due to the value of IR1000 in the municipalities during this wave. In fact, during wave 3, the incidence rate is high everywhere, so the difference in incidence rate between two neighboring municipalities is very small compared to the threshold we have defined to identify boundaries. The Crisp wombling has therefore not identified any boundaries. For all the waves, the Fuzzy wombling yields more consistent results as it does not rely on 0-1 to include segments into a boundary.

The globally smooth models revealed that the Leroux model was appropriate for incidence rate data for waves 1 and 3 while BYM is more appropriate for the incidence rate for wave 2. This result could be explained by the intensity of the spatial autocorrelation. As shown in the exploratory data analysis, wave 2 presented a higher spatial autocorrelation across municipalities (0.809) compared to wave 1 (0.330) and 3 (0.531). BYM model seems to be more suitable in the presence of high spatial correlation. This result is in line with Aswi

et al. (2020), who proved that the Leroux model performed the best in the presence of low autocorrelation.

The localized smoothing modeling approach shows that Lee and Mitchell's model is more suitable for wave 1 COVID-19 data while Lee and Sarran's model fits better wave 2 and 3 COVID-19 data. This result tells us that in the presence of weak spatial autocorrelation, the model of Lee and Mitchell performs better than Lee and Sarran. The smoothed incidence rate from the localized smooth CAR models is quite similar to the one from the globally smoothed CAR models even though the globally smoothed incidence seems a little bit more pronounced specifically for wave 2.

In wave 1, the locally smooth model has identified more boundaries compared to the globally smooth model for the Crisp wombling whereas the Fuzzy wombling provides approximately the same number of boundaries for both modeling approaches. However, the algorithm-based wombling yields more boundaries compared to the model-based. In fact, the hierarchical model corrects for uncertainty in the estimate values throughout the process rather than averaging this uncertainty out before the BLV is completed as the case in the algorithm-based wombling (Lu and Carlin, 2005). The difference between the globally and locally smooth in the Crisp wombling could be explained by the presence of sub-region spatial correlation. This presence of sub-region spatial correlation can also explain the spatial pattern in incidence rate which is more complex and cannot be partitioned into groups.

In wave 2, as in the case of wave 1, the algorithm-based wombling has identified more boundaries than the model-based wombling in wave 2. Again, this result can be explained by the fact that the modeling process corrects the uncertainty in the raw incidence rate. The difference in the number of boundaries between the globally and locally smooth is very minor here. This can be explained by the clustering observed in the map. The locally smoothing has been done mainly in two groups compared to one group for the globally smooth model, the reason why the number of boundaries is very close.

In wave 3, the locally smooth model yields more boundaries compared to the global smooth model. The same explanation is applied here. Unlike wave 1 and wave 2, the algorithm-based wombling has identified fewer boundaries compared to the model-based wombling for Fuzzy wombling. This shows that in case of a higher prevalence of the disease in the whole map, the algorithm-based can yield a lower number of boundaries compared to model-based wombling. This means that the smoothing process can also produce a lower incidence rate compared to the raw incidence in some locations.

Our results also suggest that the deprivation score is not a good covariate in the identification of boundaries in the COVID-19 incidence map.

5 Ethical thinking, societal relevance, and stakeholder awareness

The datasets involved in this project were originally obtained from the Belgian Institute for Health (Sciensano) and the Belgian statistical office (STATBEL). Confidentiality of the data was the main ethical standard related to this project, but as the data is an aggregate count over municipalities, the identity of the individual is already hidden, or the background of the individuals cannot be identified. So, no formal approval letter was obtained as confidentiality is no longer an issue.

The societal relevance of this project is to improve disease preventive and control decision-making by finding zones of significantly differing incidence or death. The results of this study can help all stakeholders (administrators of public health, decision-makers, and mayors) in the control of the pandemic to better redefine their strategies in controlling the spread of the disease and also an efficient use of material, financial and human resources. Indeed, having knowledge of the localities under boundaries can help to know which parts of the country need more intervention and thus define different intervention plans depending on whether the locality is under boundary or not. This could, for example, help to avoid the excessive use of resources in localities where the need is not so great.

The findings of this project have direct implications for many stakeholders. The most relevant stakeholders are administrators of public health and decision-makers who are directly involved in the management of diseases or health emergencies. As this project investigated the identification of the differences in adjacent municipalities and highlight those boundaries that have a high difference amongst neighboring municipalities, local authorities at the municipality levels like the mayors can also be considered as stakeholders. This project could assist them in the conception of new strategies to control the pandemic of COVID-19 or new/ future pandemics or epidemics. Indeed, the identification of boundaries can help them in designing an adequate disease management plan and also an efficient strategy for the distribution of available resources. All of this is relevant not only to the case of Belgium, but it could also be applied to other diseases other than Covid-19 or in a different setting, nation, or continent. This would be especially appealing in resource-constrained countries where financial resources are insufficient to cover or intervene across their entire territory. Such an approach could help them identify which areas require additional attention and thus manage health crises more effectively.

6 Conclusion and future work

This study presented a review of wombling techniques and applied some of the techniques to COVID-19 data in Belgium. Wombling helps to identify areas of rapid change on a map. Two main wombling techniques exist and can be applied directly to the outcome variable of interest (Algorithm-based wombling) or after accounting for uncertainty in the outcome via the modeling process (model-based wombling). The Bayesian hierarchical model, as well as the nonparametric Bayesian model, were developed in the literature to properly account for the variability in the outcome.

The spatial wombling of COVID-19 incidence in Belgium revealed the existence of boundaries during the different waves of the pandemic. The difference was more remarkable during wave 2, where the country was split into two regions: the North marked by a medium incidence and the South marked by a strong incidence. Globally algorithm-based wombling has identified more boundaries compared to model-based wombling. This can be affected by the distribution of outcomes in the general population. A uniformly or quasi-uniform distributed outcome in the study area can yield opposite behavior. It is the case of the wave 3 pandemic in Belgium where the model-based provided fewer boundaries compared to the algorithm-based. The residual wombling has shown that the identification of the boundaries may be correlated with some spatially oriented covariates. But the average deprivation score was not found as a useful covariate in the identification of boundaries in COVID-19 data. Indeed, the deprivation score is not a perfect risk factor for COVID-19. So, one of the limitations of this study is related to the absence of covariates at the municipality level in the data. It would be nice to have some potential covariates that could help to adjust the estimated incidence rate and therefore more accurate boundaries identification.

In this study, we only applied the univariate spatial wombling techniques. But as it has been seen during the pandemic that the introduction of vaccination had slowed down the spread of the disease, future works may look at the multivariate Areal Wombling for COVID-19 incidence and vaccination rate in Belgium. Also, our founding revealed that the intensity of the spatial correlation was not the same for the different waves. It would also be interesting for future studies to focus on the Spatiotemporal wombling of COVID-19 incidence. Future studies can also focus on how the intensity of the spatial autocorrelation could affect the identification of boundaries in a map.

References

Aswi, A., Cramb, S., Duncan, E., and Mengersen, K. (2020). Evaluating the impact of a small number of areas on spatial estimation. *International journal of health geographics*, 19(1):1–14.

Backer, J. A., Klinkenberg, D., and Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019-ncov) infections among travellers from wuhan, china, 20–28 january 2020. *Eurosurveillance*, 25(5):2000062.

Banerjee, S. and Gelfand, A. E. (2006). Bayesian wombling: Curvilinear gradient assessment under spatial process models. *Journal of the American Statistical Association*, 101(476):1487–1501.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43:1–20.

Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A., and Blanchet, G. (2015). Package 'spdep'. *The comprehensive R archive network*, 604:605.

Fatima, M., O'keefe, K. J., Wei, W., Arshad, S., and Gruebner, O. (2021). Geospatial analysis of covid-19: A scoping review. *International Journal of Environmental Research and Public Health*, 18(5):2336.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Fitzpatrick, M. C., Preisser, E. L., Porter, A., Elkinton, J., Waller, L. A., Carlin, B. P., and Ellison, A. M. (2010). Ecological boundary detection using bayesian areal wombling. *Ecology*, 91(12):3448–3455.

Gao, L., Banerjee, S., and Ritz, B. (2022). Spatial difference boundary detection for multiple outcomes using bayesian disease mapping. *Biostatistics*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). Bayesian data analysis (vol. 2).

Gilbert, M., Pullano, G., Pinotti, F., Valdano, E., Poletto, C., Boëlle, P.-Y., d'Ortenzio, E., Yazdanpanah, Y., Eholie, S. P., Altmann, M., et al. (2020). Preparedness and vulnerability of african countries against importations of covid-19: a modelling study. *The Lancet*, 395(10227):871–877.

Glass, G. E. (2000). Update: spatial aspects of epidemiology: the interface with medical geography. *Epidemiologic reviews*, 22(1):136–139.

Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D. S., et al. (2020). Clinical characteristics of 2019 novel coronavirus infection in china. *MedRxiv*.

Hodges, J. S., Carlin, B. P., and Fan, Q. (2003). On the precision of the conditionally autoregressive prior in spatial models. *Biometrics*, 59(2):317–322.

Jacquez, G. M., Maruca, S., and Fortin, M.-J. (2000). From fields to objects: a review of geographic boundary analysis. *Journal of Geographical Systems*, 2:221–241.

Kang, D., Choi, H., Kim, J.-H., and Choi, J. (2020). Spatial epidemic dynamics of the covid-19 outbreak in china. *International journal of infectious diseases*, 94:96–102.

Lee, D. (2017). Carbayes version 4.6: An r package for spatial areal unit modelling with conditional autoregressive priors. *University of Glasgow, Glasgow*.

Lee, D. and Mitchell, R. (2012). Boundary detection in disease mapping studies. *Biostatistics*, 13(3):415–426.

Lee, D. and Sarran, C. (2015). Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies. *Environmetrics*, 26(7):477–487.

Legewie, J. (2018). Living on the edge: neighborhood boundaries and the spatial dynamics of violent crime. *Demography*, 55(5):1957–1977.

Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191. Springer.

Li, P., Banerjee, S., Hanson, T. A., and McBean, A. M. (2015). Bayesian models for detecting difference boundaries in areal data. *Statistica Sinica*, 25(1):385.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., et al. (2020). Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England journal of medicine*.

Lu, H. and Carlin, B. P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, 37(3):265–285.

Lu, H., Reilly, C. S., Banerjee, S., and Carlin, B. P. (2007). Bayesian areal wombling via adjacency modeling. *Environmental and ecological statistics*, 14:433–452.

Ma, H., Carlin, B. P., and Banerjee, S. (2010). Hierarchical and joint site-edge methods for medicare hospice service region boundary analysis. *Biometrics*, 66(2):355–364.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery*, 88:105906.

Pfeiffer, D. U., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J., and Clements, A. C. (2008). *Spatial analysis in epidemiology*. OUP Oxford.

R Core Team (2022). *R: A Language and Environment for Statistical Computing. Version 4.2.2*. R Foundation for Statistical Computing.

Read, J. M., Bridgen, J. R., Cummings, D. A., Ho, A., and Jewell, C. P. (2021). Novel coronavirus 2019-ncov (covid-19): early estimation of epidemiological parameters and epidemic size estimates. *Philosophical Transactions of the Royal Society B*, 376(1829):20200265.

Waller, L. A. (2006). Hierarchical models for disease mapping. *Encyclopedia of Environmetrics*.

Womble, W. H. (1951). Differential systematics. *Science*, 114(2961):315–322.

Yang, Y., Lu, Q.-B., Liu, M.-J., Wang, Y.-X., Zhang, A.-R., Jalali, N., Dean, N. E., Longini, I., Halloran, M. E., Xu, B., et al. (2020). Epidemiological and clinical features of the 2019 novel coronavirus outbreak in china. *medrxiv*, pages 2020–02.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al. (2020). A novel coronavirus from patients with pneumonia in china, 2019. *New England journal of medicine*.

