Aline Martines Piroutek

NOVOS MODELOS DE VIZINHANÇA ESPACIAL E VIGILÂNCIA PROSPECTIVA ESPAÇO-TEMPO

Belo Horizonte, Novembro de 2013

Aline Martines Piroutek

NOVOS MODELOS DE VIZINHANÇA ESPACIAL E VIGILÂNCIA PROSPECTIVA ESPAÇO-TEMPO

Tese apresentada como requisito parcial
para obtenção de grau de Doutor em Estatística
pela Universidade Federal de Minas Gerais.

Orientador: Prof. Dr. Renato Martins Assunção
Co-Orientador:Profa. Dra. Denise Duarte

Programa de Pós-Graduação em Estatística
Departamento de Estatística
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais

Belo Horizonte, Novembro de 2013

...numa montanha vermelha você me resgatou.
E, no cume de outra, sonho em te reencontrar...

PARA MINHA MÃE,
OLGA,
(in memoriam)

Agradecimentos

A trajetória de um aluno de doutorado é bastante longa e árdua. Sozinhos, nós não seriamos capazes de alcançar essa vitória. Sim, digo vitória. Vitória, primeiro pela ansiedade para ingressar no programa. Vitória, por conseguir estudar de madrugada disciplinas antes assumidas por outros não "condizentes"com nossas capacidades. Vitória, pelo esforço em aprender e a tornarmos profissionais em busca de nossas "curiosidades científica". Agradeço a Deus que, através de várias pessoas em nossa volta, nos mostra o caminho e nos fortalece com sua bondade. Devo agradecer ao meu pai Mirko, pela apoio incondicional. Não sei como poderia ir em frente sem esse exemplo de vida. Pessoa dedicada a família e que se esforça integralmente para nosso sucesso. Fique tranquilo, pois você fez e faz muito mais do que o suficiente. Às minhas irmãs, eu também agradeço... À Tálita e seu "sub-bloco"(André, Carol e Daniel) por serem minha segunda casa. Obrigada pelo momentos bons e por me acolher nesse novo lar. Sempre poderei aprender com os mais experientes e desfrutar da alegria das crianças. À Jéssica, minha irmã mais nova, agradeço pela alegria e pelo carinho. Obrigada por me esperar voltar para casa e querer passar um tempo comigo. Obrigada por me escutar falar de mim mesmo quando eu não te deixava falar de você. Sei que as vezes você sofre e acha que não é valorizada, mas saiba que sinto muito orgulho de todo seu talento e de suas conquistas.

Eu não poderia me esquecer de agradecer ao meu esposo Rafael. Obrigada por fazer o pouco tempo disponível juntos valer a pena. Com você, descobri que uma palavra, um gesto carinhoso, podem mudar o mundo. Obrigada pelo exemplo de pessoa, de caráter e de persistência. Sem você, existiriam mais pedras no caminho. Devo um agradecimento especial ao Renato e à Denise. Mestres que se equilibram. O primeiro, me ofereceu a ambição e a determinação. Como Renato, existiram conquistas e decepções, mas sempre com muita admiração. Um pai acadêmico, com direito a broncas e elogios. Obrigada por todos esses anos de orientação e ensinamentos. Do outro

lado, existe a Denise, com sua doçura e paciência. Agradeço por me animar e confiar em mim. Sua calma e ajuda foram essenciais para a conclusão desse trabalho. Agradeço aos meus amigos do LESTE e do departamento: Márcia, Érica, Letícia, Marquinhos, Thaís, Jaque, Gustavo, Grazi, Bebel. Juntos, provamos que conhecimento e alegria possuem interseção. Aos professores Marcelo, Rosangela e Emília, pelo exemplo de bons profissionais. Agradeço aos meus amigos Thábatta, Cibele, Débora, Bruno, Priscila, Maumau, Adriano, Carla e Matheus. Sei que estive distantes, mas vocês continuaram me apoiando e me alegrando. Obrigada pela amizade. Agradeço todos aqueles que fizeram parte do meu caminho.

Resumo

No primeiro artigo, introduzimos um modelo denominado "Probabilistic Context Neighborhood"para um lattice de duas dimensões como uma extensão do modelo de árvore de contexto probabilística em um espaço unidimensional, preservando algumas de suas propriedades interessantes. Este modelo tem uma estrutura de vizinhança variável com uma geometria fixa, mas de raio variável. Desta forma, calculamos a cardinalidade do conjunto de vizinhos e utilizamos o Pseudo-Likelihood Bayesian Criterion (PIC) para selecionar um modelo adequado para cada amostra. Representamos, ainda, a estrutura de dependência da vizinhança como uma árvore, objetivando facilitar o entendimento da complexidade do modelo. Em seguida, propusemos um algoritmo para estimar o modelo que explora a estrutura de árvore esparsa, com o intuito de melhorar a eficiência computacional. Por fim, apresentamos uma extensão do modelo anterior, o "Non-Homogeneous Probabilistic Context Neighborhood model", permitindo uma mudança espacialmente probabilística na vizinhança de contexto à medida que nos movemos no lattice. No segundo artigo, propusemos uma abordagem Bayesiana para a problemática da escolha da matriz de vizinhança no mapeamento de doenças. Utilizamos o modelo proposto por Besag et al. (1991), no qual um dos efeitos aleatórios segue o modelo "Conditional Autoregressive"(CAR), no qual se utiliza a matriz de adjacência como vizinhança. Observamos, contudo, que a matriz de adjacência é inadequada para cenários nos quais a estrutura espacial é insuficiente para a obtenção de boas estimativas dos riscos relativos, tal como ocorre em estudos criminológicos e epidemiológicos que apresentam evidências sobre elevadas taxas de incidência em grandes centros, independentemente das taxas apresentadas pelas cidades vizinhas. Além das classes a priori, propusemos dois estimadores a posteriori para a estrutura de vizinhança. Foram apresentados vários exemplos, simulações e aplicações do nosso método, cujos resultados foram mais satisfatórios que os apresentados pelo modelo CAR. Por fim, no último artigo, propusemos um sistema de vigilância para monitorar prospectivamente a emergência de

clusters espaço-temporais em dados de doenças. O objetivo, nesse caso, é detectar um cluster logo que venha a emergir, mantendo a taxa de falsos alarmes em um nível controlado. Trata-se de um sistema de fácil compreensão e de fácil implementação, com imensa aplicabilidade pelos órgãos oficiais de saúde pública, refletindo uma modificação de proposta anteriormente feita por Rogerson (2001), que examinou um cenário de vigilância retrospectiva, olhando para o tempo mais antigo no passado que a mudança poderia ter ocorrido. Modificamos o método, de forma a levar em conta os casos prospectivos, e avaliamos o nosso sistema de vigilância em vários cenários de clusters emergentes, verificando os pressupostos de distribuição e avaliação de desempenho e impacto de diferentes tempos de emergência, formas, extensão e intensidade.

Abstract

In the first paper, we introduce the Probabilistic Context Neighborhood model for two dimensional lattices as an extension of the Probabilistic Context Tree model in one dimensional space preserving some of its interesting properties. This model has a variable neighborhood structure with a fixed geometry but varying radius. In this way we are able to compute the cardinality of the set of neighborhoods and use the Pseudo-Likelihood Bayesian Criterion to select an appropriate model given the data. We represent the dependence neighborhood structure as a tree making easier to understand the model complexity. We provide an algorithm to estimate the model that explores the sparse tree structure to improve computational efficiency. We also present an extension of the previous model, the Non-Homogeneous Probabilistic Context Neighborhood model, which allows a spatially changing Probabilistic Context Neighborhood as we move on the lattice. In the second paper, we proposed a Bayesian approach to the problem of choosing the neighborhood matrix in disease mapping. We use the model proposed by Besag et al. (1991), whose random effects follow the Autoregressive Conditional (CAR), using the adjacency matrix as neighborhood. However, the adjacency matrix is inadequate for scenarios in which the spatial structure is insufficient to obtain a precise estimative of the relative risks, as occurs in criminological and epidemiological studies that show high incidence rates in large towns, regardless the rates presented by neighboring cities. In addition to the a priori classes, we proposed two a posteriori estimators for the neighborhood structure. Finally, we presented several examples, simulations and applications of our method, which reached more satisfactory results than the CAR model. In the last paper, we propose a surveillance system to prospectively monitor the emergence of space-time clusters in point pattern of disease events. Its aim is to detect a cluster as soon as possible after its emergence and it is also desired to keep the rate of false alarms at a controlled level. It is an easily understood and easily implemented system, requiring very little input from the user. This makes it a promising candidate

to practical use by public health official agencies. Our method is a modification from a previous proposal made by Rogerson, who examined a retrospective surveillance scenario, looking for the earliest time in the past that change could have been deemed to occur. We modify his method to take into account the prospective case. We evaluated our surveillance system in several scenarios, including without and with emerging clusters, checking distributional assumptions and assessing performance impacts of different emergence times, shapes, extent and intensity of the emerging clusters.

Palavras-chaves: *Estatística espacial, mapeamento de doença, campos de Markov, modelos espaciais hierárquicos, vigilância prospectiva espaço-tempo, cluster espaço-tempo, árvores de contexto, campos de Markov com tamanho variável, vizinhança de contexto probabilística.*

Introdução

A presente tese de doutorado apresenta resultados de investigações complexas e aprofundadas na área de Estatística Espacial.

Trata-se, em verdade, de três trabalhos independentes e autônomos, e em razão de terem sido submetidos à publicação no exterior, encontram-se em formato de artigos - razão pela qual dois deles foram redigidos em inglês.

A primeira parte é intitulada "Non-Homogeneos Probabilistic Context Neighborhood estimation on two dimensional lattice by pruning spanning trees", no qual exploramos métodos de estimação de probabilidades de transição em Campos de Markov (Markov Random Field - MRF).

Assim como ocorre com as cadeias de Markov, a modelagem dos MRFs torna-se um problema quando o número de sites vizinhos é elevado.

Rissanen (1983) introduziu, para casos unidimensionais, a ideia de memória variável em modelos PCT, na qual a predição do próximo símbolo depende tão-somente da parte relevante de seu passado (contexto). Assim, o contexto de cada site varia segundo os valores de seus símbolos antecessores. O método permite, ainda, a representação do conjunto de contextos a partir de uma árvore, que sumariza toda a estrutura de dependência do passado de uma cadeia de Markov.

Existem outros estudos focados nos processos MRFs. Ao utilizarem o Pseudo-Bayesian Information Criterion (PIC), Csiszar and Talata (2006) propõem um estimador fortemente consistente nas realizações de uma crescente região finita. Locherbach, por sua vez, denomina esse processo de Variable Neighborhood Random Field (VNRF), estendendo o conceito de contexto (ou vizinhança de contexto) para um lattice com r-dimensões propondo um estimador para o raio de uma vizinhança básica (o menor círculo que contém o contexto do site).

