Journal of Machine Learning Research 20 (2019) 1-35 Submitted 12/16; Revised 9/18; Published 5/19

Bayesian Space-Time Partitioning by Sampling and Pruning Spanning Trees

Leonardo V. Teixeira LTEIXEIR@PURDUE.EDU
Department of Computer Science
Purdue University
West Lafayette, IN 47907, USA

Renato M. Assunção ASSUNCAO@DCC.UFMG.BR
Departamento de Cieência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Rosangela H. Loschi LOSCHI@EST.UFMG.BR
Departamento de Estatística
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Editor: Jennifer Dy

Abstract

A typical problem in spatial data analysis is regionalization or spatially constrained clustering, which consists of aggregating small geographical areas into larger regions. A major challenge when partitioning a map is the huge number of possible partitions that compose the search space. This is compounded if we are partitioning spatio-temporal data rather than purely spatial data. We introduce a spatio-temporal product partition model that deals with the regionalization problem in a probabilistic way. Random spanning trees are used as a tool to tackle the problem of searching the space of possible partitions making feasible this exploration. Based on this framework, we propose an efficient Gibbs sampler algorithm to sample from the posterior distribution of the parameters, specially the random partition. The proposed Gibbs sampler scheme carries out a random walk on the space of the spanning trees and the partitions induced by deleting tree edges. In the purely spatial situation, we compare our proposed model with other state-of-art regionalization techniques to partition maps using simulated and real social and health data. To illustrate how the temporal component is handled by the algorithm and to show how the spatial clusters vary along the time we presented an application using human development index data. The analysis shows that our proposed model is better than state-of-art alternatives. Another appealing feature of the method is that the prior distribution for the partition is interpretable with a trivial coin flipping mechanism allowing its easy elicitation.

Keywords: Spatial Clustering, Product Partition Models, Random Spanning Trees, Bayesian Clustering

1. Introduction

Traditional cluster analysis aims at partitioning a set of n objects into k clusters such that the clusters are composed by objects with similar valued attributes while objects from

different clusters tend to be dissimilar. When the objects are spatially located, inducing a neighborhood structure, the clustering can be constrained by their spatial contiguity. For instance, in mapping problems, it is common to aggregate neighboring small areas to partition the map into larger regions, named *spatial clusters*. The small areas within a given spatial cluster are relatively homogeneous with respect to attributes such as, for example, ecological characteristics (Sayre et al., 2014) or crime rates (Mennis and Harris, 2013). This spatial clustering problem is called *regionalization* or *spatially constrained clustering*.

Regionalization serves to a range of purposes, such as to facilitate the visualization and understanding of the geographical information, to reduce the noise introduced by outliers and inaccurate data, to make data analysis tractable, or to provide a better statistical handling of the data by reducing the effect of different populations (Wise et al., 1997). It has been used in widely different applied problems such as ecoregion delimitation (Sayre et al., 2014; McKenney et al., 2007), climate zoning (Zhang et al., 2016), environmental planning (Bernetti et al., 2011), image segmentation (Ribeiro et al., 2013), communication protocols in geo-sensor networks (Reis et al., 2007), map generalization (Ruas, 2008), zone design for health studies (Cockings and Martin, 2005; Ricketts, 1997), and enhanced sampling procedures (Martin, 1998).

There are two possible ways to carry out this aggregation. One way is through an artificial clustering, where the constructed regions are predetermined using official or normative designations (such as states, districts and counties). That is, the regionalization of small units of interest, such as counties, are arbitrarily specified as larger units such as states, defined for administrative or political reasons. This kind of aggregation is usually the expression of political will and may not take into account the information specific to the domain being studied. Another way is to perform the aggregation based on the analysis of data characteristics related to the phenomena under study. An outcome of a regionalization method can be found in the left hand side of Figure 1, which shows the three spatial clusters based on the values of lung cancer mortality rates in municipalities in the South of Brazil for the period 2008-2012.

Many different methods have been proposed to deal with the problem of spatial regionalization as a non-stochastic optimization problem. These previous works are presented in Section 1.1. Most of these regionalization techniques consider data as fixed, static values and prefix the number of clusters *a priori*. Frequently, these assumptions are inappropriate, as they do not allow for measurement error or uncertainty on the areas' measures nor the evaluation of the uncertainty of the obtained clusters. For example, consider the bladder cancer mortality rate of a small town in a given year. This value should not be considered as the most representative value for the true mortality rate since it can be severely impacted by small differences from one year to another, particularly if the area has a small population. The measured value can show a natural variability, expressed in this widely different bladder cancer mortality rate variation in two successive years in small population areas. Traditional regionalization techniques that emphasizes the similarity between the observations will be sensitive to this variability and may output an regionalization that discloses an undesirably high number of clusters. To illustrate this problem, the right hand side map in Figure 1 shows an outcome of the *Automatic Zoning Procedure* (AZP) regionalization technique proposed by Openshaw (1977) considering the bladder cancer mortality rate in municipalities in the South of Brazil for the period 2008-2012. This type of cancer is rare, generating small

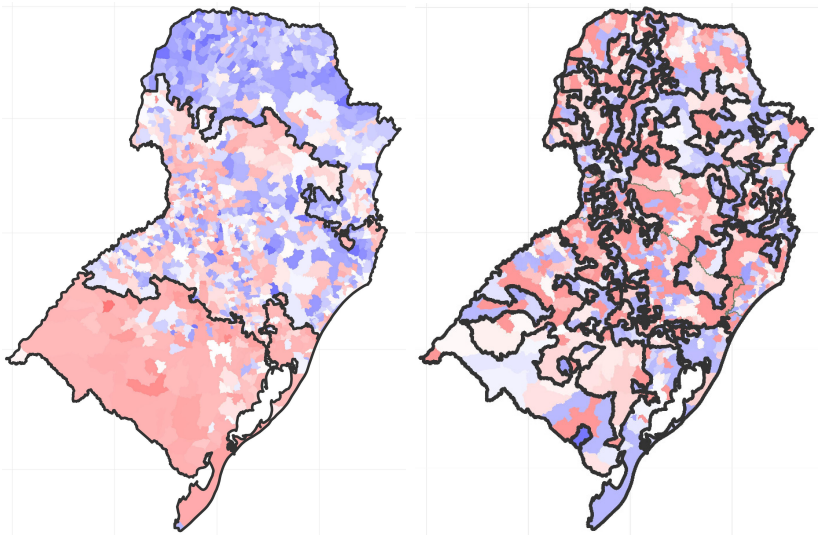


Figure 1: Example of a good (left) and a bad (right) regionalization using data from lung (left) and bladder (right) cancer mortality rates in South Brazilian municipalities for the period 2008-2012.

death counts in low population areas and thus extreme rates (high or low) in sharp contrast with their neighbors.

It is necessary to use an explicit stochastic framework to perform the regionalization if we want to take into account this natural variability in the data. An appropriate modeling approach should also allow us to quantify the uncertainty about the geographic partition. A stochastic framework accounting for such characteristics defines a random partition model. Consequently, our inference becomes more complex as there is a huge number of possible partitions that compose the parametric space.

Some attempts to achieve this goal in the spatial context have been already considered in the literature and they are reviewed in Section 1.1. In Teixeira et al. (2015), we introduce a random partition model that deals with purely spatial regionalization problems in a probabilistic way. In this article, in addition to presenting the model, we discuss in detail the elicitation of prior distributions for the partition and for the number of clusters. Such model is an alternative to the spatial product partition model (PPM) introduced by Hegarty and Barry (2008) and by Page and Quintana (2016). We also extend these models to space-time data by also accounting for spatio-temporal regionalization and provide extensive simulation results and new examples. We run a simulation study comparing the proposed model with well-known stochastic and non-stochastic regionalization techniques including Bayesian modeling alternatives. To illustrate the purely spatial case, we partition the Brazilian map based on the Human Development Index (HDI) and the Brazilian South region based on bladder and lung cancer mortality rates. The space-time case is illustrated with HDI in three decades.

The spatial and the spatio-temporal structure are both represented by a graph. The partition of graphs is not a new subject (Green and Thomas, 2013; Bornn and Caron, 2011).

Our contribution is the introduction of random spanning trees into the random partition model as a tool to handle the problem of searching in the space of possible partitions. We propose an efficient Gibbs sampler algorithm to sample from the posterior probability distribution, specially that for the partition. Conditionally on the spanning tree, the random partition has a product distribution thus defining a spatio-temporal/spatial product partition model. This strategy also greatly facilitates the scheme for sampling from the posterior distributions. By conditioning on the partitions resulting from spanning trees pruning, we substantially reduce the space of partitions or clusters to be explored. It is important to emphasize that the partition space is the set of all possible spatio-temporal/spatial partitions. The way we search this space is by selecting a new random spanning tree from the set of all possible spanning trees at each Gibbs step and then proceed with its partitioning. This guarantees that there is a positive probability that any partition will be reached in a finite time starting from any other partition in our MCMC sampling scheme.

In Section 2, we briefly review some basic concepts of graph theory we need in the construction of the proposed model. In Section 3 we introduce our proposed model and describe how we incorporate the spanning trees as a tool to drive through the partition space. In Section 4 we propose an efficient algorithm to sample from the posterior distribution. Section 5 presents a simulation study comparing the proposed model with some other well-known methods for regionalization in the spatial context. The analysis of some real data from Brazil is carried out in Section 6. Section 7 closes the paper with some final comments and the main conclusions.

1.1. Related Work

Openshaw (Openshaw, 1977) was a pioneer when he proposed *Automatic Zoning Procedure* (AZP), a heuristic method to aggregate areas that swapped regions locally improving an initial rough partitioning. Later, AZP variants were introduced using simulated annealing and tabu search (Openshaw and Rao, 1995). A modification of AZP named *Automatic Regionalization with Initial Seed Location* (ARiSeL) was presented by Duque and Church (2004). In ARiSeL, the construction of an initial feasible solution is repeated several times before running a tabu search which, according to the author, is less expensive than performing a local search. The *Self Organizing Maps* (SOM) algorithm, proposed by Kohonen (1990), is an unsupervised neural network which adjusts its weights to represent a data set distribution on a regular lattice. Although used to perform regionalization, the spatial contiguity desirable in a regionalization is not guaranteed. SOM variants have been proposed by Bação et al. (2004) and Bação et al. (2005) considering different procedures to explore the neighborhood structure. The heuristic devised by Aldstadt and Getis (2006), called AMOEBA (*A Multidirectional Optimum Ecotope-Based Algorithm*), starts with an initial area and grows it by adding neighboring areas until a local spatial autocorrelation statistic stops increasing. This process is repeated to all areas and a final step resolves overlaps. The *Max-p-regions* technique (Duque et al., 2012) does not require the previous setting of the number of spatial clusters and enforces the contiguity constraint. Clusters are formed in a such way that a regional attribute is always above certain threshold such as a minimum population or cluster area size.

The *Spatial 'K'luster Analysis by Tree Edge Removal* (SKATER) proposed by Assunção et al. (2006) is a graph based method that uses a minimal spanning tree to reduce the search space. The regions are then defined by the removal of edges from the minimal spanning tree. The removed edges are chosen to minimize a dissimilarity measure. Inspired by SKATER, Guo (2008) proposed REDCAP (*Regionalization with Dynamically Constrained Agglomerative clustering and Partitioning*) which considers six other methods to explore different connection strategies using the underlying graph structure. The model introduced in the present manuscript is also inspired by SKATER by using spanning trees as a tool to summarize the spatial connection and the variability of the data. It adopts a Bayesian approach putting prior distributions on all unknown aspects of the regionalization, including the partition itself. As a consequence, we are able to make inference on unknown parameters and, most importantly, to infer about the partition in a probabilistic way. Our method outputs a posterior distribution over all possible partitions and we can infer which ones are the most probable.

Knorr-Held and Raßer (2000) presented a Bayesian approach, named *Bayesian Detection of Clusters and Discontinuities* (BDCD), to aggregate small contiguous areas into larger regions to form spatial clusters. Focusing on the random partition, they assume a prior distribution $p(c)$ for the number c of spatial clusters. Conditioned on c, cluster centers are uniformly selected among the n available centroids. Each area is assigned to that cluster whose center is the nearest according to the number of boundaries that need to be crossed. In the Bayesian partition model (BPM) introduced by Denison and Holmes (2001), the prior distribution $p(c)$ is an uniform distribution and they use a Voronoi tessellation to determine the spatial clusters. One of the main differences between the two of them is the computational strategy to sample from the posterior distribution. While Knorr-Held and Raßer (2000) use an explicit reversible jump Markov chain Monte Carlo (MCMC) procedure (Green, 1995; Richardson and Green, 1997), Denison and Holmes (2001) proceed using a standard MCMC after integrating out some parameters.

Gangnon and Clayton (2000) proposed a different Bayesian approach for this problem by assuming a prior distribution for the number of clusters that depends on the geometry of the clusters. Such prior distribution gives more flexibility for spatial cluster analysis since, depending on our prior knowledge, it can be elicited penalizing some particular cluster aspects such as large cluster sizes or odd-shaped clusters. The algorithm proposed to approximate the posterior distribution has two components. The first is a window of plausibility, an adaptation of the Occam's window approach to model selection (Madigan and Raftery, 1994). In the second, given a window of plausibility, they use a randomized search algorithm similar to the backward elimination method used for variable selection in regression problems.

Lu and Carlin (2005) and Banerjee and Gelfand (2006) worked on a dual problem proposing a method known as boundary (or wobbling) analysis. Rather than aggregating similar areas into homogeneous regions, their method aims at identifying sharp boundaries between pairs of areas in such a way that homogeneous regions can be obtained as a byproduct. Wakefield and Kim (2013) proposed a Bayesian method for the detection of a small number of high risk zones, rather than providing a map partition. It requires the pre-specification of the maximum number of clusters, the clusters found tend to have a circular shape.

Anderson et al. (2014) proposed a two-step regionalization method. First, a hierarchical clustering method is used to define a set of possible partitions. In the second step, the

models defined by each partition are compared using a model selection tool and the best partition is thus selected.

The product partition model (PPM) introduced by Hartigan (1990) has been used for different purposes due to its flexibility in modeling heterogeneous data. However, only recently it has been considered to analyze spatial data. PPMs assume that the partition $\boldsymbol{\pi} = \{\mathcal{G}_1, \ldots, \mathcal{G}_c\}$ of a set of data is a random quantity and its main feature is to assume that the $\boldsymbol{\pi}$ distribution is a product of subjective non-negative functions $\kappa(\mathcal{G}_k)$ called prior cohesions, for $k = 1, \ldots, c$. The cohesion functions measure how likely elements in $\mathcal{G}_k$ are clustered *a priori*. The structure adopted to such cohesions defines the type of PPM we have in mind. Hegarty and Barry (2008) were the first to propose a spatial approach to PPM. They assumed that the prior cohesions of a component $\mathcal{G}_i$ of $\boldsymbol{\pi}$ is a function of the summation (over all areas in $\mathcal{G}_i$) of the number of neighboring areas not in $\mathcal{G}_i$. This cohesion may encourage maps with few contiguous clusters and discourage maps with large number of disconnected ones, which is desirable in regionalization problems. In the spatial PPM introduced by Page and Quintana (2016) the spatial dependence among the neighboring areas is incorporated into the model through both the likelihood and the prior for $\boldsymbol{\pi}$. Four prior cohesions are introduced. Differently from what was considered by Hegarty and Barry (2008), all of them are location dependent and usually dependent on the distance between areas.