A Appendix

A.1 R Codes

```
#############################################################################
# Hasselt University - Belgium
# Master Thesis in Statistics and Data Science
#
#  Topic: Review of Bayesian hierarchical areal wombling techniques
#                         with application to Covid-19
#
# June, 1 , 2023
# Edmond SACLA AIDE
# edmond.saclaaide@student.uhasselt.be ; edmond.sacla95@gmail.com
#############################################################################


### load libraries
library(sf)
library(rgeos)
library(CARBayes)
library(dplyr)
library(plyr)
library(sp)
library(spdep)
library(MASS)
library(sp)
library(stargazer)
library(sf)
library(matrixStats)
library(tidyverse)
library(scales)
library(ggplot2)
library(tidyverse)
library(SpatialEpi)
library(dplyr)
library(sp)
library(tmap)
library(sf)
library(tmap)
library(leaflet)
library(knitr)
library(coda)
library(CARBayes)

### utility functions
source(file.path(PATH, "Algo_Areal_wombling.R"))
source(file.path(PATH, "Bayesian_Areal_Wombling.R"))
source(file.path(PATH, "00-utils.R"))
source(file.path(PATH, "posterior_blv.R"))# to compute the posterior blv
```

```
### Areal Wombling
#'
#' \code{areal_wombling} for algorithmic areal wombling (Lu and Carlin 2005: 268).
#'
#' By default, \code{censusr} downloads and recodes a selected set of variables.
#' These variables include 100-300 commonly used measures from the
#'
#' @param  sp Object of type \code{SpatialPolygonsDataFrame} (sf objects as converted)
#' @param  x Vector of variable names for which we want to calculate the boundary value.
#' @param threshold Threshold for the boundary membership value (BMV). If \code{threshold} is
#'   /code{NA} (the default), /code{areal_wombling} uses fuzzy wombling. If \code{threshold} is
#'   specified (any value between 0 and 1), /code{areal_wombling} uses crisp wombling
#'   using \code{threshold} to determine boundary membership.
#' @param dist Distance function. By default the absolute difference in the response variable.
#' @return     Object of class \code{SpatialLinesDataFrame} with SpatialLines
#' @export
areal_wombling <- function(sp, x, threshold = NA, dist = function(x) abs(x[1] - x[2])) {
  # check 'reshape2:::parse_formula' for p.white.cb + p.black.cb + p.hisp.cb + p.asian.cb ~ 1
  # Coerce simple feature geometries to corresponding Spatial* objects
  if (is(sp, "sf")) sp <- as(sp, "Spatial")
  # get borders as line segments in SpatialLinesDataFrame
  sl <- border_lines(sp)
  # boundary likelihood value (BLV) and boundary membership value (BMV) to data.frame
  dots_format <- function(s, suffix)
    s %>% setNames(paste0(x, suffix)) %>%
    as.list() %>% lapply(FUN = as.formula, env = environment())
  dots_blv <- dots_format(sprintf("~ dist(sp@data[['%s']][c(i, j)])[1]", x),
  suffix = "_blv")
  print(dots_blv)
  dots_bmv <- dots_format(sprintf("~ %s_blv/max(%s_blv, na.rm = TRUE)", x, x),
  suffix = "_bmv")
  if(!is.na(threshold))
dots_bmv <- dots_format(sprintf("~ %s_blv > %s", x, threshold), suffix = "_bmv")
  sl@data <- sl@data %>%
    dplyr::group_by(i, j) %>%
    dplyr::mutate_(.dots = dots_blv) %>%
    dplyr::ungroup() %>%
    dplyr::mutate_(.dots = dots_bmv) %>%
    as.data.frame()
  # return
  return(sl)
}


###' Bayesian Areal Wombling
#'
#' \code{areal_wombling_bayesian} for Bayesian areal wombling (Lu and Carlin 2005).
#'
#' By default, \code{censusr} downloads and recodes a selected set of variables.
#' These variables include 100-300 commonly used measures from the
#'
#' @param  formula  A formula for the covariate part of the model, using the same notation as for the \
```

```
#'   \code{y ~ 1} estimates a model without covariates.
#' @param  family One of either 'binomial', 'gaussian' or 'poisson', which respectively specify a binon
#' @param  sp Object of type \code{SpatialPolygonsDataFrame} (sf objects as converted)
#' @param  phi Conditional autoregressive (CAR) prior for the random effect. Options are 'leroux' for t
#' @param threshold Threshold for the boundary membership value (BMV). If \code{threshold} is
#'   /code{NA} (the default), /code{areal_wombling} uses fuzzy wombling. If \code{threshold} is
#'   specified (any value between 0 and 1), /code{areal_wombling} uses crisp wombling
#'   using \code{threshold} to determine boundary membership.
#' @param  \dots   Arguments passed to estimation command including \code{burnin}, \code{n.sample}, \co
#' @return  Object of class \code{carbayes} from package \code{CARbayes} with two additions:
#'   First, additional element \code{borders} of class \code{SpatialLinesDataFrame}, which includes
#'   all border lines and the postior median estimates for the boundary likelihood value and the bounda
#'   membership value (boundary probability if \code{threshold} is defined). Second, \code{sampes} incl
#'   additional elements for the McMC samples of the boundary likelihood value and the boundary member
#' @export
areal_wombling_bayesian <- function(formula, family, sp, phi = "",
                                    threshold = NA, E,pop, ...) {
  if (!(phi %in% c("INDEP","IAR","leroux","BYM","LOCA1", "LOCA2")))
    stop("Incorrect prior for the random effect.")
  # Coerce simple feature geometries to corresponding Spatial* objects
  if (is(sp, "sf")) sp <- as(sp, "Spatial")
  # get borders as line segments in SpatialLinesDataFrame
  sl <- border_lines(sp)
  # polygon adjacency matrix
  W.nb  <- spdep::poly2nb(sp, row.names = rownames(sp))
  W.mat <- spdep::nb2mat(W.nb, style = "B", zero.policy = TRUE)
  # rownames(W.mat) <- NULL
  # Bayesian hierarchical model with spatially correlated random effects
  if (phi == "INDEP") m <- CARBayes::S.CARleroux(formula=formula,data=Data,
    family="poisson", W=W,rho = 0,  burnin=100000, n.sample=300000, thin=20)
  if (phi == "IAR") m <- CARBayes::S.CARleroux(formula=formula,data=Data,
    family="poisson", W=W, rho = 1,  burnin=1000, n.sample=3000, thin=20)
  if (phi == "leroux") m <- CARBayes::S.CARleroux(formula=formula,data=Data,
    family="poisson", W=W, burnin=100000, n.sample=300000, thin=20)
  if (phi == "BYM") m <- CARBayes::S.CARbym(formula=formula,data=Data,
      family="poisson", W=W, burnin=1000, n.sample=3000, thin=20)
  if (phi == "LOCA1") m <- CARBayes::S.CARdissimilarity(formula=formula,