Inspirados nos resultados obtidos para as cadeias de Markov, propusemos uma adaptação dos modelos Variable Length Markov Chain (VLMC) e Probabilistic Context Tree (PCT), adotando as

ideias de vizinhança variável em um lattice, estimação do PCT para casos unidimensionais e uso do PIC como critério de seleção. Além disso, propusemos uma nova geometria para as vizinhanças de contexto, variando o raio da estrutura (segundo os valores que apresenta). Esse nova geometria permite uma drástica redução de parâmetros livres e torna possível o cálculo da cardinalidade do conjunto de vizinhanças de contexto.

Nosso modelo, chamado de Probabilistic Context Neighborhood (PCN), possibilita representar graficamente a estrutura de dependência das vizinhanças através de uma árvore - aspecto fundamental para o entendimento do comportamento de interações. Propõe, ainda, um algoritmo de estimação das vizinhanças de contexto gerado por um PCN. Apresenta, por fim, um método de regionalização para os cenários em que as PCN não são homogêneas ao longo do lattice (Non-Homogeneous Probabilistic Context Neighborhood - NHPCN), decorrente de uma adaptação do Spatial 'K'luster Analysis by Tree Edge Removal (Skater). Para possibilitar o método de regionalização, propusemos três medidas de dissimilaridades entre árvores (probabilidade, complexidade e estrutura).

Ressalte-se que toda a metodologia desenvolvida no trabalho apresentou resultados satisfatórios nas simulações.

Na segunda parte da tese, apresentamos trabalho intitulado "Vizinhanças a priori em Campos de Markov". Toma-se como base o mapeamento de doenças, o qual tem como objetivo recuperar o comportamento de uma dada enfermidade nas áreas de uma região determinada. Cuida-se de temática importante, uma vez que estima os riscos relativos, permitindo maior efetividade na implementação de políticas públicas. Ao utilizar a abordagem Bayesiana, os riscos relativos são considerados variáveis aleatórias, o que permite incorporar uma dependência entre os eventos a partir dos campos de Markov. O modelo mais conhecido foi proposto por Besag et al. (1991), o qual se baseia na decomposição do logaritmo do risco relativo em dois efeitos aleatórios: um deles estruturado espacialmente e, o outro, não. O primeiro efeito segue o modelo Conditional Autoregressive (CAR), no qual a distribuição em uma dada área, dadas todas as outras, depende tão-somente dos valores de seus vizinhos. Nesse caso, a vizinhança é definida por meio de uma matriz, baseada na adjacência. Ressalte-se que a escolha da contiguidade baseia-se na conveniência ou na facilidade de implementação das rotinas no sistema Geographic information system (GIS).

Dessa forma, nesta primeira parte, propusemos a criação de uma classe de matrizes consider-

adas *piroris*, na qual será possível fazer inferências através de métodos computacionais (MCMC). Além disso, propusemos dois estimadores *a posteriori*, para a estrutura de vizinhança. Foram apresentados vários exemplos, simulações e aplicações do nosso método, cujos resultados foram mais satisfatórios quando comparados ao modelo CAR.

Por último, apresentamos o trabalho intitulado "Space-time prospective surveillance based on Knox local statistics", o qual aborda métodos prospectivos na detecção de clusters espaço-tempo.

Os métodos de vigilância espaço-tempo podem ser divididos em duas categorias: aqueles que usam dados de áreas e os que usam dados pontuais. Dentre os primeiros, vale ressaltar o estudo desenvolvido por Raubertas (1989) - o qual descreve procedimentos analíticos e a sua implementação na epidemiologia e na saúde pública. Kulldorff (2001), por sua vez, propõe a estatística scan para o monitoramento contínuo de doenças em dados de áreas, a qual foi adaptada pelos Neill et al.(2005), Kulldorff et al. (2005) e Assunção et al. (2007) e criticada por Woodall et al. (2008).

Para a categoria de dados pontuais, podemos mencionar o trabalho de Assunção and Correa (2009) , o qual não exige a especificação das funções puramente espacial e temporal. A proposta dos autores consiste em adaptar o método gráfico de controle introduzido por Shiryayev-Roberts para a situação espaço-tempo e utilizar o Optional Stopping Theorem para obter os valores dos parâmetros de ajuste do método.

Rogerson (2001), por sua vez, propõe a utilização individual dos pontos geo-referenciados dos eventos nos cálculos da estatística de Knox local. Trata-se de método retrospectivo apresentando significativas vantagens em relação a seus antecessores: baseia-se em uma estatística intuitiva (Knox local); não necessita de conhecimento prévio na sua aplicação; desnecessidade de modelar as funções marginais de padrão espacial e temporal; e seus parâmetros são interpretáveis e facilmente fixados.

Marshall (2007) estuda o método de Rogerson através de simulações, sob a hipótese de ausência de clusters. Ao avaliar o efeito do average run length (ARL) em cenários diversos, o autor encontra diversos problemas na estatística de knox local, razão pela qual não recomenda a utilização do método para aplicação em casos reais.

Nosso trabalho propõe um método simples, que requer pouca modelagem estatística e de fácil implementação. Sua principal contribuição está na redefinição da estatística de knox local para que passe a atuar de forma prospectiva.

Assim como em Rogerson, nossos parâmetros são interpretáveis e podem ser fixados utilizando

um dado aspecto da doença em análise. Também checamos a distribuição exponencial dos ARL_0 assumida por Rogerson. Observamos, ainda, que a aproximação para o cálculo do limiar h funciona bem quando existem pelo menos seis eventos dentro da área de busca. Além disso, as conclusões obtidas a partir das simulações são válidas tanto para cenários com densidade populacional homogênea e como para cenários com densidade não homogênea. Da mesma forma, as simulações em cenários sob a hipótese de cluster apresentaram os resultados esperados.

Em suma, o método é adequado à detecção de cluster espaço-tempo e extremamente útil em aplicações reais.

Non-Homogeneous Probabilistic Context Neighborhood estimation on two dimensional lattice by pruning spanning trees

A. Piroutek[a], D. Duarte[a,*], R. Assunção[b],

[a] *Departamento de Estatística, Universidade Federal de Minas Gerais, 31270-901. Belo Horizonte, MG, Brazil*
[b] *Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 31270-010. Belo Horizonte, MG, Brazil.*

Abstract

We introduce the Probabilistic Context Neighborhood model for two dimensional lattices as an extension of the Probabilistic Context Tree model in one dimensional space preserving some of its interesting properties. This model has a variable neighborhood structure with a fixed geometry but varying radius. In this way we are able to compute the cardinality of the set of neighborhoods and use the Pseudo-Likelihood Bayesian Criterion to select an appropriate model given the data. We represent the dependence neighborhood structure as a tree making easier to understand the model complexity. We provide an algorithm to estimate the model that explores the sparse tree structure to improve computational efficiency. We also present an extension of the previous model, the Non-Homogeneous Probabilistic Context Neighborhood model, which allows a spatially changing Probabilistic

*Corresponding author
Email addresses: apiroutek@ufmg.br (A. Piroutek), denise@est.ufmg.br (D. Duarte), assuncao@dcc.ufmg.br (R. Assunção)

Context Neighborhood as we move on the lattice.

Keywords: Context tree, Markov random fields, Variable-neighborhood random fields, Context algorithm, Probabilistic context trees, Model selection.

---

1. Introduction

In this paper we are concerned with the task of providing transition probability estimators for Markov random fields (MRF) on a two dimensional lattice by using a specific kind of neighborhood geometry with variable size. We mean by neighborhood the minimal region that determines the conditional distribution of a site subject to the values of all other sites. We also address the problem of model selection inside a class of MRF with variable neighborhood structure.

The Markov random field (MRF) on lattices is a model that has been increasingly exploited nowadays. It has for example several applications in computing. We can mention the image processing, which includes recognition, segmentation, image compression and restoration ([1], [2], [3], [4] and [5]). In statistical physics, the MRF is essential for modeling interactive particle systems [6]. In sociology, we can see several applications in polarization phenomena in society and in social networks [7]. In the area of machine learning the MRF are used in the search for hidden patterns, called learning structure [8].

Markov chain modeling could become a problem when the order dependency is not small because, in this case, the number of parameters to estimate is very large. The same problem occurs for MRF if we let the number of sites

in the neighborhood to be very large, since for each site in the lattice is associated a conditional probability that this site assumes a value according to the values showed in its neighborhood. Besides that if the neighborhood structure is fixed all over the lattice it is not possible to allow bigger dependency for sites in one region than in another. This can be a serious restriction for modeling some spatial phenomena for example.

One possible solution to this problem is to consider a variation of the MRF model analogous to Variable Length Markov Chain (VLMC) or Probabilistic Context Tree (PCT) initially proposed for Markov chains.

The PCT model for one-dimensional data was introduced by [9] in information theory for binary codes. He introduced the notion of variable memory which means that in order to predict the next symbol is not necessary to keep in memory all the past. The relevant part of the past called "context" can vary from one sequence to another. In this way the set of contexts can have substrings of different sizes and can be represented as a tree. This tree representation is very useful to understand the dependency structure of the source on the past. Processes of this class are still Markovian, but with variable memory length, producing a class of models structurally larger and richer than Markov chains of fixed order. He also introduced the context algorithm to estimate PCT which is able to compress long strings generated by a source.

In theoretical studies, we mention the work of [10], which established new results in processes of infinite dependence trough an adaptation of the Context algorithm. The consistency and some properties of BIC context tree algorithm is shown in [11], [12], [13] and [14]. The lossless compression of

digital contours is considered in [15]. They studied the problem of the chain codes of digital contours in map images. They applied the context tree based approach and provided an optimal algorithm for n-ary incomplete context tree construction. Several studies contributed to this literature in various directions [16], [17] and [18]. In a practical level, we could mention its usages in information theory, focusing on bioinformatics [19], [20], linguistics [21] and universal coding [22].

For MRF processes [23] propose an estimator for a basic neighborhood, based on a modification of the Bayesian Information Criterion (BIC) replacing likelihood by pseudo-likelihood. They also prove that this estimator, called Pseudo-Bayesian Information Criterion (PIC), under certain conditions is strongly consistent for a realization of a field in a growing finite region.

This kind of processes is called Variable Neighborhood Random Field (VNRF) model in [24] where the concept of "context" is extended to a r-dimensional lattice. They propose an estimator for the radius of the basic neighborhood (context) of a site, i.e., the smallest circle containing the context of the site. They still define an algorithm to estimate this radius, and prove the consistency of the estimator.