2. Preliminary Concepts

Consider n contiguous geographical regions such as those in Figure 2. The map is identified with an undirected graph $\mathcal{G} = (V, E)$, where V is the set of vertices or nodes representing the areas and E is the set of edges connecting pairs of vertices and representing the adjacency relationship among regions (see Figure 2, map on the top left). If there is an edge between vertices i and j we say that the corresponding areas are neighboring areas. A *path* from node v_1 to node v_m is a sequence of nodes $v_1, v_2, \ldots, v_m$ which are connected by edges $(v_1, v_2), \ldots, (v_{m-1}, v_m)$ and with $v_i \neq v_{i+1}$ for $i = 1, \ldots, n$. All vertices are distinct except, possibly, the initial and final vertices v_1 and v_m. This exceptional case, a path with $v_1 = v_m$, is a *circuit*. A graph is said to be *connected* if, for any pair of nodes v_i and v_j, there is at least one path connecting them.

A *spatial cluster* is defined as any subset of nodes forming a connected subgraph. The graph $\mathcal{G}$ is partitioned into c spatial clusters $\mathcal{G}_1, \ldots, \mathcal{G}_c$ if the clusters are disjoint and $\mathcal{G} = \bigcup_i \mathcal{G}_i$, where $1 \leq c \leq n$. A spatial partition $\pi = \{\mathcal{G}_1, \ldots, \mathcal{G}_c\}$ can be viewed as a function from the set $\{1, \ldots, n\}$ that labels the areas into the set $\{1, \ldots, c\}$ of spatial cluster labels.

A *spanning tree* $\mathcal{T}$ of a graph $\mathcal{G}$ is a fundamental concept in our work. It is a connected subgraph with no circuits containing all nodes of $\mathcal{G}$. The second and third maps on the top row in Figure 2 show two spanning trees associated with the first graph. In a spanning tree, any two nodes of $\mathcal{G}$ are connected by a unique path and the number of edges in $\mathcal{T}$ is $n - 1$. This implies that the removal of any $c - 1$ edges from $\mathcal{T}$ partitions the graph $\mathcal{G}$ into c spatial clusters. The last property makes the spanning tree a useful tool for spatial clustering problems. This can be seen in the bottom row where three different partitions are shown. The last two are based on the removal of only four edges from the corresponding spanning trees in the top row.

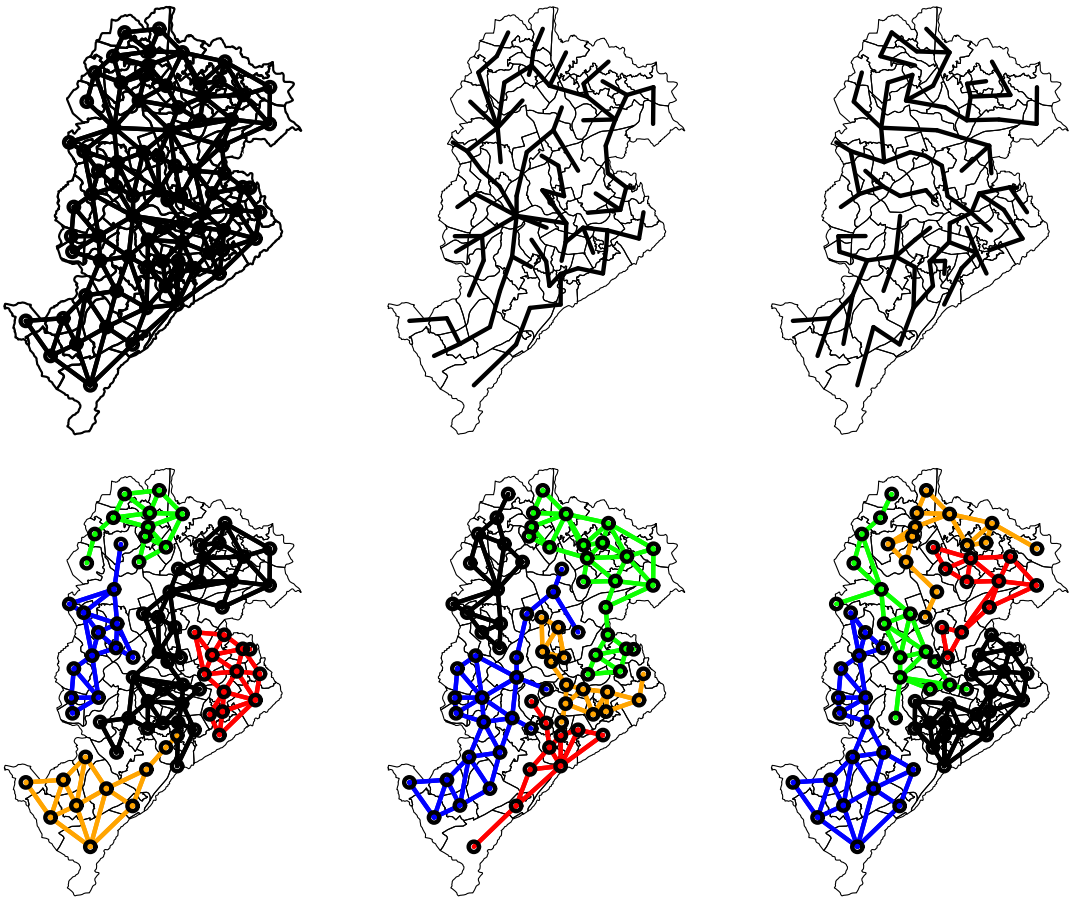


Figure 2: Map of Belo Horizonte city partitioned into 81 administrative planning units superimposed by the adjacency neighborhood graph (first map in the top row) and two spanning trees (other two maps in the top row). The bottom row shows three partitions into five spatial clusters.

Usually, there are many possible spanning trees associated with a given graph. A special kind of spanning tree can be obtained when we associate a cost or weight to each edge. The *cost of a graph* is the sum of its weights. A *minimum spanning tree* is a spanning tree with minimum cost. The minimum spanning tree is not necessarily unique. A sufficient condition for uniqueness is that the pairwise costs are distinct numbers.

Consider a spanning tree $\mathcal{T}$ and a partition π of $\mathcal{G}$ into c disjoint spatial clusters $\mathcal{G}_1, \ldots, \mathcal{G}_c$. We say that π is *compatible* with $\mathcal{T}$ if π can be obtained by pruning $c-1$ edges from $\mathcal{T}$, and we denote this by $\pi \prec \mathcal{T}$. Otherwise, they are incompatible and we write $\pi \not\prec \mathcal{T}$. The extremely large number of possible partitions of a graph is reduced tremendously by considering only those compatible with a given spanning tree. For example, there

are only $n-1$ possible partitions of a particular spanning tree into two spatial clusters while in the original graph this number is of order $O(2^n)$.

These purely spatial graph concepts can be easily extended to the space-time situation. We assume that the same map is observed for T time periods, that is, areas are not created or deleted during the observation period. Hence, we can stack the sequence of maps creating a three-dimensional lattice with nodes indexed by (t, i) where i denotes the geographical unit and t, the time. Edges between nodes (t, i) and (t, j) at the same time t are specified based only on the adjacency between the areas, as described before. Additionally, each node (t, i) is connected to itself and to all its adjacent neighboring regions at time $t + 1$. Since the remaining concepts of spanning trees, paths and others are defined for general graphs, they are also valid for this extended three-dimensional graph $\mathcal{G} = (V, E)$.

3. Spatio-Temporal PPM Driven by Spanning Trees

Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T)$ where $\boldsymbol{Y}_t = (Y_{t1}, \ldots, Y_{tn})$ and $\boldsymbol{\theta}_t = (\boldsymbol{\theta}_{t1}, \ldots, \boldsymbol{\theta}_{tn})$, respectively, denote the observable variables and the vector of parameters for the n regions of a map, at time t, $t = 1, \ldots, T$. We associate $(Y_{tr}, \boldsymbol{\theta}_{tr})$ with the node (t, r) in the graph. If the interest lies only on a purely spatial regionalization we take $T = 1$. Assume that $Y_{tj} \mid \boldsymbol{\theta}_{tj} \overset{ind}{\sim} f(Y_{tj} \mid \boldsymbol{\theta}_{tj})$, $j = 1, \ldots, n$ and $t = 1, \ldots, T$. Let $I = \{1, \ldots, nT\}$ be the set of labels for the nodes.

A major problem when partitioning a data set or a graph is the huge number of possible partitions that compose the search space. To introduce the cluster structure tackling this problem and making feasible the exploration of this space of possible partitions, let us assume a random spanning tree $\mathcal{T}$ selected from the set of all possible spanning trees associated with the graph. Denote by $\boldsymbol{\pi}$ a partition of I compatible with the selected $\mathcal{T}$ and assume that, given $\mathcal{T}$ and the partition $\boldsymbol{\pi} = \{\mathcal{G}_1, \ldots, \mathcal{G}_c\}$, there are common parameters $\boldsymbol{\theta}_{\mathcal{G}_k}$, $k = 1, \ldots, c$, such that, for all nodes with labels belonging to $\mathcal{G}_k$, we have that $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{\mathcal{G}_k}, \quad i \in \mathcal{G}_k$. To establish notation, denote by $\boldsymbol{Y}_{\mathcal{G}_k}$ the set of observations associated with the nodes in $\mathcal{G}_k$. A spatio-temporal PPM induced by random spanning trees is the joint distribution of $(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\pi}, \mathcal{T})$, with $\boldsymbol{\pi} \prec \mathcal{T}$, and denoted by $(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\pi}, \mathcal{T}) \sim$ SPPM, satisfying the following conditions:

(i) Given $\mathcal{T}$, $\boldsymbol{\pi} = \{\mathcal{G}_1, \ldots, \mathcal{G}_c\} \prec \mathcal{T}$ and $\boldsymbol{\theta}_{\mathcal{G}_1}, \ldots, \boldsymbol{\theta}_{\mathcal{G}_c}$, the observations $\boldsymbol{Y}_{\mathcal{G}_1}, \ldots, \boldsymbol{Y}_{\mathcal{G}_c}$ are independent and such that

$$Y_i \mid \boldsymbol{\theta}_{\mathcal{G}_k} \overset{iid}{\sim} f(Y_i \mid \boldsymbol{\theta}_{\mathcal{G}_k}),\ \forall\ i \in \mathcal{G}_k\,;$$

(ii) Given $\mathcal{T}$ and $\boldsymbol{\pi} = \{\mathcal{G}_1, \ldots, \mathcal{G}_c\} \prec \mathcal{T}$, the common parameters $\boldsymbol{\theta}_{\mathcal{G}_1}, \ldots, \boldsymbol{\theta}_{\mathcal{G}_c}$ are independent with joint distribution given by

$$\boldsymbol{\theta}_{\mathcal{G}_1}, \ldots, \boldsymbol{\theta}_{\mathcal{G}_c} \mid \boldsymbol{\pi}, \mathcal{T} \sim \prod_{k=1}^{c} f(\boldsymbol{\theta}_{\mathcal{G}_k})\,;$$

(iii) Given $\mathcal{T}$, the prior distribution of $\boldsymbol{\pi} \prec \mathcal{T}$ is a product distribution such that, for each partition $\{\mathcal{G}_1, \ldots, \mathcal{G}_c\}$ for $c \in \{1, \ldots, nT\}$, **we have**

$$\mathbb{P}(\boldsymbol{\pi} = \{\mathcal{G}_1, \ldots, \mathcal{G}_c\} \mid \mathcal{T}) = \frac{\prod_{k=1}^{c} \kappa(\mathcal{G}_k)}{\sum_{\mathcal{G}'_k \in \mathcal{C}(\mathcal{T})} \prod_{k=1}^{c} \kappa(\mathcal{G}'_k)}\,, \tag{1}$$

where $\kappa(\mathcal{G}_k) \geq 0$ denotes the prior cohesion associated to the subgraph $\mathcal{G}_k$ and represents how likely elements in $\mathcal{G}_k$ are clustered a priori. The summation in the denominator is over all the elements in the set $\mathcal{C}(\mathcal{T})$ which represents all the $2^{nT-1}-1$ partitions compatible with the specific spanning tree $\mathcal{T}$ in which we are conditioning.

(iv) The prior distribution of the spanning tree $\mathcal{T}$ is $\mathbb{P}(\mathcal{T})$ with the support on the set of all possible spanning trees of $\mathcal{G}$.

Usually, there is very little prior information to guide our choice of this last distribution. Hence, we assume that $\mathcal{T}$ is uniformly distributed over the space of spanning trees of the original graph.

The partition of a graph consists simply of the removal of a set of edges resulting in disjoint connected subgraphs. By conditioning on a spanning tree, this task is substantially simplified. We have only $nT-1$ edges and the removal of any $c-1$ of them immediately partition the original graph into c space-times clusters. We want to emphasize that we do not prune a single and fixed spanning tree. Rather, we randomly walk on the space of all spanning trees and, at each step, we generate a new partition. The final inference about the partition is obtained by integrating out over the spanning tree space.

With this motivation, we specify the prior cohesions as functions of ρ_e, the probability of removing a given edge e. For instance, if all edges have a common probability ρ, the prior cohesion related to group $\mathcal{G}_k$ becomes

$$\kappa(\mathcal{G}_k) = \begin{cases} (1-\rho)^{n_{\mathcal{G}_k}-1}\rho, & \text{if } k < c \\ (1-\rho)^{n_{\mathcal{G}_k}-1}, & \text{if } k = c, \end{cases} \tag{2}$$

where $n_{\mathcal{G}_k}$ is the number of remaining edges in $\mathcal{G}_k$. This prior cohesion is similar to that one considered by Barry and Hartigan (1993) to analyze problems involving the identification of contiguous clusters, such as in change point analysis in time series. In that context, this structure is a consequence of the Markovian behavior usually assumed for the change points.