  data=Data, family="poisson", W=W, Z=list(Z.ratedep=Z.ratedep), W.binary=TRUE,
  burnin=1000, n.sample=3000, thin=20) # Lee and Mitchell (2012)
  if (phi == "LOCA2") m <- CARBayes::S.CARlocalised(formula=formula, data=Data,
  family="poisson", G=3, W=W, burnin=100000, n.sample=300000, thin=20) # Lee and Sarran (2015)
  # posterior distribution of boundary likelihood value (BLV) and boundary membership value (BMV)
  # fitted values (mu): m$samples$fitted[1,] == m$samples$beta[1,] + m$samples$phi[1,]
  #blv <- posterior_blv(m$samples$fitted/Data$E, as.matrix(sl@data[,1:2])) # RR
  blv <- posterior_blv((m$samples$fitted*1000)/Data$pop,
  as.matrix(sl@data[,1:2])) # IR1000
  blv<-abs(blv)
  if ( is.na(threshold)) bmv <- t(apply(blv, 1, function(iter) iter / max(iter)))
  if (!is.na(threshold)) bmv <- blv > threshold
  # create MCMC object for blv and bmv
  mcpar          <- attr(m$samples$beta, "mcpar")
  m$samples$blv <- coda::mcmc(blv, start = mcpar[1], end = mcpar[2], thin = mcpar[3])
```

```
  m$samples$bmv <- coda::mcmc(bmv, start = mcpar[1], end = mcpar[2], thin = mcpar[3])
  # add posterior median to SpatialLinesDataFrame
  sl$blv_median <- apply(blv, 2, median)
  if ( is.na(threshold)) sl$bmv_median <- apply(bmv, 2, median) #  blv or bmv ??
  if (!is.na(threshold)) sl$bmv_mean <- colMeans(bmv)
  # return model and SpatialLinesDataFrame
  m$borders <- sl
  return(sl)
}


### Convert SpatialPolygonsDataFrame to SpatialLinesDataFrame with border segments
#'
#' \code{border_lines} converts a SpatialPolygonsDataFrame to a SpatialLinesDataFrame with one element
#'
#' @param sp Object of type \code{SpatialPolygonsDataFrame} (sf objects as converted)
#' @param longlat Use Euclidean or Great Circle distance for calculation of line length. If FALSE, Eucl
#' @return Object of class \code{SpatialLinesDataFrame} with one element for each border between neighb
#'
#' @importFrom magrittr "%<>%"
#' @importFrom magrittr "%>%"
#' @export
border_lines <- function(sp, longlat = TRUE) {
  # Coerce simple feature geometries to corresponding Spatial* objects
  if (is(sp, "sf")) sp <- as(sp, "Spatial")
  P  <- sp::polygons(sp)
  # get adjacency matrix A
  # nb <- spdep::poly2nb(sp, row.names = rownames(sp), queen = FALSE)
  # A  <- nb2mat::nb2mat(nb, style = "B", zero.policy = TRUE)
  nb <- spdep::poly2nb(sp, queen = FALSE)
  # create data.frame with adjacent areas
  greater_than <- function(a, b) a[a > b]
  data <- data.frame(i = 1:length(nb), j = NA) %>%
    group_by(i, j) %>%
    do(expand.grid(i = .$i, j = greater_than(nb[[.$i]], .$i))) %>%
    as.data.frame()
  # area borders as SpatialLines
  lines <- apply(data, 1, function(d) {
    i <- as.numeric(d["i"])
    j <- as.numeric(d["j"])
    # get list of coordinates for polygons
    c1 <- plyr::llply(P@polygons[[i]]@Polygons, sp::coordinates)
    c2 <- plyr::llply(P@polygons[[j]]@Polygons, sp::coordinates)
    # get borders for each combination of polygons
    grid <- expand.grid(s1 = 1:length(c1), s2 = 1:length(c2))
    line <- apply(grid, 1, function(obj) {
      a   <- c1[[obj["s1"]]]
      b   <- c2[[obj["s2"]]]
      # select intersecting rows
      sel <- a[, 1] %in% b[, 1] & a[, 2] %in% b[, 2]
      if(sum(sel) == 0) return(NULL)
      # create Line object for each sequence of matching coordinates
```

```
      runs <- rle(sel)
      runs <- data.frame(
        val = runs$values,
        i   = c(1, cumsum(runs$length) + 1)[-(length(runs$length) + 1)],
        len = runs$length) %>%
        dplyr::filter(val)
      # coordinates for each sequence
      pos    <- plyr::alply(as.matrix(runs), 1, . %>% {.[["i"]] : (.[["i"]] +
                  .[["len"]] - 1)})
      coords <- plyr::llply(pos, . %>% a[., , drop = FALSE])
      # remove duplicate line elements
      len_one <- plyr::laply(coords, nrow) == 1
      if (!all(len_one) & any(len_one)) {
        B <- do.call(rbind, coords)
        coords <- plyr::llply(coords, function(A) {
          if(nrow(A) > 1) return(A)
          cond <- sum(A[, 1] == B[, 1] & A[, 2] == B[, 2]) > 1
          if(cond) return(NULL)
          return(A)
        })
        coords <- coords[!sapply(coords, is.null)]
      }
# return list of Line objects (one element for each sequence of coordinates)
      plyr::llply(coords, sp::Line)
    })
    segments <- unlist(line[!sapply(line, is.null)], recursive=FALSE)
    sp::Lines(segments, ID = sprintf("i%s_j%s", i, j))
  })
  sl <- sp::SpatialLines(lines[!sapply(lines, is.null)],
        proj4string = sp::CRS(sp::proj4string(sp)))
  # SpatialLinesDataFrame from SpatialLines and data
  sldf <- sp::SpatialLinesDataFrame(sl, data, match.ID = FALSE)
  # sldf$length <- SpatialLinesLengths(sldf, longlat = longlat)
  return(sldf)
}


### posterior_blv
posterior_blv <- function(mu, adj) {
  iters <- nrow(mu)
  n <- nrow(adj)
  blv <- matrix(nrow = iters, ncol = n)

  for (i in 1:n) {
    for (iter in 1:iters) {
      blv[iter, i] <- mu[iter, adj[i, 1]] - mu[iter, adj[i, 2]]
      # blv[iter, i] <- mu[iter, adj[i, 1] - 1] - mu[iter, adj[i, 2] - 1];
    }
  }
  return(blv);
}
```

```
### Exploratory data analysis
# Moran's I : Spatial Autocorrelation diagnostic
nb <- poly2nb(Data, queen=TRUE) ## Define neighborhood
# Create neighbours list with row-standardized weights
col.W <- nb2listw(nb, style="W")
I<-moran(x=Data$IR1000,listw=col.W, n=length(nb), S0=Szero(col.W))[1]
moran.plot(x=Data$IR1000,listw=col.W, n=length(nb), S0=Szero(col.W))
# Testing based on normal approximation
moran.test(x=Data$IR1000, listw=col.W)
# Testing based on randomization
nsim <- 999; set.seed(1234)
MC<-moran.mc(x=Data$IR1000, listw=col.W, nsim)
hist(MC$res,xlab="Monte Carlo simulation of I",main="");
abline(v=MC$statistic,col="red")

### Algorithm-based wombling : Crisp wombing
Data.sp<- st_transform(x=Data, crs='+proj=longlat +datum=WGS84 +no_defs')
border<-border_lines(Data.sp, longlat = TRUE)
x<-"IR1000"
## Threshold1 : 1st Quartile = 6
summary(Data.sp$IR1000)
Algo_Crisp_w1<-areal_wombling(Data.sp, x, threshold = 6, dist = function(x)
  abs(x[1] - x[2]))
table(Algo_Crisp_w1$IR1000_bmv)
## Plot 1: Boundary values for border line segments
colors <- c("green", "red")
plot(Algo_Crisp_w1, col = colors[factor(Algo_Crisp_w1$IR1000_bmv, levels =