In this work we propose a different kind of context neighborhood geometry for the MRF. We fix a frame structure for neighborhoods of a site and allow the radius of theses frames to vary according to the values presented in the frame. The advantage is that with this geometry the number of free parameters is reduced and we are able to compute the cardinality of the set of contexts neighborhoods. In this way we can use the PIC estimator in order

to obtain an optimal model given the sample and we also present a graph representing the variable neighborhood structure in a tree format analogous to the one dimensional PCT. We call this model as Probabilistic Context Neighborhood (PCN). Based on this approach we propose an algorithm for estimating the context neighborhoods of a two-dimensional lattice generated by a PCN source. We apply our methodology to simulated data in order to show how well it recovers the parameters of the model.

Besides the estimators for the PCN model and the PCN model selection procedure we also propose a regionalization method when the PCN is not homogeneous on the entire lattice. We allow the PCN source generating the data to vary from one region to another. We call this model as Non-Homogeneous Probabilistic Context Neighborhood (NHPCN). This is done from an adaptation of the Skater method, standing for Spatial 'K'luster Analysis by Tree Edge Removal ([25], [26]), through measures of dissimilarity between trees estimated in different regions of the lattice. [27] proposed an automatic learning of a non-parametric stochastic tree edit distance (ED) to learn the optimal edit costs. They proposed two probabilistic approaches. The first one uses the joint distribution over the edit operations and builds a generative model of the tree ED. The second gives a discriminative model by using a conditional distribution. Unlike dissimilarities commonly used [28], in our work we propose dissimilarity between PCN based on the complexity, structure and conditional probabilities of the PCN from each region.

Finally, we make a simulation study to analyze the adequacy of our work in practice to black and white images. First, we present an example generating a sample via PCN in which the conditional probabilities correspond to

the probabilities of a two-dimensional Ising model [29]. In a second step, we focus our simulations on the recovery of the NHPCN from a sample generated by two different PCN's in distinct regions of the lattice. The first PCN has only one generation, i.e., the value of the site depends only on the first order neighborhood. The second tree has variable neighborhood degrees (first and second order). In our results, we found that our model was able to recover the real tree. This allows us to believe that our method is feasible and useful in practice.

2. Definitions

Let us consider a two dimensional lattice $\mathbb{Z}^2$. The points $i \in \mathbb{Z}^2$ are called sites or areas, where $||i||$ denotes the maximum norm of i, i.e. for $i = (i_1, i_2)$, $||i|| = \max(|i_1|, |i_2|)$ is the maximum of the absolute values of the coordinates of i. The cardinality of a set Δ is denoted as $|\Delta|$. We denote by $\subset$ and $\Subset$ the inclusion and strict inclusion, respectively. Subsets of $\mathbb{Z}^2$ will be denoted by uppercase Greek letters. Thus, if Λ is a finite set of sites, then $\Lambda \Subset \mathbb{Z}^2$.

A random field is a family of random variables indexed by the site i of a lattice, $\{X(i) : i \in \mathbb{Z}^2\}$, where each $X(i)$ is a random variable that takes values in a finite discrete alphabet A. We denote the set of all configurations of the random field as $\Omega = A^{\mathbb{Z}^2}$. For realizations of $X(\Delta)$, we use the notation $a(\Delta) = \{a(i) \in A : i \in \Delta\}$.

The joint distribution of the variables $X(i)$ is denoted as Q:

$$Q(a(\Delta)) = P(X(\Delta) = a(\Delta))$$

,

for $\Delta \subset \mathbb{R}^2$ and $a(\Delta) \in A^{\Delta}$.

In turn, the definition of conditional probability is given by:

$$Q(a(\Delta)|a(\Phi)) = P(X(\Delta) = a(\Delta)|X(\Phi) = a(\Phi))$$

,

for all disjoint regions Δ and Φ where $Q(a(\Phi)) > 0$.

We say that the process is a Markov random field if there exists a neighborhood Γ_i, satisfying for every $i \in \mathbb{R}^2$

$$P(X(i) = a(i)|X(\mathbb{Z}^2 \backslash i) = a(\mathbb{Z}^2 \backslash i)) = P(X(i) = a(i)|X(\Gamma_i)) = a(\Gamma_i)). \quad (1)$$

Where by a neighborhood Γ_i (of the site i) we mean a finite, central-symmetric set of sites with $i \notin \Gamma_i$.

We consider a particular type of neighborhood which we call a frame ∂_i^j defined as a square of side $2j + 1$ less a square of side $2j - 1$ with the same center i, for $j \in \mathbb{N}$. We observe that for $j = 1, 2, ..m$ the frames ∂_i^j are nested sets in the sense that $\bigcap_1^m \partial_i^j = \emptyset$ and $\bigcup_1^m \partial_i^j$ is a square region of the lattice with center in the site i and side $2m + 1$. In this work we consider $\Gamma(i) = \partial_i^j$. Since the geometry of the neighborhood is fixed we only need to know j, the order of the neighborhood, to get the conditional probabilities for a given site i. To simplify notation sometimes we write only ∂^j omitting the site i whenever it is clear. We say that a configuration $a(\partial_i^j)$ is a realization of the process on the subset ∂_i^j. We also denote the union of frames $\bigcup_{s=m}^n \partial^s = (\partial^m \partial^{m+1} \dots \partial^n)$ as $\partial^{m,\dots,n}$. The concatenation of the two configurations $a(\partial^{1,\dots,k})$ and $a(\partial^{m,\dots,n})$ is denoted by $a(\partial^{1,\dots,n})$, which is possible only if $m = k + 1$.

Then we say that $a(\partial^{1,\dots,k})$ is a suffix of $a(\partial^{1,\dots,n})$ if $a(\partial^{1,\dots,n})$ is a concatenation of $a(\partial^{1,\dots,k})$ and $a(\partial^{k+1,\dots,n})$. This defines a order in the space of configurations denoted by $a(\partial^{1,\dots,n}) \succeq a(\partial^{1,\dots,k})$. If the cardinality $|a(\partial^{n,\dots,m})| \geq 0$, then the $a(\partial^{1,\dots,k})$ is a proper suffix.

Definition 1. The subset $\mathcal{T} \subset \cup_{j=1}^{\infty} A^{\partial^{1,\dots,j}}$ is a neighborhood tree if no $a(\partial^{1,\dots,j}) \in \mathcal{T}$ is a suffix of any other $a(\partial^{1,\dots,k}) \in \mathcal{T}$ for $j < k \in \mathbb{N}$.

When a neighborhood tree does not contain proper suffixes it is called irreducible and denoted by $\mathcal{T} \in \mathcal{I}$. The depth of a neighborhood tree is defined as $d(\mathcal{T}) = max_j\{\partial^j \in \mathcal{T}\}$.

Definition 2. A finite configuration $a(\partial^j) \in A^{|\partial^j|}$ is a context neighborhood of a Markov random field if $Q(a(\partial^j) > 0$ and

$$\begin{aligned} P(X(i) = a(i)|X(\mathbb{Z}^2 \backslash i) = a(\mathbb{Z}^2 \backslash i)) &= P(X(i) = a(i)|X(\partial^j) = a(\partial^j)) \\ &= Q(a(i)|a(\partial^j)) \end{aligned} \tag{2}$$

for every $i \in \mathbb{R}^2$ and $j \in \mathbb{N}$. We say that j is the *order* of the context neighborhood $a(\partial^j)$.

This means that the site i depends only on ∂^j and there is no need to inspect the entire lattice to decide the value assumed by $X(i)$. If P satisfy 2, we call this process as Probabilistic Context Neighborhood (PCN).

A set of all context neighborhoods is a neighborhood tree and we denote it by $\mathcal{T}_0$. It is noteworthy that since in this approach the geometry of the

context neighborhood is fixed, the only variation is in j unlike [23] and this implies that we have much less parameters in the model.

Our goal is to estimate the order j of the context neighborhood of a site i, considering that it may change from one site to another according to the change in configurations.

In Figure 1 we can see the structures of neighborhoods of first to third order. Bigger orders may be understood analogously.

From now on we focus on the space of binary states due to its simplicity and because it allows the study of the interesting case of black and white images. An extension to larger state spaces is straightforward. Besides that we consider that only the number of black (or white) sites in the frame is sufficient to provide the conditional probability given the configuration. This is not a restriction in the model and could be easily changed. It is instead an illustration of the model to more realistic situations such as in Ising Model.

Figures 2 and 3 are examples for a lattice with $A = \{-1, 1\}$, where $X(i) =$ -1, if the observed value of *site* i is white, and $X(i) = 1$ if it is black.

By using this type of neighborhood it is possible to draw a PCN analogous to a PCT in one dimensional case as can be seen in the example of Figure 4. We observe that the PCN preserves several of the characteristics of the PCT. We list some of them in the following. Its root is drawn on top of the tree and represents the value of the site i. The first generation node are drawn from the root down and represents the first order neighborhoods. If the first order information is not sufficient to provide a conditional probability for a site then the second order neighborhood is drawn adding a frame to this first order neighborhood and connected to it as a branch. Each node represents

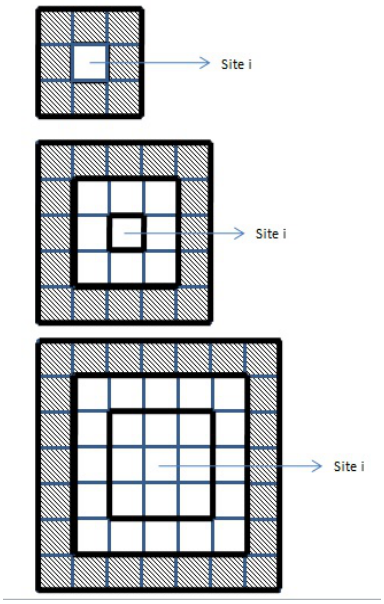


Figure 1: Structure neighborhood for j = 1,2 and 3 respectively.

an added frame.