Considering the prior cohesions in (2), given $\mathcal{T}$ and ρ, the prior probability of $\boldsymbol{\pi} = \{\mathcal{G}_1, \ldots, \mathcal{G}_c\}$ is given by

$$\mathbb{P}(\boldsymbol{\pi} \mid \mathcal{T},\ \rho) = \begin{cases} \rho^{(c-1)}(1-\rho)^{(nT-c)}, & \text{if } \boldsymbol{\pi} \prec \mathcal{T} \\ 0, & \text{otherwise.} \end{cases}$$

It is more relevant to obtain the unconditional prior distributions for the partition $\boldsymbol{\pi}$ and for the number C of clusters. Let $N_{\mathcal{T}}$ be the total number of spanning trees associated with $\mathcal{G}$ and $N_{\mathcal{T}}(\boldsymbol{\pi})$ the total number of spanning trees compatible with a partition π. It follows that, given ρ, the prior distributions for $\boldsymbol{\pi}$ and C are given respectively by

$$\begin{aligned} \mathbb{P}(\boldsymbol{\pi} \mid \rho) &= \sum_{\mathcal{T}} \mathbb{P}(\boldsymbol{\pi} \mid \mathcal{T},\ \rho)\mathbb{P}(\mathcal{T}) \\ &= \sum_{\mathcal{T}} \rho^{(c-1)}(1-\rho)^{(nT-c)}\mathbb{P}(\mathcal{T})\ I[\boldsymbol{\pi} \prec \mathcal{T}] \\ &= \rho^{(c-1)}(1-\rho)^{(nT-c)}\frac{N_{\mathcal{T}}(\boldsymbol{\pi})}{N_{\mathcal{T}}} \end{aligned}$$

and

$$\mathbb{P}(C = c \mid \rho) = \sum_{\boldsymbol{\pi}} I[\boldsymbol{\pi}, c]\mathbb{P}(\boldsymbol{\pi} \mid \rho) = \binom{nT-1}{c-1}\rho^{(c-1)}(1-\rho)^{(nT-c)}\frac{N_{\mathcal{T}}(\boldsymbol{\pi})}{N_{\mathcal{T}}}$$

where the last sum is over all partitions $\boldsymbol{\pi}$ and $I[\boldsymbol{\pi}, c]$ is an indicator function assuming 1, if the partition $\boldsymbol{\pi}$ has c clusters, and 0, otherwise.

Assuming *a priori* that $\mathbb{P}(\rho = 1/2) = 1$, it becomes clear the dependence of the prior distribution for $\boldsymbol{\pi}$ on the graph topology since $\mathbb{P}(\pi) = 2^{-nT+1}N_{\mathcal{T}}(\boldsymbol{\pi})N_{\mathcal{T}}^{-1}$. As $N_{\mathcal{T}}(\boldsymbol{\pi})$ depends on $\boldsymbol{\pi}$, we have that $\mathbb{P}(\boldsymbol{\pi})$ is not uniform over all possible partitions. As a consequence, the prior distribution of the number of clusters C is also influenced by the graph structure. Figure 3 shows a simple graph with three different partitions. The left hand side partition is given by $\pi_1 = \{\mathcal{G}_1, \mathcal{G}_2\} = \{\{1, 2, 3\}, \{4, 5, 6\}\}$ and $N_{\mathcal{T}}(\boldsymbol{\pi}_1) = 9$. The middle partition is $\pi_2 = \{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\} = \{\{1, 2, 3\}, \{4, 5\}, \{6\}\}$ and $N_{\mathcal{T}}(\boldsymbol{\pi}_2) = 6$. The number $N_{\mathcal{T}}(\boldsymbol{\pi})$ does not depend only on the number c of spatial clusters in $\boldsymbol{\pi}$. The right hand side partition is $\pi_3 = \{\mathcal{G}_1, \mathcal{G}_2\} = \{\{1, 2, 3, 4, 5\}, \{6\}\}$, with the same number of clusters as $\boldsymbol{\pi}_1$ but with $N_{\mathcal{T}}(\boldsymbol{\pi}_3) = 6$.

It will be more common to assume a prior distribution for ρ such as Beta(r, s). Then

$$\begin{aligned}
\mathbb{P}(\boldsymbol{\pi}) &= \frac{N_{\mathcal{T}}(\boldsymbol{\pi})}{N_{\mathcal{T}}}\frac{\Gamma(r+s)\Gamma(r+nT-c)\Gamma(s+c-1)}{\Gamma(r)\Gamma(s)\Gamma(nT+r+s-1)} \\
\mathbb{P}(C = c) &= \binom{nT-1}{c-1}\frac{N_{\mathcal{T}}(\boldsymbol{\pi})}{N_{\mathcal{T}}}\frac{\Gamma(r+s)\Gamma(r+nT-c)\Gamma(s+c-1)}{\Gamma(r)\Gamma(s)\Gamma(nT+r+s-1)}\,.
\end{aligned}$$

Although the prior distribution for C is dependent on the graph topology, the expected number of clusters *a priori* is not. Given $\mathcal{T}$ and ρ, the number of removed edges from the

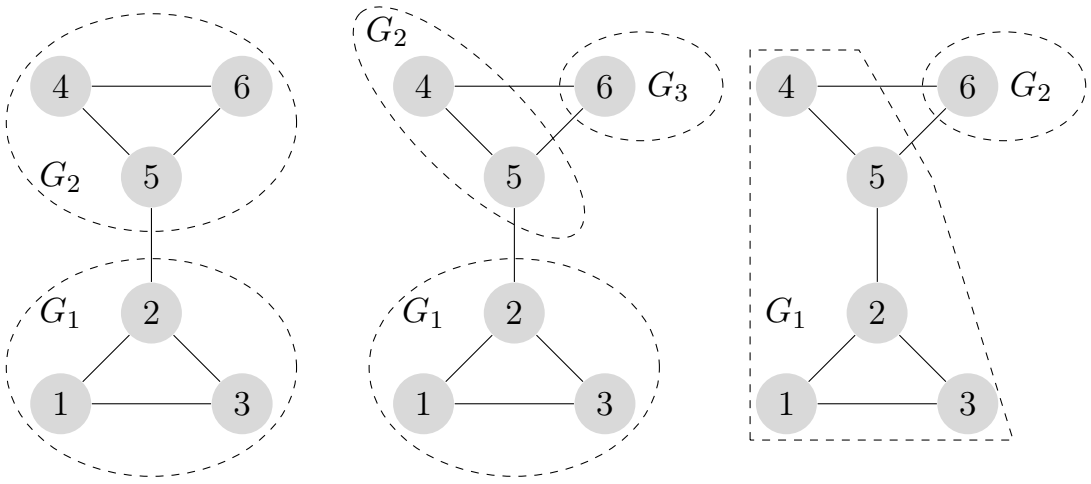


Figure 3: A graph composed by six nodes and with three different partitions $\boldsymbol{\pi}_1$ (left), $\boldsymbol{\pi}_2$ (middle), and $\boldsymbol{\pi}_3$ (right). The number of trees compatible with each partition is $N_{\mathcal{T}}(\boldsymbol{\pi}_1) = 9$, and $N_{\mathcal{T}}(\boldsymbol{\pi}_2) = N_{\mathcal{T}}(\boldsymbol{\pi}_3) = 6$.

tree compatible with the partition has a Binomial distribution with parameters $nT-1$ and ρ. Consequently, the expected number C of clusters is

$$\mathbb{E}(C) = \mathbb{E}\left(\mathbb{E}(C \mid \mathcal{T}, \rho)\right) = \mathbb{E}\left((nT-1)\rho + 1\right) = (nT-1)E(\rho) + 1\,. \tag{3}$$

As (3) reveals, the prior distribution of ρ determines how many clusters we expect *a priori*. A larger value for $\mathbb{E}(\rho)$ would indicate that we expect a high number of clusters, while a smaller value has the opposite effect. Note that the choice of a non informative uniform prior for ρ would stimulate the partition of the graph into a large number of clusters, around $\mathbb{E}(C) = (nT+1)/2$.

4. Sampling Partitions Using Spanning Trees

The great advantage of removing edges from spanning trees for regionalization is the reduction of the search space of partitions. Originally, we have to move through the huge space of all partitions of the graph avoiding those that do not respect the spatial constraint. By pruning the spanning trees, the search space of partitions is drastically reduced and all partitions generated in this way will, by definition, respect the spatial constraint. After all, the spanning tree is built from the neighborhood structure and, therefore, the clusters formed will be spatially or spatio-temporally connected.

To sample from the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{\pi}, \mathcal{T}, \rho \mid \boldsymbol{Y})$, we introduce an MCMC algorithm. The full conditional distributions of $\boldsymbol{\theta}_i$, $i = 1, \ldots, nT$, ρ are, respectively, given by:

$$\begin{aligned} \boldsymbol{\theta}_i \mid \boldsymbol{\pi}, \mathcal{T}, \rho, \boldsymbol{Y} &\sim \mathbb{P}(\boldsymbol{\theta}_{\mathcal{G}^*} \mid \boldsymbol{Y}_{\mathcal{G}^*}), \;\; \text{if} \;\; i \in \mathcal{G}^*, \\ \rho \mid \boldsymbol{\theta}, \boldsymbol{\pi}, \mathcal{T}, \boldsymbol{Y} &\sim \text{Beta}(r + c - 1, s + nT - c)\,. \end{aligned}$$

The full conditional distribution for $\mathcal{T}$ assuming a general prior cohesion is given by

$$\mathbb{P}(\mathcal{T} \mid \boldsymbol{\theta}, \boldsymbol{\pi}, \mathcal{T}, \boldsymbol{Y}) \propto \begin{cases} \left(\sum_{\mathcal{C}(\mathcal{T})} \prod_{k=1}^{c} \kappa(\mathcal{G}_k)\right)^{-1}, & \text{if } \boldsymbol{\pi} \prec \mathcal{T} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Assuming the prior cohesion specified in (2), this full conditional distribution assumes the simple uniform distribution

$$\mathbb{P}(\mathcal{T} \mid \boldsymbol{\theta}, \boldsymbol{\pi}, \mathcal{T}, \boldsymbol{Y}) \propto \begin{cases} 1, & \text{if } \boldsymbol{\pi} \prec \mathcal{T} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

because, in this case, the denominator in (1) is constant and equal to 1 for any spanning tree $\mathcal{T}$:

$$\sum_{\mathcal{C}(\mathcal{T})} \prod_{k=1}^{c} \kappa(\mathcal{G}_k) = \sum_{c=1}^{nT} \binom{nT-1}{c-1} \rho^{(c-1)} (1-\rho)^{(nT-c)} = 1\,.$$

A more efficient algorithm that has a fast convergence is obtained when we integrate out $\boldsymbol{\theta}$ and ρ from the full conditional distribution $\mathbb{P}(\boldsymbol{\pi} \mid \boldsymbol{\theta}, \mathcal{T}, \rho, \boldsymbol{Y})$ thus obtaining

$$\mathbb{P}(\boldsymbol{\pi} \mid \boldsymbol{Y}, \mathcal{T}) = \begin{cases} f_{\mathcal{T}}(\boldsymbol{Y}) \frac{\Gamma(r+s)\Gamma(s+nT-c)\Gamma(r+c-1)}{\Gamma(r+s+nT-1)\Gamma(r)\Gamma(s)}, & \text{for } \pi \prec \mathcal{T} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $f_{\mathcal{T}}(\boldsymbol{Y}) = \prod_{k=1}^{c} \int_{\Theta} f(\boldsymbol{Y}_{\mathcal{G}_k} \mid \theta_{\mathcal{G}_k}) f(\theta_{\mathcal{G}_k}) \, \mathrm{d}\theta_{\mathcal{G}_k}$. We reinforce that $f_{\mathcal{T}}(\boldsymbol{Y})$ is calculated only for partitions $\boldsymbol{\pi}$ compatible with a given spanning tree $\mathcal{T}$. The posterior distribution for the partition $\boldsymbol{\pi}$ should be obtained by marginalizing over all spanning trees compatible with a given partition: $\mathbb{P}(\boldsymbol{\pi} \mid \boldsymbol{Y}) = \sum_{\mathcal{T} \prec \boldsymbol{\pi}} \mathbb{P}(\boldsymbol{\pi} \mid \boldsymbol{Y}, \mathcal{T}) \, \mathbb{P}(\mathcal{T} \mid \boldsymbol{Y})$. This would imply the need to additionally derive the posterior $\mathbb{P}(\mathcal{T} \mid \boldsymbol{Y})$. Rather than following this approach, we decide to adopt a different strategy, by sampling from the joint posterior of compatible pairs of $\mathcal{T}$ and $\boldsymbol{\pi}$. The challenge is to sample partitions and the spanning trees that are compatible to each other. The proposed strategy to achieve this is presented in the following.

4.1. Sampling a Partition Compatible with the Current Tree

The transformation suggested by Barry and Hartigan (1993) inspired the Gibbs sampler we describe here. A partition $\boldsymbol{\pi}$ of the graph can be transformed into a vector $\boldsymbol{U}$ of $nT - 1$ dependent binary variables, given a compatible spanning tree $\mathcal{T}$. The coordinate U_i of $\boldsymbol{U}$ is 1 if the i-th edge is not removed from the tree to form $\boldsymbol{\pi}$, and 0, otherwise. With this multidimensional random variable, we can sample from the posterior of $\boldsymbol{\pi}$ using Gibbs sampler. Let $\boldsymbol{U}_{-i} = \{U_1, \cdots, U_{i-1}, U_{i+1}, \cdots, U_{nT-1}\}$. To decide if each edge in the tree should be removed or not, it is sufficient to know the ratio between the $\mathbb{P}(U_i = 1 \mid \boldsymbol{U}_{-i}, \mathcal{T}, \boldsymbol{Y})$ and $\mathbb{P}(U_i = 0 \mid \boldsymbol{U}_{-i}, \mathcal{T}, \boldsymbol{Y})$. For the model assumed here such ratio is given by

$$R_i = \frac{f_{\mathcal{T}}^{(1)}\left(\boldsymbol{Y}_{\mathcal{G}_k}\right)}{f_{\mathcal{T}}^{(0)}\left(\boldsymbol{Y}_{\mathcal{G}_k}^{(L)}\right) f_{\mathcal{T}}^{(0)}\left(\boldsymbol{Y}_{\mathcal{G}_k}^{(R)}\right)} \frac{(r + c - 2)}{(s + nT - c)} \,,$$

where $f_{\mathcal{T}}^{(1)}\left(Y_{\mathcal{G}_k}\right)$ is the prior predictive of the group formed when the edge i is present, whereas $f_{\mathcal{T}}^{(0)}\left(\boldsymbol{Y}_{\mathcal{G}_k}^{(L)}\right)$ and $f_{\mathcal{T}}^{(0)}\left(\boldsymbol{Y}_{\mathcal{G}_k}^{(R)}\right)$ are the predictive distributions for the observations of the two groups formed when the edge i is removed from the tree $\mathcal{T}$ pruned by $\boldsymbol{U}_{-i}$.

Thus, we can sample from the distribution of U_i simply by sampling a uniform value $u \sim \text{Uniform}(0, 1)$ and using the following accept/reject criterion:

$$U_i = \begin{cases} 1, & \text{if } R_i \geq \frac{u}{1-u} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

4.2. Sampling a tree compatible with the current partition

We need to sample trees that are compatible with the current partition. The valid trees are uniformly distributed over the subset of trees compatible with the partition as shown in (5). The generation of uniformly distributed spanning trees **without constraints** is a subject of study since 1989 (Broder, 1989). Wilson algorithm (Wilson, 1996) is the main algorithm in the literature and it is based on a certain random walk on the adjacency graph. Under the constraint that the generated trees must be compatible with the current partition at each step of the Gibbs sampler, Wilson's algorithm could be used using a rejection sampling approach. Uniformly select a spanning tree from the adjacency graph $\mathcal{G}$ and reject it if it is not compatible with the current partition. Otherwise, accept it. This is a very expensive way to generate uniformly from the set of spanning trees compatible with $\boldsymbol{\pi}$. For example, suppose that $\mathcal{G}$ is a regular 26×26 rectangular lattice split into two spatial clusters, the

upper half and the lower half. We did not obtain a single compatible spanning tree after uniformly generating 3.8 million spanning trees. Another example is the Brazilian map partitioned into 5564 municipalities (see Figure 7a), again split only into two groups, one being the municipalities within the three most Southern states of Brazil (see Figure 13 for the states' boundaries). This partitioning of the map has 67 edges connecting municipalities of different groups. After 1.5 million simulations by Wilson's method, none was compatible with that partition.