 c("FALSE", "TRUE"))], lwd = 0.03,  main= "Threshold=6;  Wave 1 (T=138)")
# The spatial lines object does not include Belgium's municipalities boundaries.
# Let's add them
Crisp_w1 <- st_union(Data.sp)
plot(as(Crisp_w1, "Spatial"), lwd = 0.5, add = TRUE)
legend(x = "bottomleft", legend = c("TRUE","FALSE"), lty = c(1, 1),
       col = c(2, 3), lwd = 1)
## Plot 2 : Boundary values for border line segments for areal units
# aggregate from line segments to block group level
wave1_blv <- bind_rows(
  group_by(Algo_Crisp_w1@data, i) %>%
    summarise_at(vars(ends_with("blv")), max, na.rm = TRUE),
  group_by(Algo_Crisp_w1@data, j) %>%
summarise_at(vars(ends_with("blv")), max, na.rm = TRUE) %>% dplyr::rename(i = j)
) %>%
  group_by(i) %>%
  summarise_at(vars(ends_with("blv")), max, na.rm = TRUE)
b_wave1 <- bind_cols(Data.sp, select(wave1_blv, -i))
b_wave1=b_wave1%>%mutate(BMV= case_when(IR1000_blv>=6~"TRUE",IR1000_blv<6 ~ "FALSE"))
table(b_wave1$BMV)
sel <- is.finite(b_wave1$IR1000_blv)
plot(as(b_wave1[sel,], "Spatial"), lwd = 0.1, col = colors[factor(b_wave1$BMV,
levels = c("FALSE", "TRUE"))], main= "Threshold=10; Wave1 (43 municipalities)")
legend(x = "bottomleft", legend = c("TRUE","FALSE"), col = c(2, 3), pch=16)
```

```
### Algorithm-based wombling : Fuzzy wombing
Data.sp<- st_transform(x=Data, crs='+proj=longlat +datum=WGS84 +no_defs')
border<-border_lines(Data.sp, longlat = TRUE)
x<-"IR1000"
Algo_Fuzzy_w1<-areal_wombling(Data.sp, x, threshold = NA, dist = function(x)
  abs(x[1] - x[2]))
table(Algo_Fuzzy_w1$IR1000_bmv)
# cutting point 1 : 50%
cc<-Algo_Fuzzy_w1$IR1000_bmv
cc=as.data.frame(cc)
ccc<- cc %>%mutate(BMV= case_when(cc>=0.50~"TRUE",cc<0.5 ~ "FALSE"))
Algo_Fuzzy_w1$BMV<- ccc$BMV
table(Algo_Fuzzy_w1$BMV)
## Plot 1: Boundary values for border line segments
colors <- c("green", "red")
plot(Algo_Fuzzy_w1, col = colors[factor(Algo_Fuzzy_w1$BMV,
levels = c("FALSE", "TRUE"))], lwd = 0.03, main= "Wave 1 (T=41)")
# The spatial lines object does not include Belgium's municipalities boundaries.
Fuzzy_w1 <- st_union(Data.sp)
plot(as(Fuzzy_w1, "Spatial"), lwd = 0.5, add = TRUE)
legend(x = "bottomleft", legend = c("TRUE","FALSE"), lty = c(1, 1),
       col = c(2, 3), lwd = 1)

## Plot 2 : Boundary values for border line segments for areal units
# aggregate from line segments to block group level
wave1_blv <- bind_rows(
  group_by(Algo_Fuzzy_w1@data, i) %>%
    summarise_at(vars(ends_with("bmv")), max, na.rm = TRUE),
  group_by(Algo_Fuzzy_w1@data, j) %>%
summarise_at(vars(ends_with("bmv")), max, na.rm = TRUE) %>% dplyr::rename(i = j)
) %>%
  group_by(i) %>%
  summarise_at(vars(ends_with("bmv")), max, na.rm = TRUE)
b_wave1 <- bind_cols(Data.sp, select(wave1_blv, -i))
b_wave1=b_wave1%>%mutate(BMV= case_when(IR1000_bmv>=0.50~"TRUE",
                        IR1000_bmv<0.50 ~ "FALSE"))
table(b_wave1$BMV)
sel <- is.finite(b_wave1$IR1000_bmv)
plot(as(b_wave1[sel,], "Spatial"), lwd = 0.1, col = colors[factor(b_wave1$BMV,
            levels = c("FALSE", "TRUE"))], main= "Wave1 (56 municipalities)")
legend(x = "bottomleft", legend = c("TRUE","FALSE"), col = c(2, 3), pch=16)

### Global CAR modeling
formula <- Data$CaseObs~ offset(log(Data$E)) # observe ~ expected
## models
model1 <- S.CARleroux(formula=formula,data=Data,  family="poisson",  W=W,
rho = 0,  burnin=100000, n.sample=300000, thin=20) # indep
model2 <- S.CARleroux(formula=formula,data=Data,  family="poisson",  W=W,
burnin=100000, n.sample=300000, thin=20) # Leroux
model3 <- S.CARbym(formula=formula,data=Data,  family="poisson",  W=W,
 burnin=100000, n.sample=300000, thin=20) # BYM
```

```
INDEP<-print(model1$modelfit)
Leroux<-print(model2$modelfit)
Bym<-print(model3$modelfit)
Global_smo<-rbind(INDEP,Leroux,Bym)

Data$IR1000g<-(model2$fitted.values*1000)/Data$pop
Data$Residg<-model2$residuals$response # extract residuals

## Inference
# model
model2_1 <- S.CARleroux(formula=formula,data=Data,  family="poisson",  W=W,
 burnin=100000, n.sample=300000, thin=20) # Leroux
# convergence diagnostic 1: stat in -1.96, 1.96 suggest convergence (Geweke.diag)
print(model2_1 )
# convergence diagnostic 2: traceplot comparing the results from the multiple
# chains in coda
model2_2 <- S.CARleroux(formula=formula,data=Data,  family="poisson",  W=W,
burnin=100000, n.sample=300000, thin=20) # Leroux
model2_3 <- S.CARleroux(formula=formula,data=Data,  family="poisson",  W=W,
burnin=100000, n.sample=300000, thin=20) # Leroux

beta.samples <- mcmc.list(model2_1$samples$beta, model2_2$samples$beta,
model2_3$samples$beta)

# convergence diagnostic 3: potential scale reduction factor (PSRF, Gelman et al.
# (2003) ; value less than 1.1 is suggestive of convergence
gelman.diag(beta.samples)
beta.samples.matrix <- rbind(model2_1$samples$beta, model2_2$samples$beta,
model2_3$samples$beta)
colnames(beta.samples.matrix) <- colnames(model2_1$X)
round(t(rbind(apply(beta.samples.matrix, 2, mean), apply(beta.samples.matrix,
    2, quantile, c(0.025, 0.975)))), 5)


### Locally smoothing modelling

## Define dissimilarity metric
Depriv <- Data$Depriv
Z.Depriv <- as.matrix(dist(Depriv, diag=TRUE, upper=TRUE))
formula <- Data$CaseObs~ offset(log(Data$E)) # observe ~ expected
# Lee and Mitchell (2012)
model4 <-S.CARdissimilarity(formula=formula, data=Data, family="poisson", W=W,
 Z=list(Z.Depriv=Z.Depriv), W.binary=TRUE, burnin=10000,
 n.sample=30000, thin=20) #

# Lee and Sarran (2015)
model5 <- S.CARlocalised(formula=formula, data=Data,family="poisson", G=3, W=W,
burnin=10000,n.sample=30000, thin=20)
Loca1<-print(model4$modelfit)
Loca2<-print(model5$modelfit)
Locally_smo<-rbind(Loca1,Loca2)
Data$IR1000l1<-(model4$fitted.values*1000)/Data$pop
```

```
Data$Residl1<-model4$residuals$response
Data$IR1000l2<-(model5$fitted.values*1000)/Data$pop
Data$Residl2<-model5$residuals$response

## Inference 1
model4_1 <- S.CARdissimilarity(formula=formula, data=Data, family="poisson", W=W,
Z=list(Z.popdep=Z.popdep),W.binary=TRUE, burnin=100000, n.sample=300000, thin=20)

# convergence diagnostic 1: stat in -1.96, 1.96 suggest convergence (Geweke.diag)
# convergence diagnostic 2: traceplot comparing the results from the multiple
# chains in coda
model4_2 <- S.CARdissimilarity(formula=formula, data=Data, family="poisson", W=W,
Z=list(Z.popdep=Z.popdep),W.binary=TRUE, burnin=100000, n.sample=300000, thin=20)