In the example showed in Figure 4 we can see that the PCN has context neighborhoods of order 1, 2 and 3. There are seven context neighborhood of the first order with 0, 1,2,4,5,7 and 8 black sites respectively (neighborhoods with 3 and 6 black sites are not contexts). Therefore, if we observe only one black site in the neighborhood, the probability of the site being black is known. The same is true for 0, 2, 4, 5 and 8 black sites. If in the first order neighborhood there are 3 or 6 black sites we must continue "down" in the PCN and look at the configurations of the second order. Once we did that, we noticed that the configurations of the first order with 3 and 6 black sites have 3 and 5 children respectively. The children of the configuration with 3 black sites are a context neighborhood and have no children. But

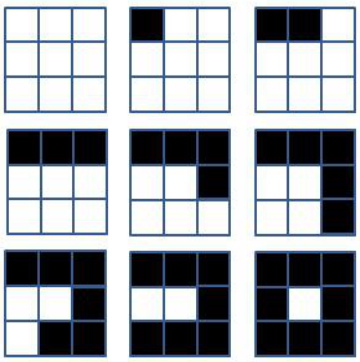


Figure 2: Possible realizations for a first order structure

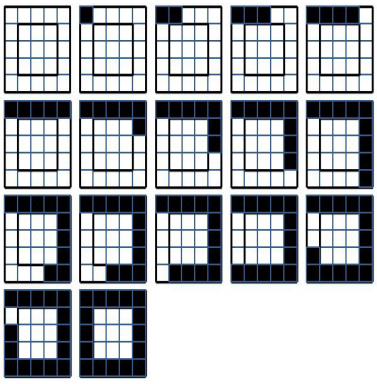


Figure 3: Possible realizations for a second order structure

not all children of the configuration with 6 black sites are considered context neighborhood. There is a child (with 8 black sites) which also has children. In Figure 4 we note that their children have 0 and 4 black sites in the third order and are context neighborhoods. In summary, this PCN has a total of 16 context neighborhoods where 7 of which have first order, 7 have second order and 2 have third order. For each context neighborhood a conditional probability of the central site being black (or white) is assigned.

Note that parent pattern are necessarily contained in children pattern:

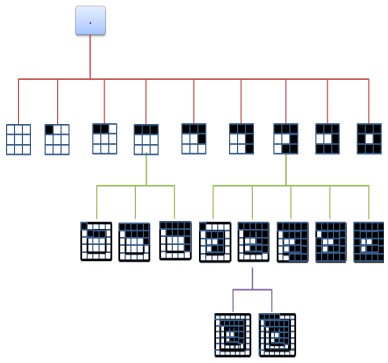


Figure 4: Example of a Probabilistic Context Neighborhood of order 3.

the frame $(\partial^1) \subset (\partial^{1,2})$. This pattern is repeated for all the orders of the PCN, see Figure 4.

In this work, we focus on the estimation of the context neighborhood of the true $\mathcal{T}_0$, from observations of a realization of a Markov field in a finite region. This sample will be denoted as the $a(\Lambda_n)$ and represents the set of n sites under study.

Definition 3. Given a sample $a(\Lambda_n)$, the pseudo-likelihood function associated with the PCN $\mathcal{T}$ and the probability transition function Q is given by:

$$PL_{\mathcal{T}}(a(\Lambda_n)) = \prod_{a(\partial^{1,\dots,j})\in\mathcal{T}, N_n(a(\partial^{1,\dots,j}))>1} \prod_{a(i)\in A} Q(a(i)|a(\partial^{1,\dots,j}))^{N_n(a(\partial^{1,\dots,j},i))} \quad (3)$$

where

$$N_n(a(\partial^{1,\dots,j}, i)) = |\{i \in a(\Lambda_n) : a(\partial^{1,\dots,j}) \subset a(\Lambda_n), a(\partial^{1,\dots,j} \cup i) = a(\partial^{1,\dots,j}, i)\}|$$

represents the number of times that the configuration $a(\partial^{1,\dots,j})$, is observed in the sample when the site i assumes the value $a(i)$ and $N_n(a(\partial^{1,\dots,j}))$ is the number of occurrences of the configuration $a(\partial^{1,\dots,j})$ in the sample $a(\Lambda_n)$,

$$N_n(a(\partial^{1,\dots,j})) = |\{i \in a(\Lambda_n) : a(\partial^{1,\dots,j}) \subset a(\Lambda_n)\}|.$$

The estimator that maximizes the pseudo-likelihood is given by:

$$\hat{Q}(a(i)|a(\partial^{1,\dots,j})) = \frac{N_n(a(\partial^{1,\dots,j}, i))}{N_n(a(\partial^{1,\dots,j}))}.$$

Thus, given a sample $a(\Lambda_n)$, the maximum pseudo-likelihood $MPL_{\mathcal{T}}(a(\Lambda_n))$ is the pseudo-likelihood function evaluated at its maximum:

$$MPL_{\mathcal{T}}(a(\Lambda_n)) = \prod_{N_n(a(\partial^{1,\dots,j}))>1} \prod_{a(i)\in A} \frac{N_n(a(\partial^{1,\dots,j}, i))}{N_n(a(\partial^{1,\dots,j}))}^{N_n(a(\partial^{1,\dots,j},i))}, \quad (4)$$

Definition 4. Given a sample $a(\Lambda_n)$, the PIC (Pseudo-Bayesian Information Criterion) for a PCN $\mathcal{T}$ is:

$$PIC_{\mathcal{T}}(a(\Lambda_n)) = -logMPL_{\mathcal{T}}(a(\Lambda_n) + \frac{(|A|-1)|\mathcal{T}|}{2} log|\Lambda_n|,$$

we stress that, unlike [23], here it is possible to obtain a simple, closed formula for the penalty term since.

$$|\mathcal{T}| = \sum_{k=1}^{max_j a(\partial^{1,\dots,j})\in\mathcal{T}} |A|^{|a(\partial^k)|} = \sum_{k=1}^{max_j a(\partial^{1,\dots,j})\in\mathcal{T}} |A|^{8k} \quad (5)$$

Given a sample $a(\Lambda_n)$, a feasible PCN $\mathcal{T}$ is such that the order $j \leq D(n)$, where $D(n)$ is a function of the sample size, with $N_n(a(\partial^j)) \geq 1$ for every

$a(\partial^j) \in \mathcal{T}$. Besides that each configuration $a(\partial^k)$ with $N_n(a(\partial^k)) \geq 1$, $k < j$ is a suffix of some $a(\partial^j) \in \mathcal{T}$. A family of feasible PCN is denoted by $\mathcal{F}_1(a(\Lambda_n), D(n))$.

Definition 5. We define the PIC estimator of a PCN by

$$\hat{\mathcal{T}}_{PIC}(a(\Lambda_n)) = argmin_{\mathcal{T} \in F_1(a(\Lambda_n), D(n)) \cap \mathcal{I}} PIC_{\mathcal{T}}(a(\Lambda_n)),$$

The consistency of the estimator $\hat{Q}(a(i)|a(\partial^{1,\dots,j}))$ is a consequence of the corollary 2.1 in [23] in which they state the consistency of this kind of estimator for a bigger class of possible neighborhoods provided that $D(n) = (\log(\Lambda_n))^{1/4}$. Under this assumption they also prove the consistency of the PIC estimator for each neighborhood Γ.

Simulation results presented in section 3 leads us to believe that due to simplicity of the frame geometry structure if $D(n) = O(log|\Lambda_n|)$ the estimator $\hat{\mathcal{T}}_{PIC}$ is consistent because we are able to recover the real neighborhood tree using this value.

2.1. Estimation Procedure

According to equation 4, the pseudo maximum likelihood function could be factored as

$$MPL_{\mathcal{T}}(a(\Lambda_n)) = \prod_{a(\partial_i^{1,\dots,j}) \in \mathcal{T}} \tilde{P}_{MPL,\partial_i^{1,\dots,j}}(a(\Lambda_n)),$$

where

$$\tilde{P}_{MPL,\partial_i^{1,\dots,j}}(a(\Lambda_n)) = \begin{cases} \prod_{a(i) \in A} \frac{N_n\left(a(\partial_i^{1,\dots,j}, i)\right)}{N_n(a(\partial_i^{1,\dots,j}))}^{N_n(a(\partial_i^{1,\dots,j}, i))} & \text{if} \quad N_n(a(\partial_i^{1,\dots,j})) \geq 1 \\ 1 & \text{if} \quad N_n(a(\partial_i^{1,\dots,j})) = 0 \end{cases}$$

Using this factorization, we can rewrite the estimator $\hat{\mathcal{T}}_{PIC}(a(\Lambda_n))$ as

$$\hat{\mathcal{T}}(a(\Lambda_n)) = argmax_{\mathcal{T}} \prod_{\partial_i^{1,\ldots,j} \in \mathcal{T}} \tilde{P}_{\partial_i^{1,\ldots,j}}(a(\Lambda_n)),$$

where $\tilde{P}_{\partial_i^{1,\ldots,j}}(a(\Lambda_n)) = n^{\frac{|A-1|}{2}} \tilde{P}_{MPL,\partial_i^{1,\ldots,j}}(a(\Lambda_n))$.

This fact allows the computational treatment for the PIC estimator from an extension of the CTM algorithm [22], [30]. The CTM algorithm is described as follows:

Given a sample $a(\Lambda_n)$, we assign to each node a value and a binary indicator. This assignment is recursive, i.e., the value and the indicator assigned are calculated from the values assigned to the children of this node. The indicator determines the estimator, which assumes a sub-tree form, as follows:

Definition 6. Given a sample $a(\Lambda_n)$, each node (neighborhood) received recursively, from complete tree leaves , the value

$$V^D_{\partial^{1,\ldots,j}}(a(\Lambda_n)) = \begin{cases} max\{\tilde{P}_{\partial^{1,\ldots,j}}(a(\Lambda_n)) \ , \ \prod_{a(i)\in A, N_n(a(\partial^{1,\ldots,j},i))\geq 1} V^D_{\partial^{1,\ldots,j},i}(a(\Lambda_n))\}, \\ \qquad\qquad \text{if } \ 0 \leq j < D \\ \tilde{P}_{\partial^{1,\ldots,j}}(a(\Lambda_n)), \quad \text{if} \quad j = D \end{cases} \tag{6}$$

and the indicator

$$\chi^D_{\partial^{1,\ldots,j}}(a(\Lambda_n)) = \begin{cases} 1 \quad \text{if} \quad \tilde{P}_{\partial^{1,\ldots,j}}(a(\Lambda_n)) < \prod_{a(i)\in A, N_n(a(\partial^{1,\ldots,j},i)\geq 1} V^D_{\partial^{1,\ldots,j},i}(a(\Lambda_n)) \\ \quad \text{and} \quad 0 \leq j < D \\ 0 \quad \text{if} \quad \tilde{P}_{\partial^{1,\ldots,j}}(a(\Lambda_n)) \geq \prod_{a(i)\in A, N_n(a(\partial^{1,\ldots,j},i)\geq 1} V^D_{\partial^{1,\ldots,j},i}(a(\Lambda_n)) \\ \quad \text{and} \quad 0 \leq j < D \\ 0 \quad \text{if} \quad j = D \end{cases} \tag{7}$$

where $D = D(n)$.