Another possibility is to adapt Wilson's algorithm to obey the constraints imposed by the current partition. A reviewer suggested the following procedure. Suppose a clustering as in Figure 3 (middle). First sample a uniform spanning tree by Wilson's algorithm within each of $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$. Next, define a collapsed graph where the nodes are the three clusters and we set an edges between $\mathcal{G}_i$ and $\mathcal{G}_j$ if there is at least one edge between the two subgraphs. Sample a uniform spanning tree on that collapsed graph. Finally, for each edge sampled in the collapsed graph, sample one of the edge uniformly in the original bipartite graph.

While this procedure may sound promising at first, it, unfortunately fails to generate trees uniformly. What is problematic in this approach is that the number of edges between clusters can be imbalanced, which could lead to trees being sampled with different probability.

Consider the graph in Figure 4. We would first sample a uniform tree inside of each of the tree groups. In this example, there are 3 independent options for each group, so in this first step a choice is made with probability $\left(\frac{1}{3}\right)^3 = \frac{1}{27}$. Now, the collapsed graph is a complete graph of size 3 (a triangle). Once again, there is a choice between three spanning trees. Assume we uniformly select a spanning tree which contains an edge between Group A and Group B and between Group B and Group C. This choice has probability $\frac{1}{3}$. Finally, we need to uniformly choose among the corresponding edges between the clusters. For the connection between Group B and Group C, there is only one option, while for Group A and Group B, there are three choices. Uniformly selecting the edges, thus, will have a probability of $\frac{1}{3} \cdot 1 = \frac{1}{3}$. The final tree, then, will have been selected with probability $\frac{1}{27}\frac{1}{3}\frac{1}{3} = \frac{1}{247}$.

Consider now another possible combination of choices. For the first step, we select the same trees inside the groups, with probability $\frac{1}{27}$. For the second step, we choose the spanning tree of the collapsed graph which contains a connection between Group A and Group C and between Group C and Group B. This step, again, has probability $\frac{1}{3}$. Finally, to select the corresponding trees in the original bipartite graph, there is only one choice corresponding to each connection, namely (B2, C3) and (A2, C1). Therefore, the choice in this final step was made with probability 1. The final tree, then, was chosen with probability $\frac{1}{27}\frac{1}{3}1 = \frac{1}{81}$. Therefore, this method will sample the trees with a probability which is not uniform.

Instead of Wilson's algorithm, we employ another procedure to sample trees using a minimum spanning tree (MST) algorithm. We sample a new tree $\mathcal{T}$ compatible with the current partition π by assigning weights to the edges in the graph respecting this current partition in the following way. The edges that connect vertices belonging to the same cluster receive a low weight, obtained from a uniform distribution which generates low values (e.g. between 0 and 1). The edges that connect vertices belonging to different clusters receive a high weight, obtained from a uniform distribution which generates higher values (e.g.

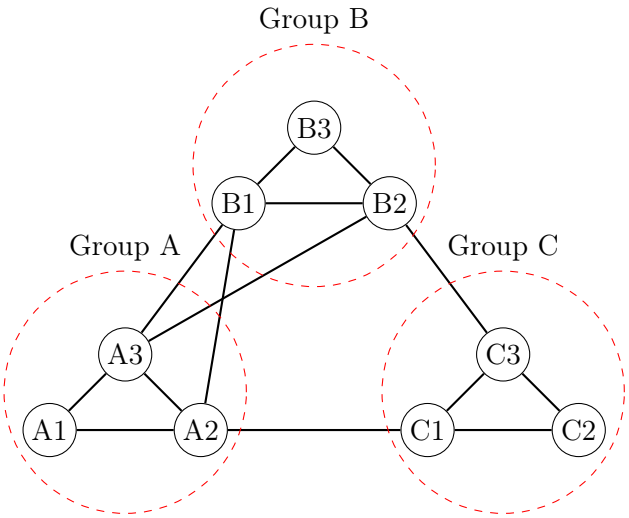


Figure 4: Counter example of graph where the approach of adapting Wilson's algorithm as a multi-step procedure to obey the constraints imposed by the partition fails to sample trees uniformly.

between 10 and 20). Once the weights are assigned, the minimum spanning tree is obtained and it is the new sampled tree, compatible with the current partition.

The reason for using these two sets of values is that the algorithm computes the spanning tree with minimum sum of weights. When we use weights in this way, we ensure that the tree will be compatible with the partition, since the edges with higher weights are added to the tree only when all possible connections through edges with a lower weight are already explored. So, it only adds a connection between clusters when all the possible connections inside a cluster have been visited.

The proposed algorithm works because the MST algorithm, either by Prim algorithm or by the Kruskal algorithm (Cormen et al., 2009), selects the edges according to their weights. Since all edges separating clusters have higher weights, they are selected only after the lower weight edges connecting vertices inside a cluster. The random attribution of weights ensures that the tree will respect the partition and, since we randomly assign weights each time we sample a tree, we get a random tree, even though the algorithm is deterministic.

This procedure does not sample exactly from a uniform distribution. To see this, consider the following example. Consider the graph in the left hand side of Figure 5 where we disregard any partition. There are 5 spanning trees that do not include the $(3-4)$ edge, each one obtained by the removal of $(3-4)$ and one of the remaining edges. There are 6 other spanning trees including the $(3-4)$ edge: we need to prune one of the three edges located at the left-hand side of $(3-4)$ and one of the two located at its right-hand side. Our method would assign independent $U(0,1)$ random variables $W_1, \ldots, W_6$ to the six edges and then select its minimum spanning tree. This is equivalent to substitute the W_i's by their rank statistics $R_1, \ldots, R_6$ where $R_i = \sum_{j=1}^{6} I[W_j \leq W_i]$ and to select the minimum spanning tree using these ranks. The total number of possible ranks assignment to the

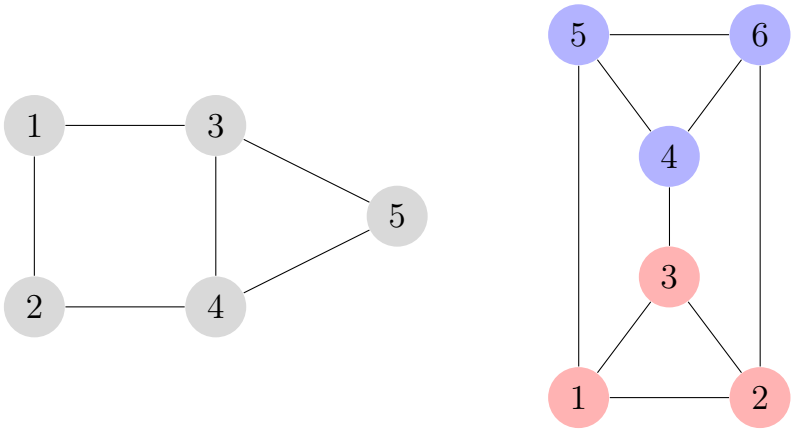


Figure 5: Two graphs to discuss different possibilities to randomly generate spanning trees.

edges is equal to the number of permutations of the six edges, that is, $6! = 720$. Each rank statistic configuration is equally likely and each one of the 11 spanning trees is associated with a subset of these 720 ranking patterns. The cardinality of these subsets should be the same if we have a uniform sampling procedure. However, 11 is not a divisor of 720 and therefore some spanning trees are selected with higher frequency than others.

However, this simple and fast procedure is a good approximation for the uniform sampling of spanning trees and we show this by means of an example. Consider the graph in the right-hand side of Figure 5 with 6 nodes divided into 2 groups (nodes 1-3 and nodes 4-6). A complete enumeration shows that there are 27 spanning trees compatible with this partition, which, if uniformly sampled, should have a selection probability equal to $1/27 \approx 0.03704$. Using Wilson's algorithm, we selected 100 thousand random spanning trees. The observed frequencies of the 27 spanning trees varied from 0.03588 to 0.03831. Using our method based on 100 thousand simulations of $U(0, 1)$ random variables for the edges within cluster and $U(10, 20)$ random variables for the edges between clusters, the observed frequencies of the spanning trees varied from 0.03588 to 0.03811. Additionally, Wilson's algorithm requires generating about 3 spanning trees before one is accepted (compatible with the partition), even in this small example, while for our approach each generated tree is guaranteed to be compatible with the partition.

5. Analysis of Simulated Data Sets

In this section, we evaluate the performance of our model for regionalization in spatial settings only. The main goal is to compare it with the following state-of-art regionalization techniques: *SKATER*, through its implementation on the `spdep` R package[1], the original Automatic Zoning Procedure (*AZP*), and its variations with Simulated Annealing (*AZP_SA*), Tabu (*AZP_TABU*), Reactive tabu (*AZP_RTABU*), the Automatic Regionalization with Initial Seed Location (*ARISEL*), the Max-p-regions Tabu model (*MAXP*) and the "A Multidirectional Optimum Ecotope-Base Algorithm" (*AMOEBA*), all implemented

1. `http://cran.r-project.org/package=spdep`

in the Python ClusterPy[2] library (Duque et al., 2011). For Poisson data sets, the proposed spatial PPM is also compared to the BDCD model, implemented in C++ code (available at `http://www.statistik.lmu.de/sfb386/software/bdcd/download.html`) and to BPM, implemented in R software by ourselves.

To define the spatial structure of the simulated data (coordinates, shapes and spatial adjacency), we consider two geographical neighborhood structures of Brazilian regions. In the Gaussian data sets, inspired by the human development index map shown in Section 6, we considered 853 Brazilian regions grouped into three clusters: two large clusters and a third small cluster, composed by 10 regions and islanded in the middle of one of the other two clusters. The small cluster mean is very different from that in surrounding regions. In Scenario 1, the observations within each cluster are *iid* generated from normal distributions. To mimic the observed data, we choose means $\mu_i = 0.73$ and precision $\tau_i = 4000$, for the small cluster, $\mu_i = 0.63$ and $\tau_i = 754.20$, for the large cluster containing the small one, and $\mu_i = 0.70$ and $\tau_i = 878.35$, for the other large cluster. In Scenario 2, the spatial structure also presents three clusters but inside each cluster, the distribution varies slightly from a region to another. The regions means within the large cluster containing the smaller one vary from 0.612 up to 0.685 and precision varies between 435 and 484. The other large cluster have $\mu_i \in (0.663, 0.726)$ and $\tau_i \in (471, 497)$. Finally, the small cluster is composed by three areas with common mean equal to 0.780 and precision 10000.

For the Poisson data sets, we were inspired by the epidemiological studies presented in Section 6. We considered smaller areas in the southern region of Brazil, the 1188 Brazilian municipalities divided into clusters. Each observation y_i simulates a cancer death count and the expected count E_i is extracted from the real data analysis and based on true population in each area. Motivated by the high mortality rates observed in the lung cancer data, we simulated Poisson data with high rates in Scenarios 3 and 4, while in Scenarios 5 and 6, we assumed low rates similar to those observed in the bladder cancer data. Rates assumed in high rate scenarios are, on average, 8 times higher than those for the low rate scenarios. Similarly to the normal case, we assume areas with the same rate (Scenarios 3 and 5) and with distinct but similar rates (Scenarios 4 and 6) within each cluster.

In Scenarios 3 and 4, the map is divided into 10 clusters with the common rates in Scenario 3 ranging from 0.45 to 1.40. The distinct rates for the 1188 regions considered in Scenario 4 range from 0.45 to 1.75. In bladder cancer inspired low rate scenarios, the map is divided into 13 clusters. The common rates in Scenario 5 has values varying from 0.24 to 1.74, while for Scenario 6, the distinct parameters for each region vary from 0.77 to 1.42.

Concerning the inference, we assume that $Y_i \mid \mu_{\mathcal{G}_k}, \tau_{\mathcal{G}_k} \overset{iid}{\sim} \text{Normal}\left(\mu_{\mathcal{G}_k}, \tau_{\mathcal{G}_k}^{-1}\right)$, for all $i \in \mathcal{G}_k$, in the normal case. We also consider the normal-gamma prior distribution with $\mu_{\mathcal{G}_k} \mid \tau_{\mathcal{G}_k} \sim \text{Normal}\left(m, [v\tau_{\mathcal{G}_k}]^{-1}\right)$ and $\tau_{\mathcal{G}_k} \sim \text{Gamma}\,(a, b)$, where $m = 0.65$, $v = 1$, $a = 400$ and $b = 1$. For the Poisson data, $Y_i \mid \phi_{\mathcal{G}_k} \overset{iid}{\sim} \text{Poisson}\,(E_i \cdot \phi_{\mathcal{G}_k})$ and $\phi_{\mathcal{G}_k} \sim \text{Gamma}(a, b)$, where $a = b = 2$. This prior distribution for $\phi_{\mathcal{G}_k}$ puts probability mass around 1, as we usually expect for the relative risk. Furthermore, 90% of its probability mass is concentrated between 0.18 and 2.37, which is a reasonable range for many human disease relative risks. As usually done in real problems, in our analysis, the expected count E_i is assumed to

2. `http://www.rise-group.org/`

be known. Besides, we assume that each edge is removed from the spanning tree with probability $\rho \sim \text{Beta}(5, 1000)$.

Our algorithm was implemented in C++ and is available upon request. For the MCMC we run a chain of 5000 iterations, skipping the first 500 samples as a *burn-in* period and, to avoid correlation, we use a thinning of 5 simulated values.

5.1. Evaluation Metrics

We consider two groups of metrics. The first group is based on the estimates of the parameters that index the data distributions. We consider the mean absolute error (MAE), the mean relative error (MRE) and the mean squared error (MSE) based on the distance between the true parameters and their estimates. Under BPM, BDCD and the proposed PPM the parameter estimates in each region are the posterior means. Under the usual regionalization methods, such estimates are obtained by averaging the observations Y_i into the cluster. For normal data, these are given by $\hat{\theta}_i = \sum_{j \in \mathcal{G}_k} Y_j / n_k \quad \forall i \in \mathcal{G}_k$ and, in the Poisson case, they are $\hat{\theta}_i = [\sum_{j \in \mathcal{G}_k} Y_j][\sum_{j \in \mathcal{G}_k} E_j]^{-1}$, $\forall i \in \mathcal{G}_k$.