model4_3 <- S.CARdissimilarity(formula=formula, data=Data, family="poisson", W=W,
Z=list(Z.popdep=Z.popdep),W.binary=TRUE, burnin=100000, n.sample=300000, thin=20)
beta.samples <- mcmc.list(model4_1$samples$beta, model4_2$samples$beta,
model4_3$samples$beta)

# convergence diagnostic 3: potential scale reduction factor (PSRF, Gelman et al.
# (2003) ; value less than 1.1 is suggestive of convergence
gelman.diag(beta.samples)
beta.samples.matrix <- rbind(model4_1$samples$beta, model4_2$samples$beta,
model4_3$samples$beta)
colnames(beta.samples.matrix) <- colnames(model4_1$X)
round(t(rbind(apply(beta.samples.matrix, 2, mean), apply(beta.samples.matrix,
2, quantile, c(0.025, 0.975)))), 5)

### Model-based wombling 1 : Crisp wombing
Data.sp<- st_transform(x=Data, crs='+proj=longlat +datum=WGS84 +no_defs')
border<-border_lines(Data.sp, longlat = FALSE)
x<-"IR1000g"
formula <- Data$CaseObs~ offset(log(Data$E)) # observe ~ expected
### Threshold1 : 1st Quartile = 6
#Bayes_Crisp_w1<-areal_wombling(Data.sp, x, threshold = 6, dist = function(x)
#                                                    abs(x[1] - x[2]))
Bayes_Crisp_w1<-areal_wombling_bayesian(formula, family="poisson", Data.sp,
          phi = "leroux", threshold = 6, E,pop)
table(Algo_Crisp_w1$IR1000_bmv)
## Plot 1: Boundary values for border line segments
colors <- c("green", "red")
plot(Algo_Crisp_w1, col = colors[factor(Algo_Crisp_w1$IR1000_bmv, levels =
c("FALSE", "TRUE"))], lwd = 0.03,  main= "Threshold=6;  Wave 1 (T=138)")
# The spatial lines object does not include Belgium's municipalities boundaries.
# Let's add them
Crisp_w1 <- st_union(Data.sp)
plot(as(Crisp_w1, "Spatial"), lwd = 0.5, add = TRUE)
legend(x = "bottomleft", legend = c("TRUE","FALSE"), lty = c(1, 1),
       col = c(2, 3), lwd = 1)
## Plot 2 : Boundary values for border line segments for areal units
# aggregate from line segments to block group level
wave1_blv <- bind_rows(
  group_by(Algo_Crisp_w1@data, i) %>%
```

```
    summarise_at(vars(ends_with("blv")), max, na.rm = TRUE),
  group_by(Algo_Crisp_w1@data, j) %>%
summarise_at(vars(ends_with("blv")), max, na.rm = TRUE) %>% dplyr::rename(i = j)
) %>%
  group_by(i) %>%
  summarise_at(vars(ends_with("blv")), max, na.rm = TRUE)
b_wave1 <- bind_cols(Data.sp, select(wave1_blv, -i))
b_wave1=b_wave1%>%mutate(BMV= case_when(IR1000_blv>=6~"TRUE",
      IR1000_blv<6 ~ "FALSE"))
table(b_wave1$BMV)
sel <- is.finite(b_wave1$IR1000_blv)
plot(as(b_wave1[sel,], "Spatial"), lwd = 0.1, col = colors[factor(b_wave1$BMV,
levels = c("FALSE", "TRUE"))], main= "Threshold=10 ; Wave1 (43 municipalities)")
legend(x = "bottomleft", legend = c("TRUE","FALSE"), col = c(2, 3), pch=16)

### Algorithm-based wombling : Fuzzy wombing
Data.sp<- st_transform(x=Data, crs='+proj=longlat +datum=WGS84 +no_defs')
border<-border_lines(Data.sp, longlat = TRUE)
x<-"IR1000"
Algo_Fuzzy_w1<-areal_wombling(Data.sp, x, threshold = NA, dist = function(x)
  abs(x[1] - x[2]))
table(Algo_Fuzzy_w1$IR1000_bmv)
# cutting point 1 : 50%
cc<-Algo_Fuzzy_w1$IR1000_bmv
cc=as.data.frame(cc)
ccc<- cc %>%mutate(BMV= case_when(cc>=0.50~"TRUE",cc<0.5 ~ "FALSE"))
Algo_Fuzzy_w1$BMV<- ccc$BMV
table(Algo_Fuzzy_w1$BMV)
## Plot 1: Boundary values for border line segments
colors <- c("green", "red")
plot(Algo_Fuzzy_w1, col = colors[factor(Algo_Fuzzy_w1$BMV,
levels = c("FALSE", "TRUE"))], lwd = 0.03, main= "Wave 1 (T=41)")
# The spatial lines object does not include Belgium's municipalities boundaries.
Fuzzy_w1 <- st_union(Data.sp)
plot(as(Fuzzy_w1, "Spatial"), lwd = 0.5, add = TRUE)
legend(x = "bottomleft", legend = c("TRUE","FALSE"), lty = c(1, 1),
       col = c(2, 3), lwd = 1)

## Plot 2 : Boundary values for border line segments for areal units
# aggregate from line segments to block group level
wave1_blv <- bind_rows(
  group_by(Algo_Fuzzy_w1@data, i) %>%
    summarise_at(vars(ends_with("bmv")), max, na.rm = TRUE),
  group_by(Algo_Fuzzy_w1@data, j) %>%
summarise_at(vars(ends_with("bmv")), max, na.rm = TRUE) %>% dplyr::rename(i = j)
) %>%
  group_by(i) %>%
  summarise_at(vars(ends_with("bmv")), max, na.rm = TRUE)
b_wave1 <- bind_cols(Data.sp, select(wave1_blv, -i))
b_wave1=b_wave1%>%mutate(BMV= case_when(IR1000_bmv>=0.50~"TRUE",
                                        IR1000_bmv<0.50 ~ "FALSE"))
table(b_wave1$BMV)
```

```
sel <- is.finite(b_wave1$IR1000_bmv)
plot(as(b_wave1[sel,], "Spatial"), lwd = 0.1, col = colors[factor(b_wave1$BMV,
     levels = c("FALSE", "TRUE"))], main= "Wave1 (56 municipalities)")
legend(x = "bottomleft", legend = c("TRUE","FALSE"), col = c(2, 3), pch=16)

### Wombling using dissimilarity metric  in CARBayes
## Adjacency matrix
W.nb <- poly2nb(Data, row.names = Data$NIS5)
W <- nb2mat(W.nb, style="B")
##  dissimilarity metric
Depriv <- Data$Depriv
Z.Depriv <- as.matrix(dist(Depriv, diag=TRUE, upper=TRUE))

## S.CARdissimilarity() :  CAR model + dissimilarity metrics
## W.binary=TRUE makes elements in W are ones or zeros corresponding to boundaries
formula <- Data$CaseObs~ offset(log(Data$E)) # observe ~ expected
model <- S.CARdissimilarity(formula=formula, data=Data,
       family="poisson", W=W, Z=list(Z.Depriv=Z.Depriv),
          W.binary=TRUE, burnin=100000, n.sample=300000, thin=20)

##  number and locations of these boundaries
# boundaries as a SpatialPoints
border.locations <- model4$localised.structure$W.posterior # matrix of border locations
Data$IR1000p <- (model4$fitted.values*1000)/Data$pop #  estimated IR1000
boundary.final <- highlight.borders(border.locations=border.locations,
    sfdata=Data) # identifies the boundary points using the CARBayes function
# highlight.borders() and formats them to enable plotting

# plotting
boundary.coordinates <- st_coordinates(boundary.final)
colours <- colorNumeric(palette = "YlOrRd", domain = Data$IR1000p)
map1 <- leaflet(data=Data) %>% addTiles() %>%
addPolygons(fillColor = ~colours(IR1000p), color="", weight=1, fillOpacity = 0.7)%>%
  addLegend(pal = colours, values = Data$IR1000p, opacity = 1, bins = 5,
            title="CAR dissi Womb") %>%
  addCircles(lng = ~boundary.coordinates[,1], lat = ~boundary.coordinates[,2],
             weight = 1, radius = 2) %>%
  addScaleBar(position="bottomleft")
map1
##############################################################################
```