The pruning procedure is done starting from the root. If any of the first order neighborhood result in an indicator equal to zero, we keep these nodes and cut the entire second order configuration, which are connected to it. In other words, we exclude children that have parents with indicator equal to zero. We adopt the same procedure to the second order generation nodes that were not pruned: by cutting the third order configuration that has parents with indicator equal to zero. After the pruning procedure, all the nodes of the resulting tree have indicator equal to one and all leaves have indicator equal to zero [12].

2.2. Non Homogeneous Probabilistic Context Neighborhood- NHPCN

In section 2 we propose the PCN model for the two-dimensional lattice, and described a computational method to estimate the PCN from a sample $a(\Lambda_n)$. By applying this methodology, we are able to estimate a single PCN that represents the generating source for all sites in the sample. We now present a model that allows us to have different PCN's according to the sub region of the lattice. We let the conditional probabilities of a site depend not only on the neighborhood but also on the sub region of the lattice in which the site is located.

Definition 7. Consider a partition of the lattice such that $\Lambda_n = \cup_{k=1}^{m} \Delta_k$, $m \leq n$, $\Delta_k \cap \Delta_l = \emptyset$ for all k and l. For each subregion Δ_k and $i \in \Delta_k$

$$P_{\Delta_k}(X(i) = a(i)|X(\mathbb{Z}^2 \backslash i) = a(\mathbb{Z}^2 \backslash i)) \quad = \quad P_{\Delta_K}(X(i) = a(i)|X(\partial^j) = a(\partial^j))$$

$$= \quad Q_{\Delta_k}(a(i)|a(\partial^j)), \tag{8}$$

where $a(\partial^j)$ is a context neighborhood. We call this process as Non-homogeneous Probabilistic Context Neighborhood (NHPCN).

Then in this model we have that

- Sites of the same sub-region should become from the same PCN.
- Sites of different sub-regions should become from dissimilar PCN.

Thus using a NHPCN one might ask if the PCN representing the north of a region is not equal to the PCN representing the south. The same for east and west. But two adjacent sites could have been generated from the same PCN or from a similar PCN.

Our goal now is to propose a procedure for grouping sites into homogeneous and contiguous regions. Thus each region would contain a set of sites governed by the same PCN, and different regions will have different PCN. The problem is to find the appropriate partition of the sample. To solve this problem, we use the SKATER (Spatial K'luster Analysis by Tree Edge Removal) method, suggested by [26], that transforms a region (or lattice) into a graph and splits this graph according to the cost of keeping its edges together [31]. We propose to use a cost based on PCN dissimilarities.

2.3. NHPCN Estimation

In this section we describe the strategy to transform the regionalization of a lattice in a problem of graph partitioning, through the SKATER algorithm ([25], [26]). First it is necessary to introduce a few concepts:

We represent the region of interest through a graph $G = (V, E)$ where each site i is represented by a node (or vertice) v_i and two neighboring sites, i and j, are connected by an edge (or line) (v_i, v_j). We denote V as a set of vertices and E as a set of edges of the graph G. In the present work, two areas are considered neighbors if they have a common border. A *path* of v_i to v_k is a sequence of distinct nodes $v_i, v_j, \ldots, v_k$ which are connected by edges $(v_i, v_j), (v_j, v_l), \ldots, (v_m, v_k)$. A graph is connected if, going from a node v_i to any other node v_j, there is at least one path from v_i to v_j. A *cycle* in a graph is a path where the start node and the end node are the same. A *tree* is a graph that does not have cycles.

A *spatial cluster* is a subset of connected sites. A region is partitioned into spatial clusters when they are disjoint. Our goal is to partition the graph G into c spatial clusters $G1, G2, ..., Gc$, where $\cup_k^c G_k = G$.

A *spanning tree* (ST) of a graph G is a tree connecting all the n sites of G, whereas two sites are connected by one single path and the number of edges is $n - 1$. Removing an edge from the tree, we have two sub-graphs connected which are candidates to become spatial clusters.

Each site $i, i = 1, \ldots, n$, has a vector of attributes (characteristics) given by $\mathbf{b_i} = \{b_{1i}, \ldots, b_{mi}\}$. We associate a cost $d(i, j)$ between the sites (i, j) as the measure of the dissimilarity among these sites through $\mathbf{b_i}$ and $\mathbf{b_j}$. The measure of dissimilarity depends on the situation. When the attributes have comparable scales, a usual choice is the Euclidian distance between the vectors of attributes $\mathbf{b_i}$ and $\mathbf{b_j}$. A *minimum spanning tree* (MST) is a spanning tree with minimal cost, where cost is measured as the sum of dissimilarities between all sites of the tree. If the costs between any sites and

its neighbors are distinct, the MST is unique [32].

The MST is obtained from the graph by the Prim algorithm [33].

In our problem, the attribute of each site is a PCN, as we proposed in section 2. Considering all these concepts, the SKATER procedure can be described in 5 steps:

1. Turn the lattice into a graph;
2. Calculate the cost of the edges (dissimilarity);
3. Reduce the graph to a minimum spanning tree [31] ;
4. Prune the MST and
5. Repeat the previous item until the desired number of clusters is reached.

Figure 5 shows a scheme of the procedure.

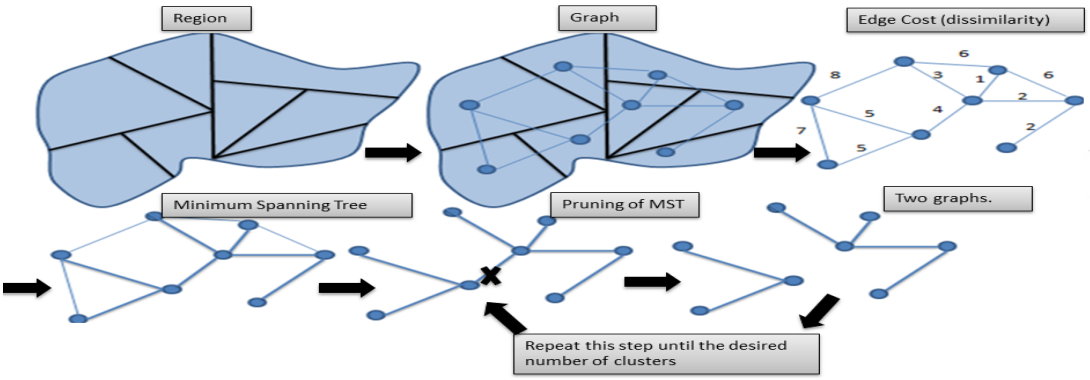


Figure 5: Procedure Skater

We are going to explain the procedure in more details in the following section.

2.3.1. Pruning the MST

After construction of the MST, we move to a step in which we are going to make the pruning of the edges. To partition the MST into c regions, it is

necessary to remove $c-1$ edges of the tree. Each cluster will be a resulting tree connected to all sites, without cycles.

To create a partition of n sites in c trees, we use a strategy of hierarchical division. Initially, all sites belong to a single tree. As the edges of the MST are removed, a set of trees appears disconnected. At each iteration, a tree is divided into two trees by cutting an edge, until we reach the number of clusters previously stipulated.

In this step, to allow more homogeneous and balanced regions in terms of number of sites, the cost is given by:

$$\text{Edge cost } l = SSD_{\mathcal{T}_{MST}} - SSD_l,$$

where

SSD_{MST} is the sum of squares of deviations associated with the MST, given by:

$$SSD_{MST} = \sum_{j=1}^{m} \sum_{i=1}^{n_{MST}} d(b_{ij} - \bar{b}_j),$$

where n_{MST} is the number of sites the MST, b_{ji} is the value of the j-th attribute of the site i; m is the number of attributes considered in the analysis and $\bar{b}_j$ is the average value of j -th attribute in all sites of MST.

SSD_l is the sum of two parts, MST_1 and MST_2, of the two sub-trees generated by the removal of the edge l oh the MST:

$$SSD_l = SSD_{MST_1} + SSD_{MST_2}$$

To obtain the sum of squares of deviations for the two sub-trees we calcu-

lated the average values of m attributes in the same way as in the calculation of SSD_{MST} but considering only attributes related to objects belonging to each sub-tree of MST, MST_1 and MST_2.

After finding each edge cost we remove the one that has minimum cost. Then we repeat the process on each of the sub-tree until a stopping criterion is reached (for example, the desired number of classes).

2.3.2. Dissimilarity between PCN's

The SKATER method is used to make a regionalization of the sample into sub regions. We note that it is possible to work with a vector $\mathbf{b} = \{b_1, \ldots, b_m\}$ for each site. In our work, we will consider as an attribute a sub-PCN as defined in section 2. Thus for each site i, $i = 1, ..., n$ we define a PCN $\mathcal{T}_i$ such that

- $\mathcal{T}_i$ represents a PCN for a sub-lattice centered in the site i with a square geometric shape and base equal to m.
- $\mathcal{T}_i$ has the same properties of the PCN previously defined.

After setting the attribute of each site we define a dissimilarity measure between them given by.

$$d(\mathcal{T}_i, \mathcal{T}_j) = w_1 d_s(\mathcal{T}_i, \mathcal{T}_j) + w_2 d_p(\mathcal{T}_i, \mathcal{T}_j) + w_3 d_c(\mathcal{T}_i, \mathcal{T}_j) \tag{9}$$

where $0 \leq w_k \leq 1$, $\sum_{k=1}^{3} w_k = 1$, $d_s(\mathcal{T}_i, \mathcal{T}_j)$ represents the difference in structure, $d_p(\mathcal{T}_i, \mathcal{T}_j)$ the difference in probability, and $d_c(\mathcal{T}_i, \mathcal{T}_j)$, the difference in complexity. The three differences are defined as:

$$d_s(\mathcal{T}_i, \mathcal{T}_j) = \frac{\left|\{a(\partial^{1,\dots,k}) \in \mathcal{T}_i\} \Delta \{a(\partial^{1,\dots,l}) \in \mathcal{T}_j\}\right|}{|\{a(\partial^{1,\dots,k}) \in \mathcal{T}_i\}| + |\{a(\partial^{1,\dots,l}) \in \mathcal{T}_j\}|}, \tag{10}$$

$$\begin{aligned} d_p(\mathcal{T}_i, \mathcal{T}_j) &= \frac{1}{2|A||\{a(\partial^{1,\dots,l}) \in (\mathcal{T}_i \cup \mathcal{T}_j)\}|} \times \\ &\sum_{a(k)\in A, a(\partial^{1,\dots,l})\in(\mathcal{T}_i \cup \mathcal{T}_j)} \left(\frac{N_{ni}(a(\partial^{1,\dots,l}, k)) - N_{nj}(a(\partial^{1,\dots,l}, k))}{m}\right)^2 \end{aligned} \tag{11}$$

and finally

$$d_C(\mathcal{T}_i, \mathcal{T}_i,) = \frac{\left|C_{\mathcal{T}_i} - C_{\mathcal{T}_j}\right|}{C_{max}}, \tag{12}$$

where $C_{\mathcal{T}_i}$ is the complexity of the i-th tree:

$$C_{\mathcal{T}_i} = \sum_{a(\partial^{1,\dots,k})\in\mathcal{T}_i} k^2 \left|\{a(\partial^{1,\dots,k}) \in \mathcal{T}_i\}\right|. \tag{13}$$

In two-dimensional lattices we have:

$$C_{max} = \sum_{k=1}^{D(n)} k^2 |A|^k, \tag{14}$$

and $D(n)$ is the maximal order for j. In this work we consider $D(n) = \log(|\lambda_n|)$. Finally, we denote the average value of the attribute as $\bar{\mathcal{T}}$, and define it as:

$$\bar{\mathcal{T}} = \text{argmin}_{\mathcal{T}_i} \sum_{j=1}^{n_{MST}} d(\mathcal{T}_j, \mathcal{T}_i). \tag{15}$$

3. Simulation - NHPCN

In this section we present simulations focusing on two goals

- Generating a lattice from a Probabilistic Context Neighborhood;
- Obtaining a Probabilistic Context Neighborhood from a lattice.