The other set of metrics is based on the difference between the true clusters used to generate the data and the clusters obtained by the different methods. Each pair of areas is labeled as positive if they are in the same cluster and as negative if they are in different clusters. We looked at the number of true positives (TP) classifications, the number of pairs that are in the same cluster both, in the true partition as well as in the estimated one. False positive classifications (FP) is the number of pairs assigned to distinct clusters but that actually are in the same cluster in the true partition. True negatives (TN) classification is the number of pairs correctly assigned to distinct clusters. Finally, false negatives (FN) classification is the number of pairs incorrectly assigned to distinct clusters. In our evaluation, we consider the rand measure (RAND), the F_1 score (F1), the Jaccard index (JI) and the Fowlkes-Mallows index (FM) which are given, respectively, by

$$\begin{aligned} RAND &= [TP + TN] \cdot [TP + FP + FN + TN]^{-1}, \\ F1 &= 2 \cdot P \cdot R \cdot [P + R]^{-1}, \\ JI &= TP \cdot [TP + FP + FN]^{-1}, \\ FM &= \sqrt{P \cdot R}, \end{aligned}$$

where $P = TP \cdot [TP + FP]^{-1}$ and $R = TP \cdot [TP + FN]^{-1}$.

Most of the traditional regionalization techniques considered in our analysis requires the pre-definition of the number of clusters to be generated as an input. In all cases, we use three different values: the true number c of clusters as well as 3 more and 3 less (1 less for the normal case) clusters than the true number. The max-p-regions model clusters a set of geographic areas into the maximum number of homogeneous regions such that the value of a spatially extensive regional attribute is above a predefined threshold value. In clusterPy we measure heterogeneity as the within-cluster sum of squares from each area to the attribute centroid of its cluster. The parameter values shown in the tables for this method correspond to this threshold.

To summarize the posterior information about the random partition provided by our method, we take the underlying graph of the spatial structure of the data and, for each edge

connecting neighboring areas, we compute how often they are in the same cluster based on the generated partitions. This percentage is assigned to each edge. Then, we trim the edges by removing all those which are below a certain threshold. Once the infrequent edges are removed, the remaining components of the graph define the clusters. The reasoning is that the removed edges are exactly those which are frequently crossing the borders between clusters in the sampled partitions. By removing them, the bulk of the clusters frequently present in the sampled partitions remain connected in the graph. We consider three thresholds: 70%, 80% and 90%. The same strategy is used for the BPM and BDCD methods.

5.2. Results

Tables 1 to 3 show the model fit measures for the proposed model and the competitor methods for normal and Poisson data. For each evaluation measure, the best outcome is underlined and shown in bold.

In almost all the simulated data sets, our model outperformed all the other methods. The only scenarios where our method had inferior results were in the normal data set with common parameter, where the MAE for the *SPPM* was the second best and the Poisson data set with low rate and common parameter where *SKATER* had the lowest MRE. However, in both cases, the *SPPM* had better performance according to all other metrics we consider to evaluate the models. Furthermore, in these two exceptional cases, even in those losing metrics, *SPPM* had a very close value to the better competitors.

In all data sets, particularly in the Poisson data sets, the error metrics (MAE, MSE, MRE) for our method were from 1.5 to 5 times smaller than the other methods. For *SKATER*, *ARISEL* and the *AZP*-type methods, we notice that the bias in the parameter estimates (either the normal mean or the Poisson rate) depends on the number of clusters we assume to implement the method as well as on the type of scenario. We notice, for instance, that *SKATER* produced less biased estimates for such parameters in all scenarios if the assumed number of cluster is $c - 3$ (and $c - 1$ for the normal case), the exception occurring in Scenario 1. Moreover, in Scenario 3, less biased estimates are obtained by all such methods, except *SKATER*, if the number of clusters is correctly assumed. This bias may be explained by the poor capacity of these methods to identify correctly the clusters. The non-stochastic methods were favored in the sense that they received good information about the number of clusters. We either provided the exact number or a close one. In some cases, even with the correct number of clusters, the *AZP*-type methods had a worse performance than with the wrong number of clusters.

Our method also achieved better results if compared to the Bayesian models *BPM* and *BDCD* (only implemented for Poisson models). *BPM* had a very poor performance for both, cluster identification and parameter estimation, in all four Poisson scenarios. *BDCD* and the proposed *SPPM* are comparable with respect to the parameter estimation, with proposed model usually providing only slightly better estimates. However, the proposed model has much better performance for cluster identification, except for Scenario 3 and thresholds 70% and 80%, and Scenarios 4 and 6 and thresholds 70%.

Table 1: Model fit summaries for all methods, Normal data, Scenarios 1 and 2.

Method	Parameter	MAE	MSE	MRE	RAND	F1	JI	FM
				Common parameters				
SPPM	70%	0.00466	**0.00010**	**0.00701**	**96.50**	**96.51**	**93.26**	**96.51**
	80%				96.30	96.26	92.79	96.27
	90%				94.74	94.58	89.72	94.65
SKATER	$c-1$	0.00745	0.00024	0.01132	92.08	92.22	85.56	92.22
	c	0.00503	0.00016	0.00771	94.10	94.08	88.82	94.09
	$c+3$	0.00909	0.00027	0.01389	89.24	88.56	79.47	88.79
AZP	$c-1$	0.03526	0.00125	0.05304	50.34	66.92	50.28	70.87
	c	0.00488	0.00017	0.00751	92.52	92.34	85.77	92.39
	$c+3$	0.01158	0.00035	0.01799	84.98	83.11	71.10	83.85
AZP_SA	$c-1$	0.00488	0.00015	0.00751	96.03	96.07	92.44	96.07
	c	0.01902	0.00071	0.02865	67.15	62.10	45.04	62.93
	$c+3$	0.00849	0.00028	0.01304	81.07	77.92	63.82	79.14
AZP_TABU	$c-1$	0.02972	0.00104	0.04536	60.17	67.43	50.86	68.50
	c	0.00702	0.00021	0.01093	87.63	86.55	76.29	86.95
	$c+3$	0.01171	0.00037	0.01805	86.35	84.84	73.67	85.45
AZP_RTABU	$c-1$	0.02680	0.00108	0.03943	60.49	63.33	46.34	63.47
	c	0.00683	0.00023	0.01056	93.50	93.59	87.96	93.60
	$c+3$	0.00530	0.00022	0.00807	92.39	92.17	85.49	92.23
ARISEL	$c-1$	**0.00456**	0.00015	0.00701	96.02	96.07	92.43	96.07
	c	0.00945	0.00024	0.01481	88.77	87.92	78.45	88.23
	$c+3$	0.00542	0.00025	0.00824	86.91	85.72	75.01	86.14
AMOEBA	*None*	0.02497	0.00074	0.03770	70.89	64.90	48.04	66.45
MAXP	10	0.01243	0.00036	0.01879	51.45	7.89	4.10	19.15
	100	0.01551	0.00058	0.02317	65.74	55.54	38.44	58.36
				Distinct parameters				
SPPM	70%	**0.00705**	**0.00015**	**0.01065**	50.37	66.58	49.90	70.32
	80%				**90.97**	**90.84**	**83.22**	**90.86**
	90%				90.33	89.63	81.21	89.92
SKATER	$c-1$	0.01086	0.00037	0.01624	87.45	87.55	77.85	87.55
	c	0.01131	0.00039	0.01702	86.06	85.93	75.33	85.94
	$c+3$	0.01263	0.00041	0.01927	82.92	81.75	69.14	81.99
AZP	$c-1$	0.01213	0.00046	0.01822	85.08	85.40	74.51	85.40
	c	0.03308	0.00130	0.04894	52.96	59.88	42.74	60.49
	$c+3$	0.01162	0.00043	0.01741	86.41	86.51	76.23	86.51
AZP_SA	$c-1$	0.01102	0.00037	0.01651	87.88	88.07	78.69	88.08
	c	0.01256	0.00049	0.01891	73.41	69.46	53.21	70.33
	$c+3$	0.01281	0.00053	0.01931	76.11	74.31	59.12	74.57
AZP_TABU	$c-1$	0.02277	0.00113	0.03506	65.63	69.11	52.80	69.42
	c	0.01512	0.00066	0.02280	78.56	78.98	65.27	78.99
	$c+3$	0.01566	0.00055	0.02374	73.43	68.39	51.96	69.77
AZP_RTABU	$c-1$	0.01386	0.00054	0.02068	83.51	83.73	72.02	83.73
	c	0.01317	0.00050	0.01978	84.15	84.15	72.63	84.15
	$c+3$	0.01394	0.00051	0.02076	79.09	77.26	62.95	77.61
ARISEL	$c-1$	0.01206	0.00044	0.01808	85.87	86.12	75.62	86.12
	c	0.01149	0.00040	0.01722	87.01	87.17	77.26	87.17
	$c+3$	0.01477	0.00047	0.02199	77.31	73.55	58.16	74.69
AMOEBA	*None*	0.03450	0.00144	0.05156	61.44	52.00	35.14	53.76
MAXP	10	0.01180	0.00034	0.01751	51.58	8.63	4.51	19.86
	100	0.01774	0.00053	0.02647	60.15	43.46	27.76	48.12

Table 2: Model fit summaries for all methods, Poisson data with high rate, Scenarios 3 and 4.

Method	Parameter	MAE	MSE	MRE	RAND	F1	JI	FM
				Common parameters				
SPPM	70%	**0.03054**	**0.00422**	**0.03797**	49.14	42.08	26.65	51.24
	80%				76.54	59.65	42.50	63.71
	90%				**90.59**	**74.28**	**59.09**	**74.32**
SKATER	$c-3$	0.11544	0.02046	0.12527	74.76	54.92	37.86	58.11
	c	0.12579	0.03044	0.13527	75.50	54.12	37.10	56.62
	$c+3$	0.12618	0.03409	0.12778	77.03	54.06	37.04	55.77
AZP	$c-3$	0.14481	0.03003	0.15500	69.35	43.77	28.01	46.01
	c	0.10819	0.02484	0.11362	74.69	57.25	40.11	61.45
	$c+3$	0.13351	0.02751	0.14128	79.76	51.97	35.11	52.26
AZP_SA	$c-3$	0.17650	0.04653	0.21710	51.07	42.05	26.62	50.49
	c	0.07600	0.01620	0.08459	76.36	51.36	34.55	52.70
	$c+3$	0.11834	0.03139	0.12526	74.29	47.72	31.34	49.07
AZP_TABU	$c-3$	0.19176	0.06519	0.23379	49.84	36.52	22.34	42.84
	c	0.10617	0.02122	0.11570	73.60	51.01	34.23	53.47
	$c+3$	0.13359	0.02907	0.14364	81.11	48.89	32.35	48.90
AZP_RTABU	$c-3$	0.18137	0.05388	0.22075	53.29	41.53	26.20	48.90
	c	0.10348	0.02615	0.11161	78.00	47.37	31.03	47.60
	$c+3$	0.13224	0.03324	0.14001	78.60	50.97	34.20	51.45
ARISEL	$c-3$	0.13210	0.02655	0.14694	66.90	47.90	31.49	52.43
	c	0.10389	0.02524	0.11519	73.19	55.76	38.66	60.23
	$c+3$	0.12787	0.03623	0.13810	77.63	46.42	30.22	46.64
AMOEBA	*None*	0.16781	0.06013	0.18080	69.82	37.53	23.10	38.48
MAXP	10	0.08963	0.01851	0.10448	81.91	10.06	5.30	20.76
	100	0.12921	0.02821	0.15403	82.93	46.76	30.52	47.50
BPM	70%	0.35841	0.30151	0.53348	49.90	41.00	25.79	49.33
	80%				49.90	41.00	25.79	49.33
	90%				49.90	41.00	25.79	49.33
BDCD	70%	0.03637	0.00528	0.04562	74.61	57.73	40.58	62.21
	80%				83.63	61.93	44.85	62.41
	90%				82.33	14.34	7.72	25.57
				Distinct parameters				
SPPM	70%	**0.04759**	**0.00544**	**0.05465**	51.13	42.04	26.62	50.46
	80%				**89.02**	**70.40**	**54.32**	**70.41**
	90%				88.95	65.78	49.01	66.73
SKATER	$c-3$	0.09821	0.02133	0.10367	82.90	62.00	44.93	62.83
	c	0.10061	0.02474	0.10698	84.04	61.96	44.89	62.28
	$c+3$	0.10650	0.02619	0.11220	85.39	62.13	45.06	62.15
AZP	$c-3$	0.11725	0.02275	0.12573	72.38	46.57	30.35	48.37
	c	0.09930	0.01941	0.10935	71.55	50.88	34.12	54.29
	$c+3$	0.11223	0.02456	0.12804	74.61	43.48	27.78	44.05
AZP_SA	$c-3$	0.12101	0.02373	0.14110	70.97	39.75	24.81	40.73
	c	0.09564	0.01968	0.10301	76.76	56.78	39.65	59.50
	$c+3$	0.07958	0.01855	0.08658	83.88	56.34	39.22	56.35
AZP_TABU	$c-3$	0.10113	0.01917	0.11041	69.77	50.18	33.49	54.22
	c	0.14293	0.03248	0.17143	64.21	36.72	22.49	38.95
	$c+3$	0.09638	0.01827	0.10530	76.34	46.68	30.45	47.22
AZP_RTABU	$c-3$	0.11068	0.01971	0.12723	67.98	48.34	31.87	52.53
	c	0.10050	0.02104	0.11083	68.54	49.37	32.78	53.70
	$c+3$	0.12937	0.02724	0.14326	78.26	44.22	28.39	44.25
ARISEL	$c-3$	0.09152	0.01625	0.09966	76.59	48.94	32.40	49.74
	c	0.11073	0.02656	0.11981	79.64	51.77	34.93	52.07
	$c+3$	0.09866	0.02000	0.10814	76.19	51.87	35.01	53.40
AMOEBA	*None*	0.17683	0.05720	0.19835	67.94	35.11	21.29	36.13
MAXP	10	0.09528	0.01730	0.10998	81.93	10.68	5.64	21.19
	100	0.13384	0.02775	0.16170	80.19	41.26	26.00	41.56
BPM	70%	0.35298	0.26530	0.51927	50.58	37.93	23.40	44.61
	80%				50.58	37.93	23.40	44.61
	90%				50.61	37.87	23.36	44.52
BDCD	70%	0.05641	0.00676	0.06367	67.69	46.47	30.27	50.13
	80%				83.00	29.55	17.34	35.75
	90%				82.76	16.59	9.04	29.32

Table 3: Model fit summaries for all methods, Poisson data with low rate, Scenarios 5 and 6.