A.2 Figures

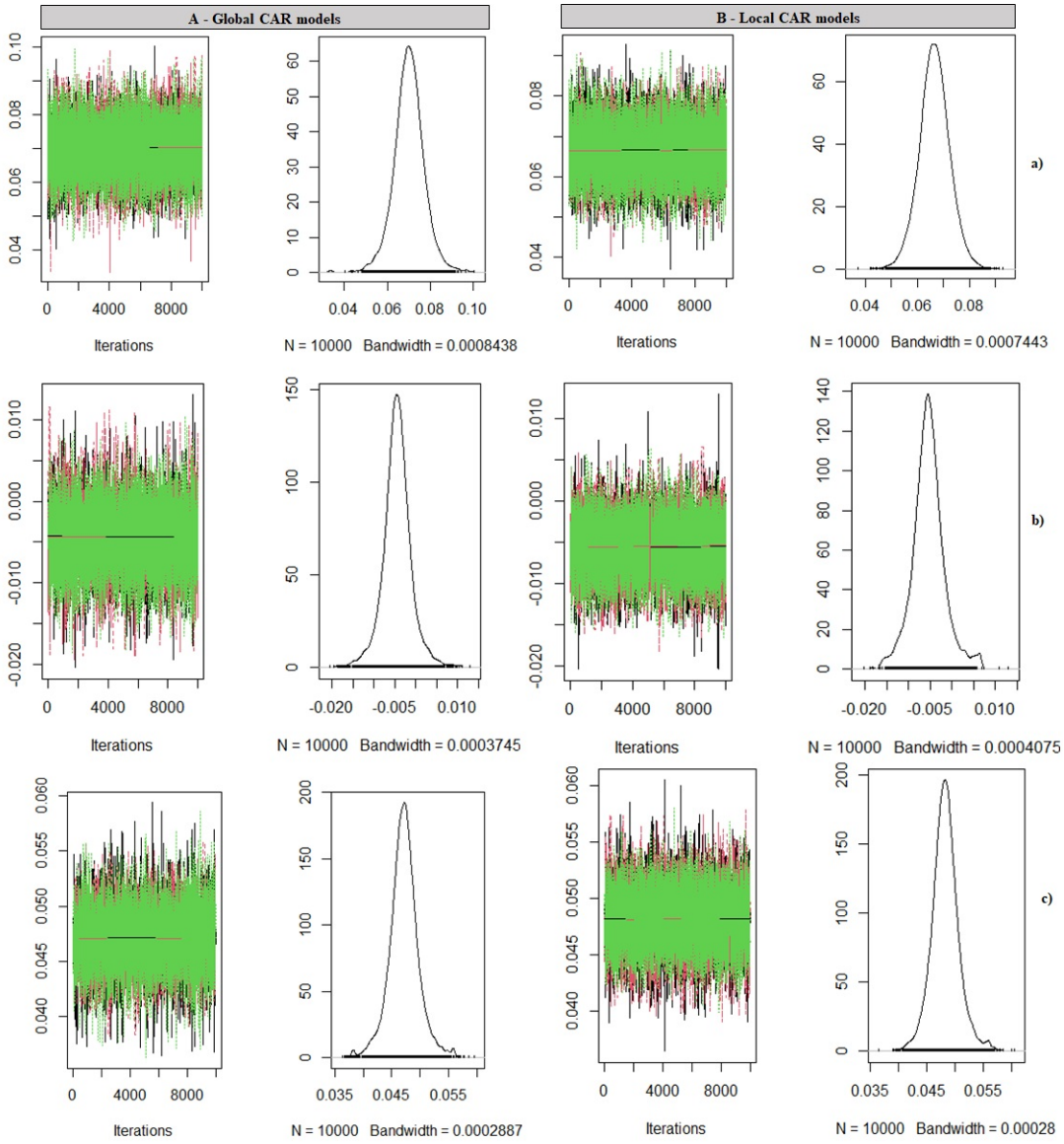

Figure A.2.1: Convergence of the Markov chains 1 (a: wave1; b: wave2; c: wave3)