Our simulations are all based on the Bivariate Ising Model [29], due to its simple formulation and exact solutions on regular lattices. The Ising Model considers an interaction system of particles (sites), located on a regular lattice. Each site can have one of two orientations, labeled as magnetic spin up (+1) and down (-1).

In this model, each particle interacts only with its nearest neighbors. The contribution of each particle in the total energy of the system depends on the orientation of its spin when compared with its neighbors. Adjacent particles that have the same spin, either both -1 or $+1$, are in a state of lower energy than those with opposite spins. The likelihood of a site being white is a function of the number of neighboring sites black and the parameter β:

$$P\big(X(i) = 1 | X(\mathbb{Z}^2 \backslash i)\big) = \frac{1}{1 + e^{-2\beta s_i}},$$

where s_i is the number of black neighbors minus the number of white neighbors of the site i, $i\ in \mathbb{R}^2$, $X(i)$ is a random variable that can assume the values $\{-1, 1\}$ and $X(i) : i \in \mathbb{R}^2$ is a random field.

The neighborhood of the Ising Model is fixed and can be observed in Figures 2 and 3.

Thus, if $\beta > 0$, the more black neighbors, the higher the probability of being black. Furthermore, the higher β is, more neighbors are going to be similar.

3.1. Generating samples

As an example, Figure 6 shows the simulation results for 9 PCT with same tree structure but different transition probabilities. We considered 64×64 sites of a regular lattice and burn-in of 500 and 1000 iterations. The PCN tree used to generate the sample lattice is represented in the left side.

In the right side of Figure 6 each scenario represents a sample generated with a given value of the parameter $\beta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

3.2. Estimating the PCN

We generate a sample of the Ising Model with 50×50 sites with $\beta = 0.25$, a PCN tree of first-order (shown in the left side of Figure 6) and burn-in of 500.

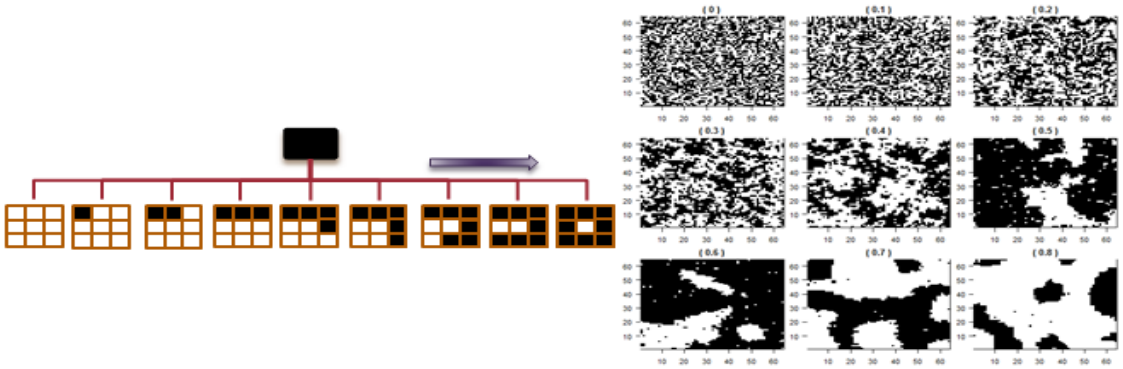


Figure 6: Simulation of Lattices from a Probabilistic Context Neighborhood based on the Ising Model with $\beta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

We can observe that our methodology could recover the same neighborhood structure behind the sample (as shown in the right side of the Figure 7). The two bars that appears below each leaf have an area equal to one and represent the conditional probabilities given the observed neighborhoods. The first bar represents the true conditional probability and the second bar represents the estimation of the conditional probability. Thus, the larger the black portion of the bar, the greater the probability of observing a black site i, given the observed neighborhoods. Concluding, the more similar the two bars are, better is the model and closest are the estimates. Thus our methodology has recovered the true model very well.

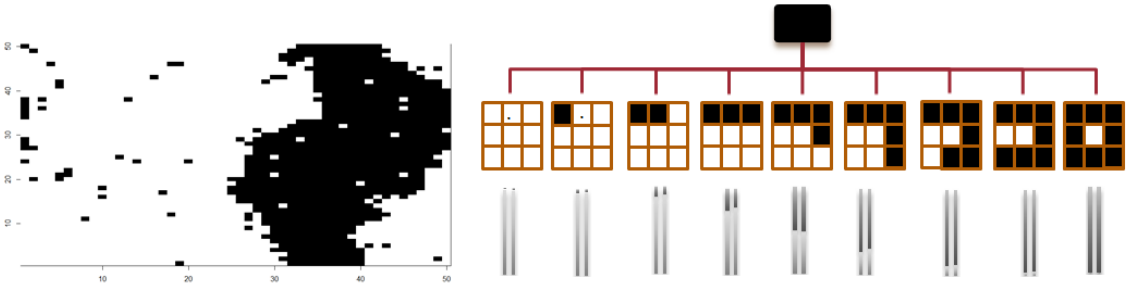


Figure 7: The left hand represents the sample obtained and the right hand represents the estimates of the PCN.

In the second simulation, we generated a sample with 50×50 sites and $\beta = 0.25$. Figure 8 shows the PCN. Due to space limitation, we chose not to draw all context neighborhoods in the PCN of Figure 8. In their place, we created a grey scale that corresponds to the ratio of black in the context neighborhood. Thus, the whiter is the color, the less black sites exist in the context neighborhood. On the contrary, darker the color, the bigger the amount of black sites. Figure 9 shows the scale explained above based on few examples of first and second neighborhoods order.

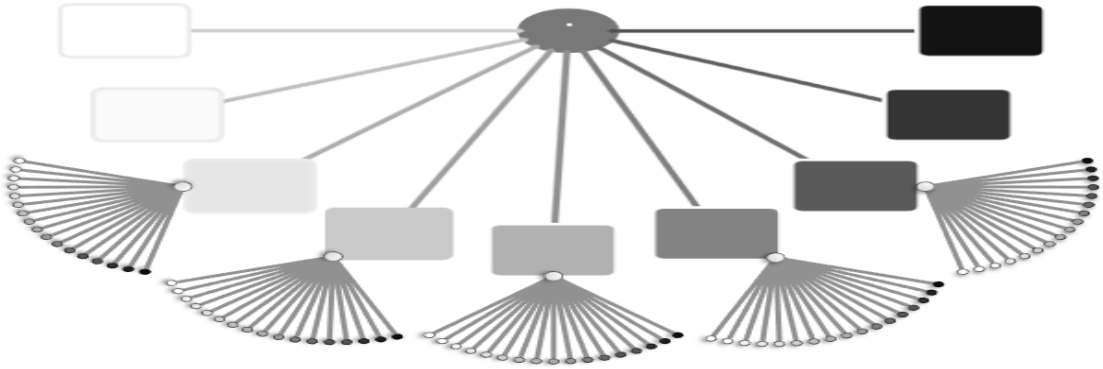


Figure 8: True NHPCN -order one and two- with conditional probabilities based on the Ising Model and $\beta = 0.25$

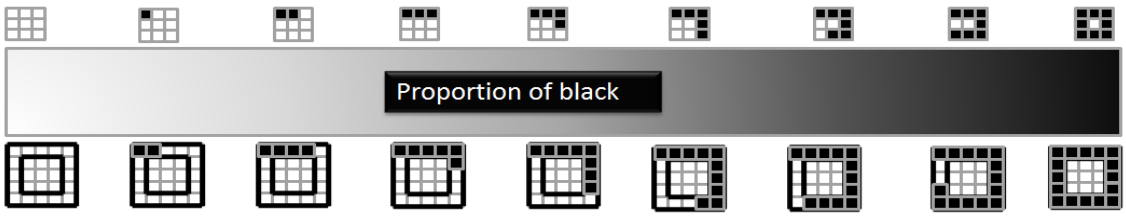


Figure 9: Grey scale representing the proportion of black sites in the context neighborhood.

The sample generated from the structure shown in Figure 10 can be observed in Figure 11. It is clear that the estimated structure is very close to the real. Furthermore, comparing the three bars of the first order neighborhood, we observe that the estimates of the conditional probabilities are quite close to the actual probabilities. Finally, the bars of

the second-order neighborhood were omitted, but it is important to say that they follow the same trend of the first order.

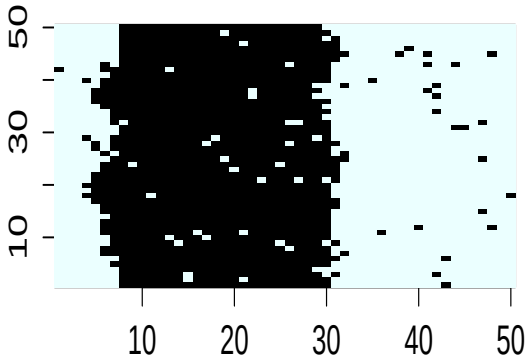


Figure 10: Sample obtained in the second order structure.

4. Analysis Via Monte Carlo Simulation

In this section, we are going to analyze the behavior of the estimation method through Monte Carlo Simulation. We generated 100 samples described in section as 3.2 and estimated 100 PCN models. The results are exhibited in Table 1. The first column corresponds to context neighborhoods of first order, using values of s (the number of black sites, minus the number of whites sites). The second and third columns show the true conditional probability and the average estimated conditional probability in each simulation. Finally, the last column represents the standard deviation of the estimated conditional probabilities.