Method	Parameter	MAE	MSE	MRE	RAND	F1	JI	FM
				Common parameters				
SPPM	70%	**0.07669**	**0.01905**	0.49370	65.43	78.99	65.27	80.78
	80%				65.98	79.17	65.52	80.87
	90%				**79.50**	**82.46**	**70.16**	**82.99**
SKATER	$c-3$	0.22172	0.12678	**0.36023**	49.73	52.20	35.32	53.72
	c	0.26417	0.16666	0.41019	46.72	47.16	30.85	49.27
	$c+3$	0.29476	0.23519	0.44595	45.43	44.73	28.81	47.16
AZP	$c-3$	0.24460	0.14382	0.74205	44.84	47.85	31.45	49.16
	c	0.23630	0.15310	0.60011	56.83	58.58	41.42	60.46
	$c+3$	0.27275	0.21656	0.78907	43.98	43.27	27.61	45.62
AZP_SA	$c-3$	0.22391	0.11542	0.74414	45.51	47.05	30.76	48.77
	c	0.18481	0.14430	0.66085	49.64	55.47	38.38	56.10
	$c+3$	0.17373	0.14279	0.67033	44.04	41.78	26.40	44.65
AZP_TABU	$c-3$	0.18362	0.12962	0.63674	51.98	60.42	43.29	60.58
	c	0.30429	0.18741	0.46918	45.31	38.46	23.81	43.40
	$c+3$	0.27051	0.19779	0.85028	47.51	45.23	29.23	48.42
AZP_RTABU	$c-3$	0.27233	0.17007	0.78498	44.19	44.95	28.99	46.86
	c	0.30857	0.20849	0.81854	43.09	43.67	27.93	45.59
	$c+3$	0.30699	0.21135	0.86227	43.84	35.09	21.28	40.60
ARISEL	$c-3$	0.29914	0.19268	0.81843	43.80	43.72	27.97	45.86
	c	0.28959	0.18852	0.43265	45.82	43.18	27.53	46.35
	$c+3$	0.30599	0.22116	0.88723	41.76	33.93	20.43	38.56
AMOEBA	*None*	0.53662	0.56779	0.66286	42.83	33.89	20.40	39.23
MAXP	10	0.16617	0.05139	0.60488	36.08	3.69	1.88	13.04
	100	0.12539	0.03611	0.68845	39.68	20.78	11.59	29.41
BPM	70%	0.18454	0.05587	0.67587	65.02	78.80	65.02	80.63
	80%				64.61	78.47	64.56	80.24
	90%				46.72	58.23	41.08	58.24
BDCD	70%	0.09166	0.01986	0.51715	63.98	77.84	63.72	79.44
	80%				36.77	5.67	2.92	16.60
	90%				35.02	0.12	0.06	2.44
				Distinct parameters				
SPPM	70%	0.05172	0.00532	0.05553	65.02	78.80	65.02	80.63
	80%				71.80	81.23	68.39	81.97
	90%				**82.34**	**84.77**	**73.57**	**85.40**
SKATER	$c-3$	0.17415	0.14741	0.19465	58.07	72.63	57.02	73.47
	c	0.21371	0.17941	0.24005	53.24	67.56	51.02	67.89
	$c+3$	0.24092	0.24516	0.27006	49.64	63.36	46.37	63.45
AZP	$c-3$	0.23537	0.14779	0.26733	43.52	53.24	36.28	53.39
	c	0.24854	0.17059	0.28081	42.97	38.19	23.60	41.85
	$c+3$	0.18657	0.19945	0.20594	48.30	51.36	34.56	52.69
AZP_SA	$c-3$	0.19384	0.11809	0.21444	45.09	38.37	23.74	43.20
	c	0.14145	0.11201	0.15493	48.34	59.72	42.57	59.72
	$c+3$	0.22061	0.22617	0.24595	39.32	29.60	17.37	34.37
AZP_TABU	$c-3$	0.20738	0.16752	0.22732	46.57	48.66	32.16	50.25
	c	0.21042	0.21353	0.22851	54.15	57.56	40.41	58.79
	$c+3$	0.30981	0.25766	0.34688	40.37	30.94	18.30	35.87
AZP_RTABU	$c-3$	0.24173	0.11225	0.26920	49.53	57.09	39.95	57.41
	c	0.26007	0.22601	0.29338	41.96	40.90	25.71	43.23
	$c+3$	0.28364	0.24032	0.32242	42.29	51.31	34.51	51.56
ARISEL	$c-3$	0.26916	0.18284	0.30273	42.89	46.43	30.23	47.59
	c	0.26754	0.19789	0.30433	40.72	48.16	31.72	48.62
	$c+3$	0.30933	0.26337	0.34726	42.22	37.86	23.35	41.28
AMOEBA	*None*	0.50147	0.55240	0.56333	42.28	34.78	21.05	39.38
MAXP	10	0.14022	0.03872	0.15519	35.97	3.46	1.76	12.44
	100	0.08065	0.01122	0.08758	40.55	23.89	13.57	31.97
BPM	70%	0.22147	0.06551	0.24927	52.61	58.85	41.70	59.35
	80%				52.61	58.85	41.70	59.35
	90%				41.58	43.90	28.13	45.33
BDCD	70%	**0.04487**	**0.00426**	**0.04738**	65.12	78.77	64.98	80.55
	80%				40.61	16.44	8.95	29.42
	90%				35.04	0.17	0.09	2.88

6. Case Studies

The analysis described in this section illustrates the usefulness of our method in three relevant applications. The first one is a purely spatial case with Gaussian data and aims at the regionalization of the Brazilian municipalities according to their Human Development Index (HDI) in 2010. The second application is also purely spatial but assumes a Poisson distribution for death counts. It aims at the regionalization of the Southern Brazilian municipalities according to lung and bladder cancer mortality. We restrict the analysis to the Brazilian region where the mortality data is not affected by under reporting. The third application is a spatio-temporal situation with Gaussian data. We analysed the HDI as in the first application but now we describe its evolution from 1991 to 2010. In the three analysis, for the MCMC, we ran chains of size 10000, skipping the first 1000 samples as the *burn-in* period and take a lag of 10 to avoid correlation.

6.1. HDI Data: A Spatial Regionalization with Gaussian SPPM

The HDI is an index combining life expectancy, education and income measures developed by the United Nations Development Programme (UNDP) and it is often used to rank countries. Together with UNDP, the *Instituto de Pesquisa Econômica e Aplicada* (IPEA) and *Fundação João Pinheiro* developed a version of the HDI incorporating additional variables extracted from the demographic census to evaluate the Brazilian municipalities.

The data set is composed of the HDI of 5564 municipalities of Brazil (see Figure 7) in 2010. which is assumed to be normally distributed with mean and variance changing over the space. We consider the neighborhood structure computed through the geographic adjacency. As prior distribution for the cluster parameters we assume a conjugate Normal-Gamma distribution such that $\mu_{\mathcal{G}_k} \mid \tau_{\mathcal{G}_k} \sim N\left(0.65, (0.04\tau_{\mathcal{G}_k})^{-1}\right)$ and $\tau_{\mathcal{G}_k} \sim \text{Gamma}\,(100, 1)$. This prior distribution for the precision concentrates its probability mass around 100, which yields a standard deviation of about 0.1 for the observations of the clusters. Consequently, the cluster means are most probably around 0.65, with a deviation of 0.5. This spans most of the range the HDI can take, which is from 0.0 to 1.0. *A priori*, we assume that each edge is removed from the tree with probability $\rho \sim \text{Beta}(2, 7)$. As a result, the expected number of clusters in the map is around 1236. This is likely to be much larger than one would expect. However, we selected these values for two reasons. First, we want to allow a possible large variability in the parameters to be expressed by the prior. Second, we want to verify if the data provide enough evidence to shrink the distribution of the number of clusters.

The posterior distribution for the number c of clusters has mode equal to 12 and varies between 10 and 16, the 5% and 95% quantiles, respectively. The posterior probability of edge removal ρ has mean 0.0025 and median 0.0022.

In Figure 6 we show a random sample of partitions generated by our algorithm. The main difference between samples is on the frontier between the clusters. As expected, the boundaries are hard to be established because, for this type of data, we do not expect a clearly defined and sharp transition in the data.

To summarize the posterior partition distribution, we present Figure 7b. It is constructed by deleting the boundary between neighboring municipalities that belong to the same cluster in at least 80% of the sampled partitions. Therefore, the boundaries that remain indicates

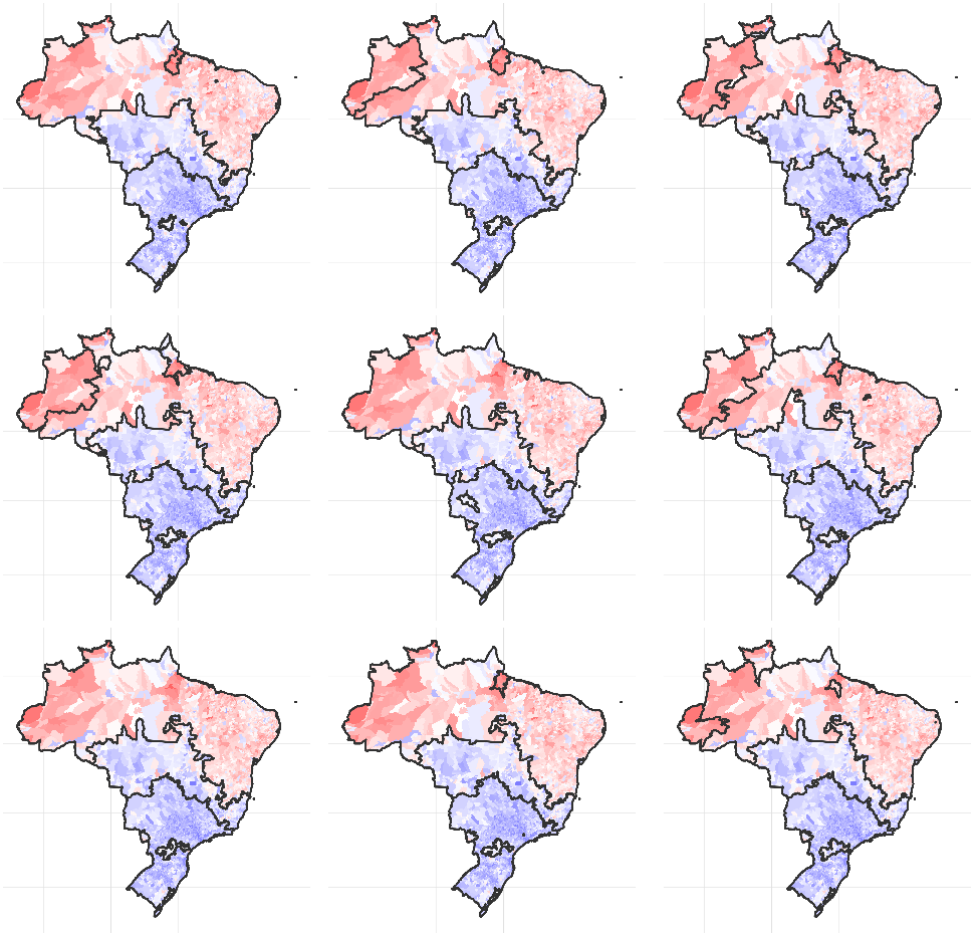
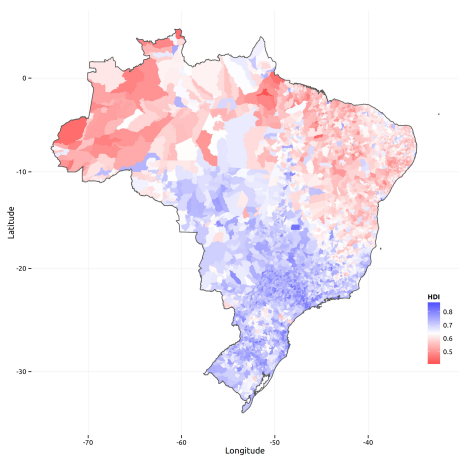


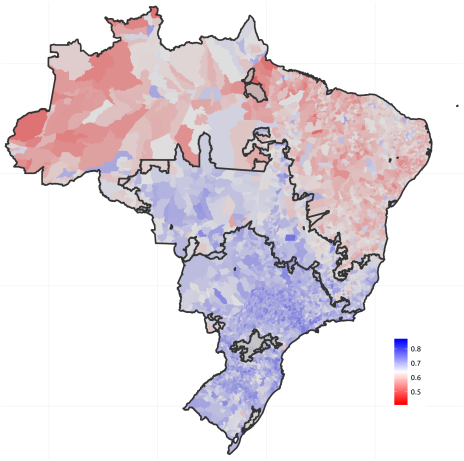
Figure 6: Some of the sampled partitions (HDI data)

areas that cannot be sharply allocated together into a cluster with its neighbor. It does not mean that the area necessarily stays isolated from their neighbors most of the time. It means that it is often allocated to clusters with different area components. Lines clearly splitting large portions of the map indicate sharp transition zones. Another posterior partition summary is given in Figure 7c . Each area is colored according to the size of cluster to which it belongs. Darker colors indicate that the area belong to a small cluster.

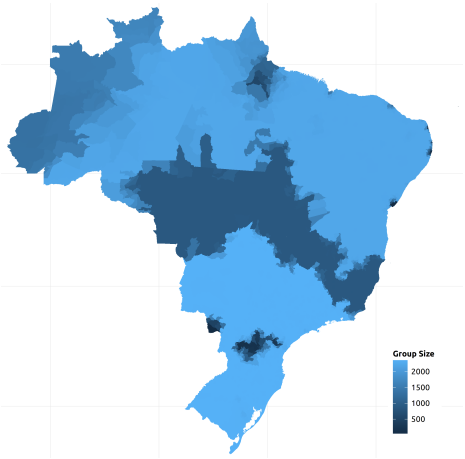
These figures show the presence of three large regions in the country. Additionally, we can see a number of small groups in the frontier and some tiny clusters located on the Northeast coast, hard to visualize on the figure. We also have two small clusters isolated in the middle of the big clusters, both with average HDI. The Southern one is surrounded by better HDI areas while the Northern one is relatively better than its neighbors. Although the main groups are well defined, the separation between them is not, causing a certain level of noise. In fact, this discloses the natural characteristic of this kind of data where there is a transition between two distinct groups but frontiers are not well defined. Another result is related to the tiny clusters found on the Northeast coast. These areas are frequently assigned to small clusters, with size within 9000 squared kilometers in at least 90% of the



(a) Map of Brazilian municipalities and their HDI data



(b) Summary of the sampled partitions for Brazilian municipalities (HDI data)



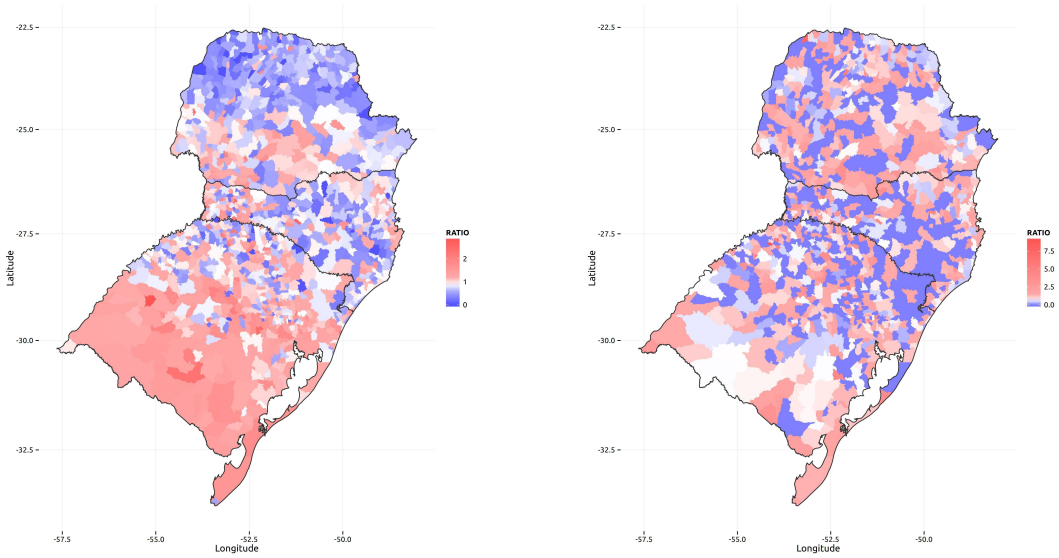
(c) Average group size for each municipality in the sampled partitions

Figure 7: Posterior regionalization of Brazilian municipalities according to their HDI

sampled partitions. Such areas are the capital of states in the northeast region and their neighboring towns. The HDI in the northeast region is generally very low, as can be seen in Figure 7a, but the capital cities and their surroundings are more developed and have a stronger economy than other regions on the State.