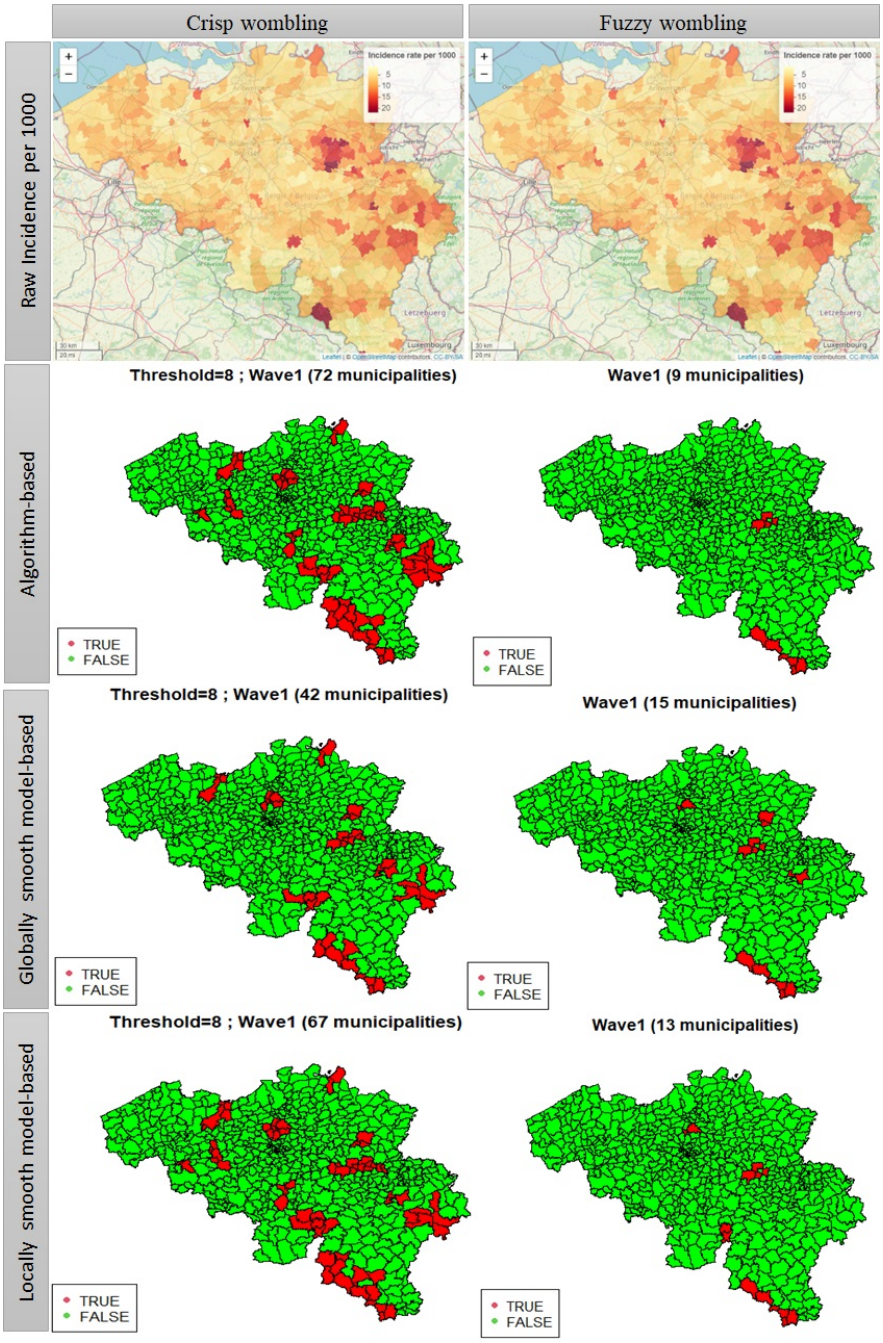

Figure A.2.2: Model-based wombling for wave 1 (municipalities in boundaries

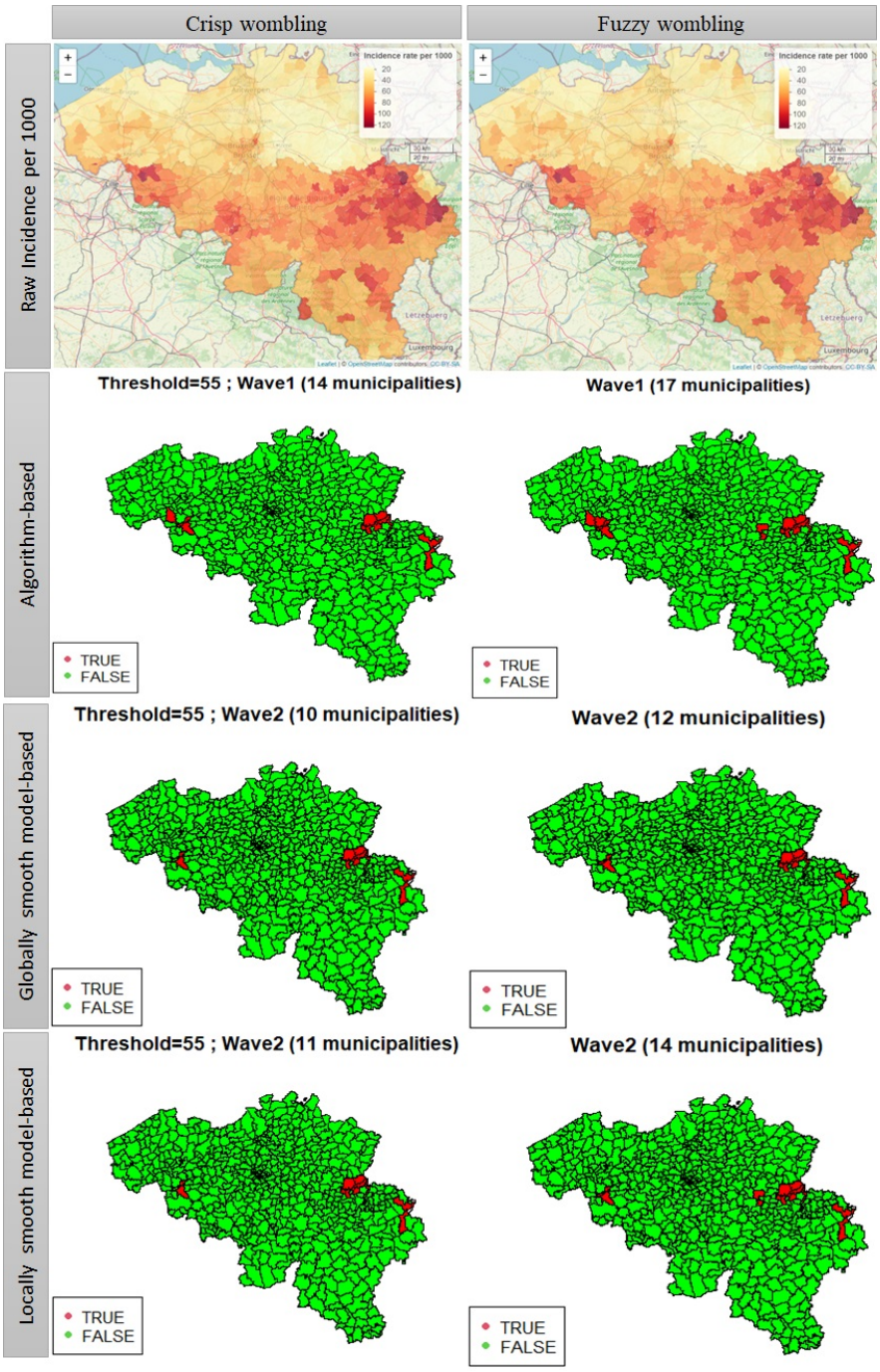

Figure A.2.3: Model-based wombling for wave 2 (municipalities in boundaries

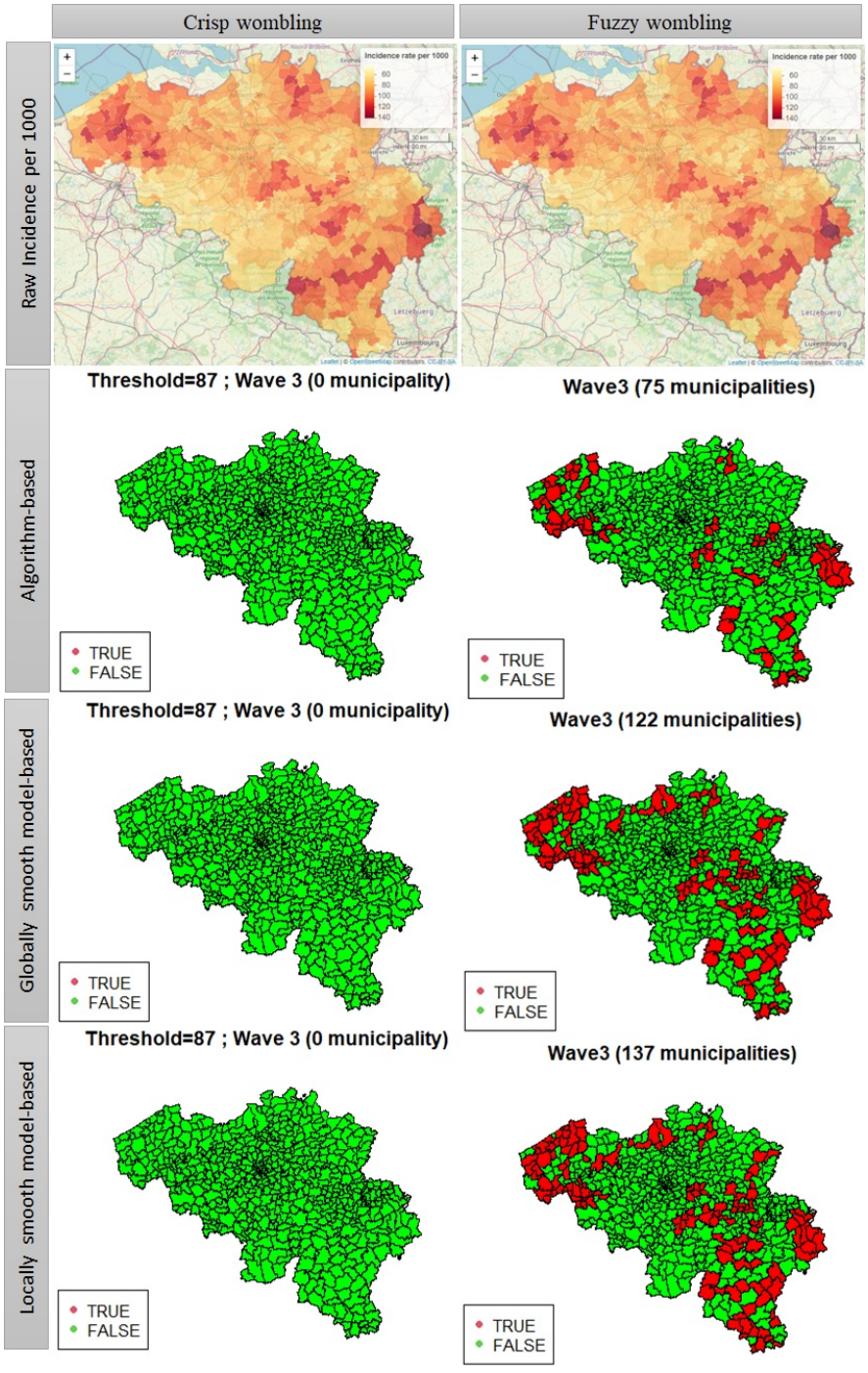

Figure A.2.4: Model-based wombling for wave 3 (municipalities in boundaries