We conclude that the PCN found in the simulations are very similar to the true PCN. Observe that retrieved neighborhood structure, which can be observed in the first column of Table 1 and the estimates of the conditional probabilities are very close to the real, which can be observed in the second and third columns.

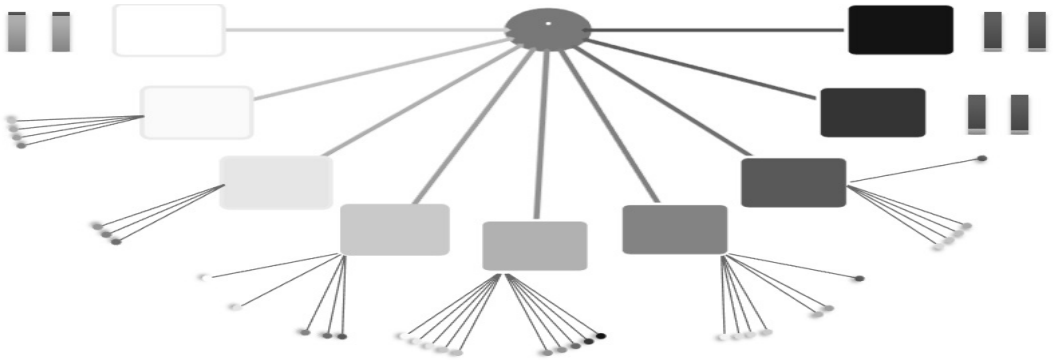


Figure 11: PCN estimated.

Table 1: Results of simulation

Black-White	True probability	Estimates	S.d
0	0.0180	0.0181	0.0057
1	0.0474	0.0454	0.0185
2	0.1192	0.1235	0.0369
3	0.2689	0.2718	0.0552
4	0.5000	0.4955	0.0517
5	0.7311	0.7359	0.0521
6	0.8808	0.8782	0.0356
7	0.9526	0.9531	0.0177
8	0.9821	0.9813	0.0065

5. Discussion

In this paper we propose a Markov model with variable neighborhood in two dimensional lattice, adapting the ideas of one-dimension PCT estimation and using the PIC to perform model selection. By using this model it is possible to present a graphical representation in a tree format of the neighborhood dependency structure of a site in the lattice.

This tree representation is crucial to understand data interactive behavior.

We are able to calculate for this new model named Probabilistic Context Neighborhood the cardinality of the set of context neighborhoods and consequently the number of free parameters because the geometric form of the neighborhoods in this model is fixed. Furthermore we propose an algorithm to estimate the PCN based on the PIC estimator proposed in [23], which uses pseudo-likelihood instead of the likelihood for MRF.

Another contribution of this work is a regionalization scheme based on dissimilarity between Probabilistic Context Neighborhoods. We propose dissimilarity measures that can be more efficient than [28] to capture differences, because it takes into account several characteristics of the PCN's. In addition to the dissimilarity in probability (commonly used), we also include dissimilarity in structure and complexity. These three dissimilarities are the foundation of our method for regionalization if we consider that the lattice is not homogeneous and have different PCN depending on the sub region. We call this process as Non Homogeneous Probabilistic Context Neighborhood. The regionalization method is based on the SKATER procedure, using the PCN as attributes of each site.

Finally, we tested the method in different scenarios. First, we generated a sample from PCN, evaluating the performance of the method using conditional probabilities based on two dimension Ising Model [29]. Second, we analyze the recovery of a NHPCN with two different neighborhood dependencies in two sub regions with order one and order two PCN. We have found that the estimates were very close to the real source in both cases.

We conclude then that our methodology provides good estimates for PCNs and it is also capable to identify sub regions on a two dimensional lattice.

References

[1] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images Pattern Analysis and Machine Intelligence, IEEE Transactions on IEEE 4 (1984) 721-741.

[2] I. Y. Kim, H. S. Yang, An integrated approach for scene understanding based on Markov random field model Pattern Recognition, Elsevier 28 (1995) 1887-1897.

[3] R. Kindermann, J. L. Snell, Markov random fields and their applications, American Mathematical Society Providence RI 1, 1980.

[4] S. K. Kopparapu, U. B. Desai, Bayesian approach to image interpretation, Springer, Powai India, 2001.

[5] S. Z. Li, Markov random field modeling in image analysis, Springer-Verlag London Limited, 2009.

[6] P. L. Dobruschin, The description of a random field by means of conditional probabilities and conditions of its regularity, Th. Prob. and Its Appl. 13 (1968) 197-224.

[7] O. Frank, D. Strauss, Markov graphs, Journal of the american Statistical association, Taylor and Francis Group 81 (1986) 832-842.

[8] S. Parise, M. Welling, Structure learning in markov random fields Advances in Neural Information Processing Systems, Citeseer 29 (2006) 54.

[9] J. Rissanen, A universal data compression system Information Theory, Transactions on IEEE 29 (1983) 656-664.

[10] F. Ferrari,A. J. Wyner, Estimation of general stationary processes by variable length markov chains, Scandinavian Journal of Statistics 30 (2003) 459-480.

[11] I. Csiszár, P. C. Shields, The consistency of the BIC Markov order estimator, The Annals of Statistics, Institute of Mathematical Statistics 28 (2000) 1601-1619.

[12] I. Csiszár, Z. Talata, Context tree estimation for not necessarily finite memory processes, via BIC and MDL, Information Theory, IEEE Transactions on, IEEE 52 (2006) 1007-1016.

[13] A. Garivier, Redundancy of the context-tree weighting method on renewal and Markov renewal processes Information Theory, IEEE Transactions on, IEEE 52 (2006) 5579-5586.

[14] Z. Talata, T. Duncan, Unrestricted BIC context tree estimation for not necessarily finite memory processes Information Theory, ISIT , IEEE International Symposium on, 2009, 724-728.

[15] A. Akimov, A. Kolesnikov, P. Frä nti, Lossless compression of map contours by context tree modeling of chain codes, Pattern Recognition, Elsevier 40 (2007) 944-952.

[16] P. Bühlmann, Efficient and adaptive post-model-selection estimators, Journal of statistical planning and inference, Elsevier 79 (1999) 1-9.

[17] P. Bühlmann, Model selection for variable length Markov chains and tuning the context algorithm, Annals of the Institute of Statistical Mathematics, Springer 52 (2000) 287-315.

[18] A. Garivier, F. Leonardi, Context tree selection: A unifying view, Stochastic Processes and their Applications, Elsevier 121 (2011) 2488-2506.

[19] G. Bejerano, G. Yona, Variations on probabilistic suffix trees: statistical modeling and prediction of protein families, Bioinformatics, Oxford Univ Press 17 (2001) 23-43.

[20] J. R. Busch, P. A. Ferrari, A. G. Flesia, R. Fraiman, S. P. Grynberg, F. Leonardi, Testing statistical hypothesis on random trees and applications to the protein classification problem, The Annals of Applied Statistics, JSTOR 3 (2009) 542-563.

[21] A. Galves, C. Galves, J. E. Garcia, N. L. Garcia, F. Leonardi, Context tree selection and linguistic rhythm retrieval from written texts, The Annals of Applied Statistics, Institute of Mathematical Statistics 6 (2012) 186-209.

[22] F. M. J. Willems, Y. M. Shtarkov, T. J. Tjalkens, The context-tree weighting method: Basic properties, Information Theory, IEEE Transactions on, IEEE 41 (1995) 653-664.

[23] I. Csiszár, Z. Talata, Consistent estimation of the basic neighborhood of Markov random fields, Ann. Statist. 1 (2006) 123-145.

[24] E. Löcherbach, E. Orlandi, Neighborhood radius estimation for variable-neighborhood random fields, Stochastic Processes and their Applications, Elsevier 121 (2011) 2151-2185.

[25] J. P. Lage, R. M. Assunção, E. A. Reis, A minimal spanning tree algorithm applied to spatial cluster analysis, Electronic Notes in Discrete Mathematics, Elsevier 7 (2001) 162-165.

[26] R. M. Assunção, M. C. Neves, G. Câmara, C. Da Costa Freitas, Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees, International Journal of Geographical Information Science, Taylor and Francis 20 (2006) 797-811.

[27] M. Bernard, L. Boyer, A. Habrard, M. Sebban, Learning probabilistic models of tree edit distance, Pattern Recognition, Elsevier 41 (2008) 2611-2629.

[28] G. Mazeroff, V. De, C. Jens, G. Michael, G. Thomason, Probabilistic trees and automata for application behavior modeling, 41st ACM Southeast Regional Conference Proceedings, 2003.

[29] R. Peierls, On Isings model of ferromagnetism, Proc. Camb. Phil. Soc 32 (1936) 477-481.

[30] F. M. Willems, Y. M. Shtarkov, T. J. Tjalkens, Context-tree maximizing, Proc., Conf. Information Sciences and Systems, 2000, 7-12.

[31] M. Maravalle, B. Simeone, R. Naldini, Clustering on trees Computational Statistics and Data Analysis, Elsevier 24 (1997) 217-234.

[32] A. V. Aho, J. E. Hopcroft, J. D. Ullman, Estructura de datos y algoritmos, Addison Wesley Iberoamericana, SA Washington. Cap 6 (1988) 200-251.

[33] R. C. Prim, Shortest connection networks and some generalizations, Bell system technical journal 36 (1957) 1389-1401.

Vizinhanças a priori em Campos de Markov

Aline Piroutek, Renato Assunção e Denise Duarte

25 de setembro de 2013

Resumo

1 Introdução

O mapeamento de doenças tem sido amplamente usado em análises epidemiológicas e nas intervenções realizadas no âmbito da saúde pública. Essa ferramenta é bastante utilizada para descrever variações espaciais das taxas de incidências de doenças, identificar áreas com risco inesperadamente altos e apresentar um mapa de risco de uma região permitindo a adoção de políticas públicas mais eficientes. Uma recente revisão de mapeamento de doenças pode ser visto em [1], [2] e [3], e aplicações em variadas patologias podem ser vistas em [4], [5], [6],[7], [8], [9], [10] e [11].