6.2. Lung and Bladder Cancer Data: A Spatial Regionalization with Poisson SPPM

To illustrate the use of the Poisson SPPM we consider the number of deaths by bladder and lung cancer in the south region of Brazil. These two types of cancers were selected due to their different incidence rate. Bladder cancer is almost an order of magnitude rarer than lung cancer. Thus, the incidence rate for bladder cancer is more affected by small variations.



(a) Lung cancer (b) Bladder cancer

Figure 8: Ratio between the actual and expected number of deaths by lung and bladder cancer in Brazilian south region

This scenario is a good example where the benefits of using a stochastic method instead of traditional approaches may be more easily perceived.

For both data sets, we obtained the total number of deaths by age group and gender, in each municipality, in the years 2008 - 2012. Data are available in the DATASUS website (`http://datasus.saude.gov.br/`). We also obtained demographic information of the same years, for the same age groups and gender, from IBGE. We summed over the years the number of deaths in each area to obtain the total number of events in the period generating one single count Y_i for each municipality. We assume that, conditional on a parameter θ_i, the random variable Y_i follows a Poisson distribution with mean $\mu_i = E_i\theta_i$. The values of E_i represent the expected number of deaths in area i if the risk was spatially constant in each age and sex class. That is, after obtaining the age-sex-specific mortality rates for the entire map, we applied them on each municipality respecting their demographic distribution by age and sex. We end up with the expected number of events in each municipality under this spatially homogeneous hypothesis. The θ_i parameters are called relative risks and they represent multiplicative factors with respect to the baseline E_i. In Figure 8 we show the ratio between the observed and the expected number of deaths by cancer.

We assume that, a priori, the cluster mortality rate has a Gamma distribution with parameters $a = 1.1$, and $b = 1.1$. This distribution concentrates its mass around 1.1, with a variance of 0.91, so the relative risk is concentrated in values mostly between 0 and 2. This seems reasonable as we expect the incidence rate to deviate from the expected value by a factor smaller than 2. We also assume that $\rho \sim \text{Beta}(5, 1000)$ and thus the expected number of cluster *a priori* is 6.9. The proposed model is compared to ARISEL and SKATER.

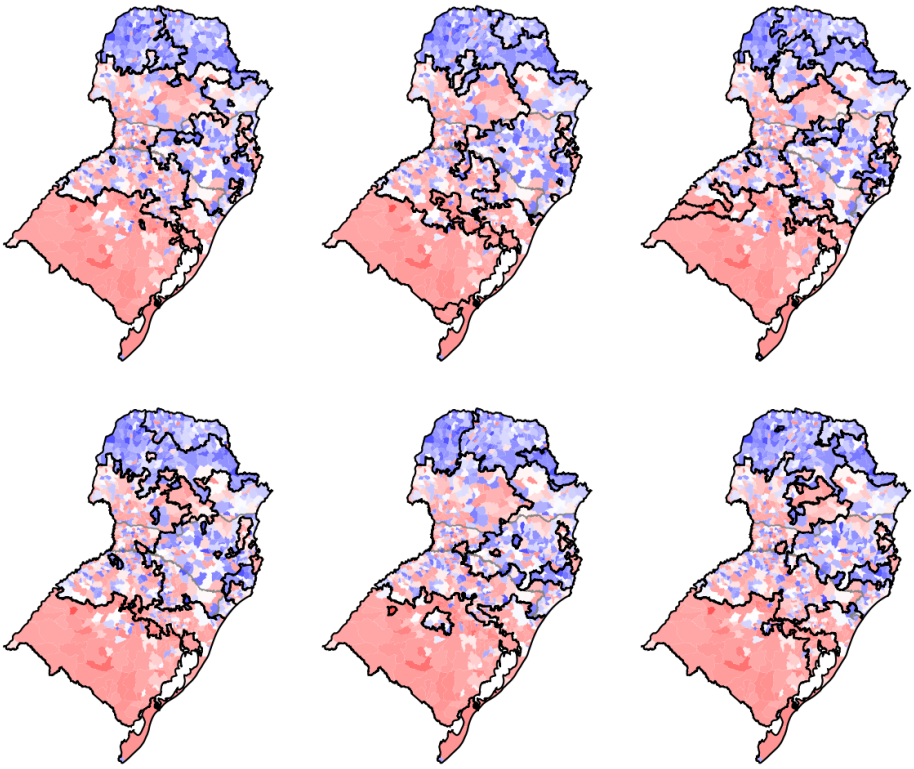


Figure 9: Some of the sampled partitions (lung cancer)

Inspired by our findings by applying the proposed model, for such approaches, we fixed the number the clusters equal to 4.

The posterior distribution for the number of clusters c has mode 29 (lung) and 13 (bladder) with 5% and 95% quantiles given by 18 and 34 (lung) and 7 and 22 (bladder). The edge removal parameter ρ has posterior mean and median around 0.013 and 0.008 for the lung and bladder examples, respectively.

In Figures 9 and 10 we show some of the partitions sampled by our algorithm for the lung and bladder cancer data sets, respectively.

As in the normal case, we summarize the posterior distribution of these partitions in Figure 11. Maps in (a) and (b) that are constructed taking into consideration the neighboring municipalities belonging to the same cluster in at least 85% of the sampled partitions. Maps in (c) and (d) represent the average size of the cluster to which each region belongs.

Our method is able to find the easily spotted clusters in the extreme North and South regions in Figure 8a for the lung cancer. However, it goes beyond that by also finding additional clusters that divide the central area into West and East. In the bladder cancer case, our model finds evidence in favor of the existence of clusters in the West and South regions. These findings are hardly identified by visual inspection of Figure 8b or by using some of the traditional approaches.

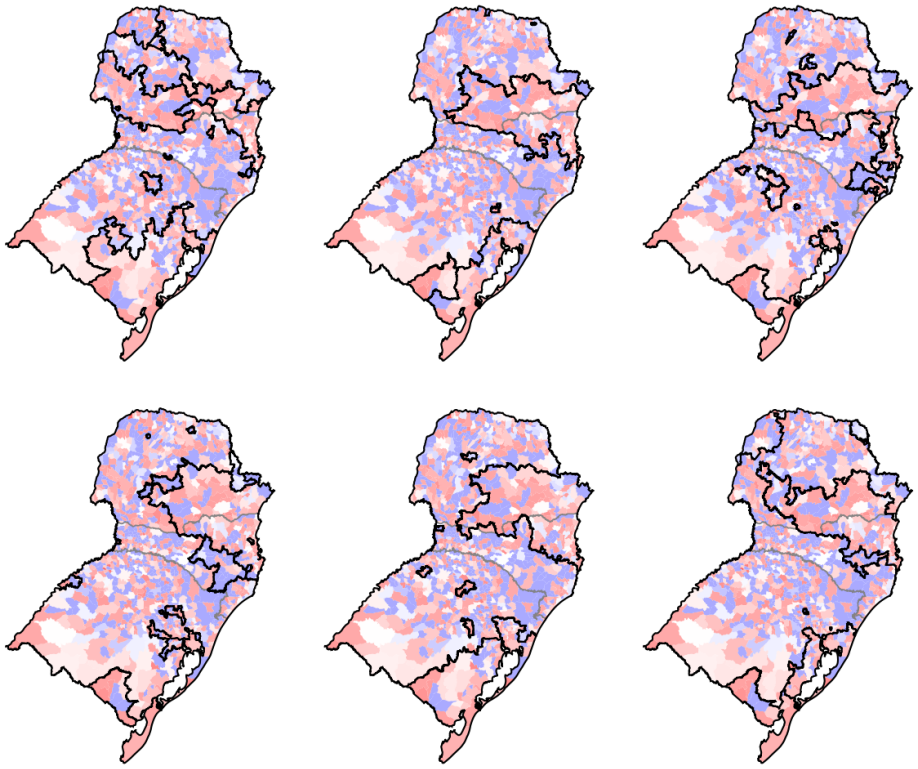
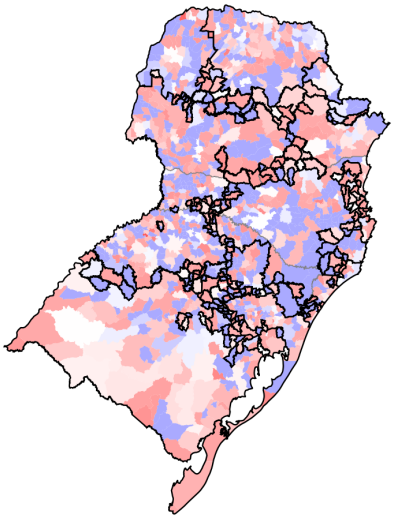


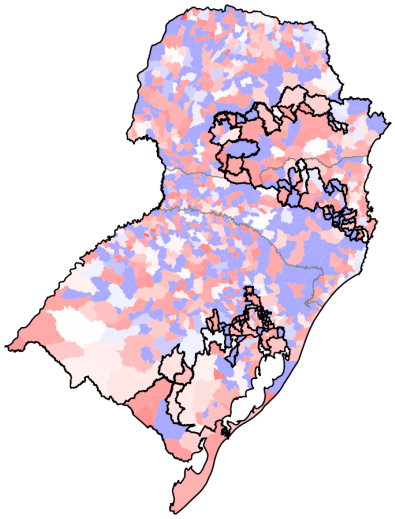
Figure 10: Some of the sampled partitions (bladder cancer)

As we can see in Figure 12, the *ARiSEL* and *SKATER* techniques result in regionalizations that capture only the most obvious visual aspects of the raw data maps. These deterministic methods were run with 4 groups. The *SKATER* method was still able to separate the top from the bottom of the lung cancer map, while *ARiSEL* detected only the southern boundary and with a more jagged line. For the bladder cancer, however, neither method was able to detect the region in the northeast of the map. They seemed to be more sensitive to local variations in the rate. This is the practical exemplification of what we expected. Given that these methods do not use a statistical model, they are more susceptible to this kind of problem where the population incidence rate is small.

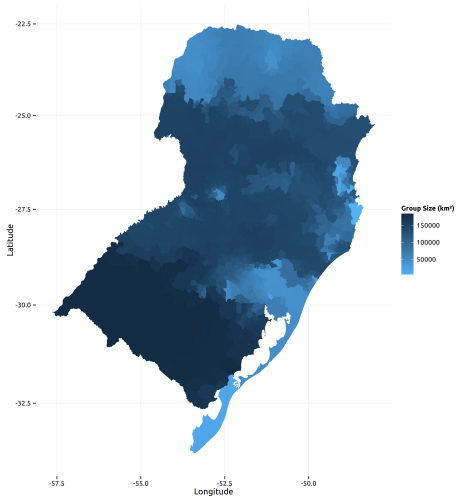
Lung cancer is linked to smoking and the rate increases going south, which could be related to the colder climate. This explains the clusters in the two extreme parts of the North-South gradient in Figure 11a. With the help of hindsight, we can see that the right portion of the middle region in this figure is where some of the large cities are located, and presents a lower rate than its western counterpart. The deterministic methods either separates out only the extreme south region and fail to identify the regions in the center. A more visible difference between the methods is in Figure 11b. SPPM identified a cluster in the northeast of the map, where Curitiba (capital of Paraná state) and the most populated cities of the state of Santa Catarina are located. All the other methods failed to find this



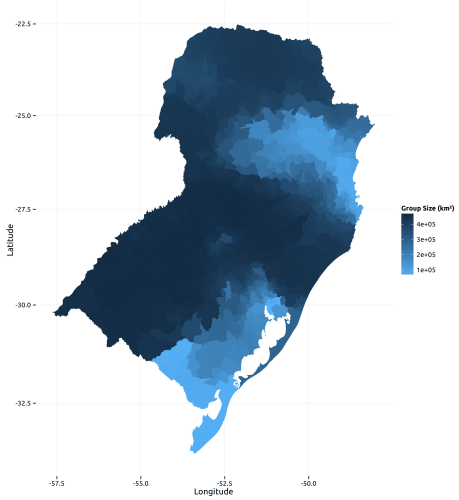
(a) Summary of the sampled partitions for each region, Lung cancer



(b) Summary of the sampled partitions for each region, Bladder cancer

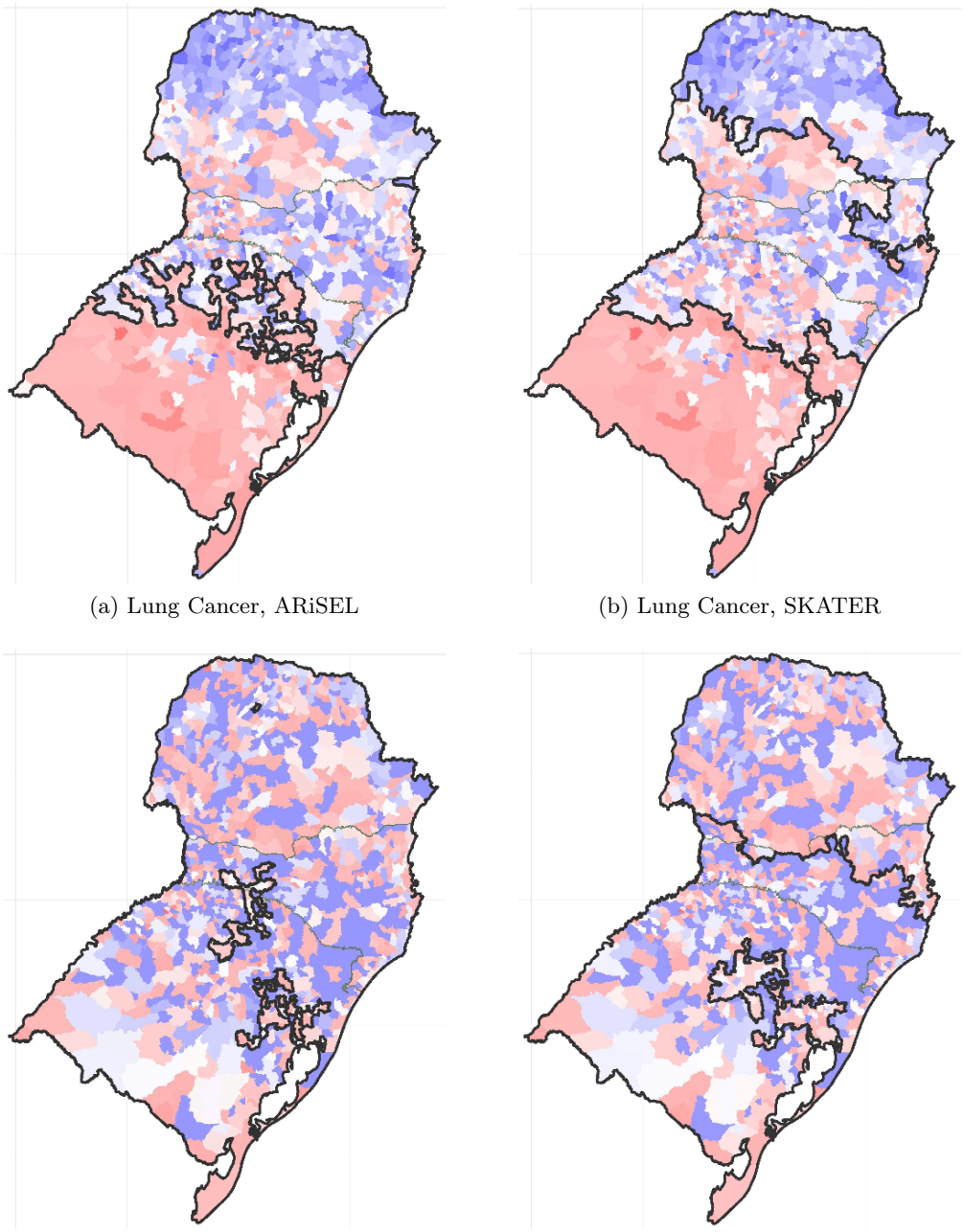


(c) Average group size for each region, Lung cancer



(d) Average group size for each region, Bladder cancer

Figure 11: Posterior regionalization of municipalities in Brazilian south region according to the number of deaths by lung and bladder cancer, proposed model



(a) Lung Cancer, ARiSEL

(b) Lung Cancer, SKATER

(c) Bladder Cancer, ARiSEL

(d) Bladder Cancer, SKATER

Figure 12: Regionalization of municipalities in the Brazilian south region according to the number of deaths by lung and bladder cancer, ARiSEL and SKATER methods.

cluster and instead identified noisy small regions, which is a demonstration of how these methods are sensitive to the variations of the data.

6.3. HDI Data: A Space-Time Regionalization

In this section, we provide a short example of our space-time method using the state-level HDI for the 26 Brazilian states and the Federal district. We have data for three different time moments, the 1991, 2000, and 2010 Census years, which implies in 81 graph nodes. We used the state-level graph to simplify the analysis and the interpretation of the results. We considered the same prior choices as in Section 6.1: $\mu_{\mathcal{G}_k} \mid \tau_{\mathcal{G}_k} \sim N\left(0.65, (0.04\tau_{\mathcal{G}_k})^{-1}\right)$, $\tau_{\mathcal{G}_k} \sim \text{Gamma}\,(100, 1)$, and $\rho \sim \text{Beta}(2, 7)$.

Figure 13 shows the HDI in each year in the top row and the summary of our posterior distribution in the bottom row. We delete the boundary between neighboring states that belong to the same cluster in at least 70% of the sampled partitions. It is clear the improvement of the Brazilian socio-economic condition over the three decades. In the first decade, the small Federal District was isolated from the other regions presenting the highest HDI level. In the second decade, this cluster is enlarged including the states in southern part of the map. Such regions, in fact, presented an improvement in their socio-economic conditions if compared with the first decade under analysis, getting closer of the Federal District. Another small cluster was identified separating Amapá state, located in the North, from the other regions. After three decades, all Brazilian states are assembled in a single cluster.

7. Conclusion

In this work we dealt with the problem of regionalization, an important type of clustering problem which arises in many areas. We proposed a new product partition model that accounts for spatial and spatio-temporal clustering. The innovative aspect of this paper is the use of random spanning trees as a tool to reduce the search space of partitions. This random spanning tree trick allowed us to represent the random partitions as fixed dimensional vector with binary elements. We presented a posterior sampling algorithm for the proposed model that keeps fixed the dimension of the parametric space and resorts to a regular Gibbs sampling procedure.

We evaluated of our method using simulated and real data. In the simulated study, we compared our approach to available implementations of non-stochastic optimization methods and Bayesian proposals and showed how our results were consistently superior. This was particularly noticed in low rate Poisson generated data. We applied our technique to perform the regionalization of Brazilian municipalities based on the human development index and on bladder and lung cancer mortality data. We discussed how the results we obtained were suitable for the domain subjects and how the use of our model obtained better and more meaningful results.

The model proposed has some limitations. The temporal component is not dealt with typical time series techniques and this prevents us to make inference in a prospective way. More importantly, the prior distribution on the set of spanning trees is an artificial tool to make inference on the partitions and these spanning trees do not have an obvious interpretation making difficult a prior distribution elicitation for $\mathcal{T}$. However, the spanning tree plays

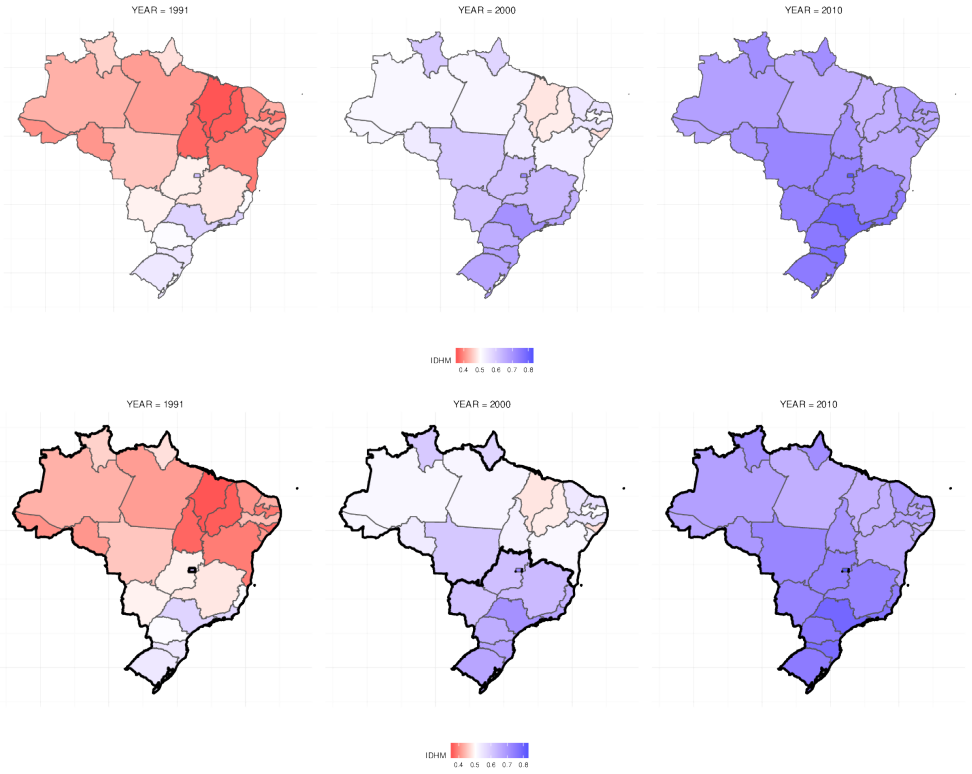


Figure 13: Map of Brazilian states and their HDI data (first row) and the evolution of the posterior summaries for cut point 0.7 (second row), over the years 1991, 2000, and 2010.

an important role in the prior elicitation for the partition. Coupled with the simplification of the computational search for a good partition by restricting this search to n edges, it allowed us to write the prior cohesion as functions of ρ, the prior probability of removing edges from the tree.

In conclusion, we proposed a stochastic model for the problem of space-time regionalization that captures much of the prior reasoning one could have for its formation. We introduced the use of spanning trees to provide an effective sampling algorithm. Our model is flexible enough to accommodate different types of data and provides good results.

Acknowledgments

The authors would like to thank CNPq, CAPES and FAPEMIG, three Brazilian funding agencies, for partial support. We also thank the comments provided by Pedro O. S. Vaz de Melo and Fábio Cozman on a first version of this paper and the comments provided by the editor and reviewers, which contributed to improve the paper. We also want to thank Vinícius Fernandes dos Santos for suggesting the counter-example illustrated in the

left hand-side of Figure 4. The research for this project was conducted while the first author was a student at the Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, where he obtained his Master of Science degree in Computer Science.

References

Jared Aldstadt and Arthur Getis. Using amoeba to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, 38(4):327–343, 2006.

Craig Anderson, Duncan Lee, and Nema Dean. Identifying clusters in Bayesian disease mapping. *Biostatistics*, 15(3):457–469, 2014.

Renato M. Assunção, Marcos C. Neves, Gilberto Câmara, and Corina da Costa Freitas. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811, 2006.

Fernando Bação, Victor Lobo, and Marco Painho. Geo-self-organizing map (geo-som) for building and exploring homogeneous regions. In *Geographic Information Science*, pages 22–37. Springer, 2004.

Fernando Bação, Victor Lobo, and Marco Painho. Applying genetic algorithms to zone design. *Soft Computing*, 9(5):341–348, 2005.

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

Daniel Barry and John A. Hartigan. A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.

Iacopo Bernetti, Christian Ciampi, and Sandro Sacchelli. Minimizing carbon footprint of biomass energy supply chain in the province of florence. *Italian Journal of Forest and Mountain Environments*, 66(4):321–329, 2011.

Luke Bornn and Francois Caron. Bayesian clustering in decomposable graphs. *Bayesian Analysis*, 6(4):829–846, 12 2011. doi: 10.1214/11-BA630.

Andrei Broder. Generating random spanning trees. In *30th Annual Symposium on Foundations of Computer Science, 1989*, pages 442–447. IEEE, Oct 1989. doi: 10.1109/SFCS.1989.63516.

Samantha Cockings and David Martin. Zone design for environment and health studies using pre-aggregated data. *Social Science & Medicine*, 60(12):2729 – 2742, 2005. doi: http://dx.doi.org/10.1016/j.socscimed.2004.11.005.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009. ISBN 0262033844, 9780262033848.

David G.T. Denison and Christofer C. Holmes. Bayesian partitioning for estimating disease risk. *Biometrics*, 57(1):143–149, 2001.

Juan C. Duque and Richard L. Church. A new heuristic model for designing analytical regions. In *North American Meeting of the International Regional Science Association, Seattle*, 2004.

Juan C. Duque, Boris Dev, Alejandro Betancourt, and Jose L. Franco. *ClusterPy: Library of spatially constrained clustering algorithms, Version 0.9.9.* RiSE-group (Research in Spatial Economics). EAFIT University., Colombia, 2011. URL `http://www.rise-group.org`.

Juan C. Duque, Luc Anselin, and Sergio J. Rey. The max-p-regions problem. *Journal of Regional Science*, 52(3):397–419, 2012.

Ronald E. Gangnon and Murray K. Clayton. Bayesian detection and modeling of spatial disease clustering. *Biometrics*, 56(3):922–935, 2000.

Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Peter J. Green and Alun Thomas. Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, 100(1):91–110, 2013. doi: 10.1093/biomet/ass052.

Diansheng Guo. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, 22 (7):801–823, 2008.

John A. Hartigan. Partition models. *Communications in Statistics-Theory and Methods*, 19 (8):2745–2756, 1990.

Avril Hegarty and Daniel Barry. Bayesian disease mapping using product partition models. *Statistics in Medicine*, 27(19):3868–3893, 2008.

Leonhard Knorr-Held and Günter Raßer. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21, 2000.

Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

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

David Madigan and Adrian E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

David Martin. Optimizing census geography: the separation of collection and output geographies. *International Journal of Geographical Information Science*, 12(7):673–685, 1998.

Daniel W. McKenney, John H. Pedlar, Kevin Lawrence, Kathy Campbell, and Michael F. Hutchinson. Beyond traditional hardiness zones: Using climate envelopes to map plant range limits. *BioScience*, 57(11):929–937, 2007. doi: 10.1641/B571105.

Jeremy Mennis and Philip Harris. Spatial contagion of male juvenile drug offending across socioeconomically homogeneous neighborhoods. In *Crime Modeling and Mapping Using Geospatial Technologies*, pages 227–248. Springer, 2013.

Stan Openshaw. A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers*, pages 459–472, 1977.

Stan Openshaw and Liang Rao. Algorithms for reengineering 1991 census geography. *Environment and planning A*, 27(3):425–446, 1995.

Garritt L. Page and Fernando A. Quintana. Spatial product partition models. *Bayesian Analysis*, 11(1):265–298, 03 2016. doi: 10.1214/15-BA971.

Ilka A. Reis, Gilberto Câmara, Renato Assunção, and Antônio M. V. Monteiro. Data-aware clustering for geosensor networks data collection. In *Anais XIII Simpósio Brasileiro de Sensoriamento Remoto*, pages 6059–6066, 2007.

Patricia B. Ribeiro, Roseli A.F. Romero, Patrícia R. Oliveira, Homero Schiabel, and Luciana B. Verçosa. Automatic segmentation of breast masses using enhanced ica mixture model. *Neurocomputing*, 120:61–71, 2013.

Sylvia Richardson and Peter J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997.

Thomas C. Ricketts. *Using geographic methods to understand health issues*. Agency for Health Care Policy and Research, Dept. of Health and Human Services, US Public Health Service, 1997.

Anne Ruas. Map generalization. In *Encyclopedia of GIS*, pages 631–632. Springer, 2008.

Roger Sayre, Jack Dangermond, Charlie Frye, Randy Vaughan, Peter Aniello, Sean Breyer, Douglas Cribbs, Dabney Hopkins, Richard Nauman, William Derrenbacher, et al. A new map of global ecological land units—an ecophysiographic stratification approach. *Washington, DC: Association of American Geographers*, 2014.

Leonardo V. Teixeira, Renato M. Assunção, and Rosangela H. Loschi. A generative spatial clustering model for random data through spanning trees. In *2015 IEEE International Conference on Data Mining*, pages 997–1002, Nov 2015. doi: 10.1109/ICDM.2015.106.

Jonathan Wakefield and Albert Kim. A Bayesian model for cluster detection. *Biostatistics*, 14(4):752–765, 2013.

David B. Wilson. Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 296–303, New York, NY, USA, 1996. ACM. ISBN 0-89791-785-5. doi: 10.1145/237814.237880.

Steve Wise, Robert Haining, and Jingsheng Ma. Regionalisation tools for the exploratory spatial analysis of health data. In Manfred M. Fischer and Arthur Getis, editors, *Recent Developments in Spatial Analysis*, Advances in Spatial Science, pages 83–100. Springer Berlin Heidelberg, 1997.

Ying Zhang, Semu Moges, and Paul Block. Optimal cluster analysis for objective regionalization of seasonal precipitation in regions of high spatial-temporal variability: Application to Western Ethiopia. *Journal of Climate*, 29(10):3697–3717, 2016.