O modelo mais popular para estimar os riscos relativos foi proposto por Besag, York e Mollié [12]. Esse modelo tem sido estudado em diversos campos da estatística, sendo estendido em métodos de sobrevivência ([13], [14]), modelos multivariados ([14], [15], [16], [17]), modelos espaço-tempo [18], [19], [20], [21]), entre outros. Em seu trabalho, Besag et al. utiliza a decomposição do risco relativo em dois efeitos aleatórios: espacialmente estruturado e espacialmente não estruturado. O primeiro efeito é baseado no modelo conditional autoregressivo (CAR), no qual a dependência espacial é expressa através de uma estrutura Markoviana. Isso significa que o valor do efeito aleatório de uma área, dado o valor de todas as outras, dependente somente de um conjunto reduzido de áreas chamado vizinhança. O segundo efeito assume v.a's iid normal multivariada. Uma característica desse modelo, assim como da maioria dos modelos focados no mapeamento de doenças, é a fixação de uma estrutura de vizinhanças. Mais especificamente, essa estrutura determinística é expressa através de uma matriz pouco flexível e comumente baseada somente em vizinhança de adjacência. A razão para esta escolha inclue sua simplicidade, conveniência e fácil obtenção através de rotinas do GIS (Geographic information system).

Muitos exemplos podem ser levantados apresentando relações entre a taxa de incidência de uma doença e fatores não espaciais ([22], [23],[24] e [25]) , o que torna inadequada a utilização apenas da vizinhança de adjacência.

Em particular, daremos ênfase, neste trabalho, à análise dos casos bronquite e bronquilite aguda devido à crença na relação entre taxa de mortalidade e o tamanho da população,uma vez que, quanto maior a cidade, maior é nível de poluição do ar ([26],[27] e [28]).

As contagens foram realizadas nas 127 microrregiões dos estados do Paraná, do Rio Grande do Sul, de Santa Catarina e de São Paulo, no período de agosto de 2010 até agosto de 2011, cujos dados encontram-se disponibilizados no site do Ministério da Saúde - Sistema de Informações Hospitalares do SUS (SIH/SUS) - e do IBGE - www.ibge.gov.br/home/estatistica/estimativa2010/default.shtm.

Com o fim de motivar, dividimos as cidades em 4 grupos, e plotamos os box-plots dos logaritmos naturais das taxas de mortalidade. O primeiro grupo, representado no primeiro box-plot da Figura 1, é formado pelas microrregiçoes com o tamanho da população pertencente ao primeiro quartil. O segundo grupo, representado no segundo box-plot da Figura 1, é constituído por microrregiões com população pertencente ao segundo quartil, e assim sucessivamente.

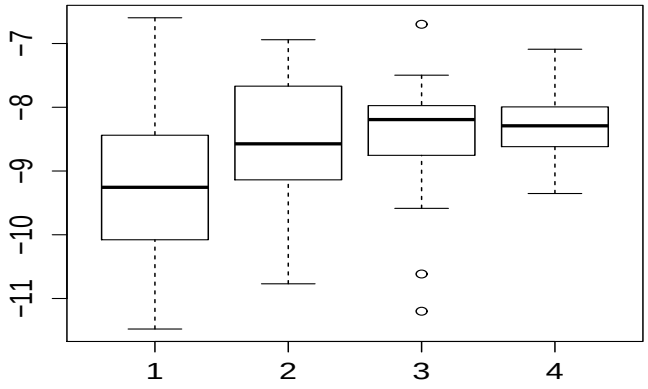


Figura 1: Box-plots do logaritmo natural das taxas de bronquite e bronquilite aguda. Os grupos 1,2,3 e 4 são formados pelas microrregiões com a população pertecente ao primeiro, segundo, terceiro e quarto quartil, respectivamente.

Analisando os dados, fica fácil concluir que, quanto maior a população da cidade, maior é taxa de mortalidade de bronquite e bronquilite agudas.

É de se suspeitar que, nesse caso, a vizinhança baseada na adjacência não seja uma boa opção. Portanto, seria muito útil dispor de um modelo que permite que a estrutura de vizinhança adapta-se automaticamente de acordo com a evidência de dados observado.

De acordo com abordagens em redes bayesianas, aprender a estrutura de um grafo (uma representação da matriz de vizinhança) é uma tarefa onerosa devido ao grande número de grafos possíveis em um conjunto de nós (ou variáveis). Uma das soluções nesse cenário é a utilização de métodos de optimização segundo alguma métrica ([29], [30], [31], [32], [33]). Outro tipo de procedimento tem como base a d-separação (Judea Pearl, [34]), na qual, através do uso de testes estatísticos, as relações condicionais entre nós são identificadas. A partir desse ponto, torna-se indispensável o uso de algoritmos baseados na restrição do número de arestas ou variáveis.

Ainda nessa linha, podemos citar o trabalho de Born and Caron [35] no qual é abordado o problema da aprendizagem estrutural em grafos não direcionados. Eles utilizam modelos de partição produto para encontrar clusters a partir de grafos desconexos. Além disso, eles propõem uma classe de grafos *a priori* baseados no controle de clusterização e nível de separação do grafo.

Já White and Ghosh [36] propõem uma simples extensão do modelo CAR, no qual a seleção da vizinhanças depende da distância entre áreas e de parâmetros desconhecidos. Eles estimam uma medida global que é utilizada para determina se duas áreas são consideradas vizinhas. Dessa forma, até essa distância, o valor da matriz de vizinhança entre as duas áreas é igual a 1, e depois, o valor decai exponencialmente. A escolha dos parâmetros é feita de forma a garantir duas características da matriz de covariância: ser definida positiva e esparsa.

No presente trabalho, propomos um modelo de mapeamento de doenças mais flexível em termos da estrutura de vizinhança. Nossa maior contribuição está nas propostas para as classes de matrizes de vizinhanças *a priori*. Para permitir ainda mais abrangência, propomos também a combinações entre elas. Adicionado a esse fato, usamos a abordagem bayesiana e métodos computacionais (MCM) para encontrar amostras da matriz de vizinhança. Por fim, propusemos dois estimadores *a posteriori* para as amostras das matrizes permitindo, além da estimação do risco relativo, uma análise de influência entre áreas.

2 Mapeamento de doenças

Considere uma região $\mathcal{D} \subset \Re^2$. A região $\mathcal{D}$ é um conjunto finito e enumerável de sítios geográficos (áreas) $D_1, D_2, \ldots, D_n$ onde $\mathcal{D} = D_1 \cup D_2 \cup \ldots, \cup D_n$ com $D_i \cap D_j = \emptyset$ se $i \neq j$. Para facilitar a notação vamos denotar a região D_i por i com $i = 1, \ldots, n$. Denotamos y_i o número observado de casos na região i.

Condicionadas ao vetor de parâmetros $\boldsymbol{\psi}$, modelamos as contagens como variáveis aleatóias independentes segundo uma distribuição de Poisson$(\psi_i E_i)$, onde:

- $E_i = \sum_j pop_{ij} r_j$ é o número esperado de ocorrências na área i sob a hipótese de que risco na área i seja igual ao risco na região total
- $r_j = \dfrac{\sum_{i=1}^{n} y_{ij}}{\sum_{i=1}^{n} pop_{ij}}$ é taxa de incidência da doença em toda região de estudo na faixa etária j para o sexo feminino/masculino;
- pop_{ij} é a população em risco da área i na faixa etária j.

É importante salientar que o risco relativo subjacente é representados por $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$ e, segundo a abordagem Bayesiana, será considerado como um vetor aleatório. Dessa forma, o método é baseado na distribuição *a posteriori* de ψ_i:

$$f(\boldsymbol{\psi}|y) \propto l(y_1, \ldots, y_n) f(\boldsymbol{\psi}), \tag{1}$$

onde $l(y_1, \ldots, y_n)$ é a função de verossimilhança e $f(\boldsymbol{\psi})$ a distribuição *a priori* do vetor de parâmetros $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$.

A modelagem da distribuição *a priori* $f(\boldsymbol{\psi})$ permite introduzir características dos riscos relativos.

A modelagem mais comum é feita a partir de um modelo de efeitos aleatórios, da seguinte maneira:

$$\log \psi_i = \mu + \epsilon_i, \tag{2}$$

em que μ representa a média global do risco relativo e ϵ_i representa o risco específico da i-ésima região.

Uma das distribuições mais utilizada nos estudos estatísticos foi proposto em Besag et al.(1991) [12] no qual o efeito aleatório ϵ é decomposto em duas componentes, uma não estruturada espacialmente $(\boldsymbol{\phi})$ e outra estruturada espacialmente $(\boldsymbol{\theta})$:

$$\log \psi_i = \mu + \phi_i + \theta_i,$$

O primeiro efeito, denotado pela componente $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$, pode decorrer das características individuais das áreas cuja influência se restringe às fronteiras geográficas e podem ser modeladas como efeitos aleatórios independentes (por exemplo, ações de saúde pública não compartilhadas entre regiões). Além disso, é atribuído ao vetor $\boldsymbol{\phi}$ uma distribuição conjunta *a priori* normal multivariada independente, com média 0 e variância $1/\tau_\phi$. O grau de dispersão dos efeitos aleatórios não estruturados espacialmente é controlado pelo parâmetro desconhecido τ_ϕ. Se τ_ϕ é relativamente pequeno, a variabilidade dos efeitos aleatórios $(\phi_1, \ldots, \phi_n)$ será grande em torno de sua média comum igual a zero, significando grande variabilidade dos riscos relativos. Por outro lado, se τ_ϕ é relativamente grande, haverá uma pequena variação desses efeitos em torno de zero.

O segundo efeito, representado pela componente $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, pode ser considerado como efeito aleatório devido à correlação espacial entre áreas. Assim, uma área tende a ser semelhante às áreas vizinhas em termos do risco relativo (por exemplo, índices pluviométricos e temperatura em casos de dengue). As componentes desse vetor possuem distribuições condicionais escolhidas como campos aleatórios de Markov [37] chamado *condicional autoregressivo* CAR (Besag et al.)[38]:

$$\theta_i|(\theta_{-i}, \mathbf{W}, \tau_\theta, \rho) \sim N\left(\frac{\rho \sum_j w_{ij}\theta_j}{\sum_j w_{ij}}, \frac{\tau_\theta^{-1}}{\sum_j w_{ij}}\right) \quad \text{com } i = 1, ..., n. \tag{3}$$

Aqui, $\rho \in (0, 1)$ e τ_θ denota um hiperparâmetro relacionado com a variância de θ_i dado os valores dos outros elementos de $\boldsymbol{\theta}$. O elemento w_{ij} da matriz $\mathbf{W}$ representa a estrutura (grau) de dependência espaciais fixado pelo usuário, o que define quais regiões i e j são vizinhas. A notação $i \sim j$ significa que as áreas i e j são vizinhas sendo que $w_{ij} = 0$ quando essas áreas não vizinhas. Por convenção, consideramos $w_{ii} = 0$ para todo i, ou seja, nenhuma região é vizinha de si mesma. Muitas aplicações consideram essa estrutura de vizinhança baseada em adjacência, onde $w_{ij} = 1$ se a região j é adjacente à região i, e $w_{ij} = 0$ caso contrário.

Segundo o Lemma de Brook [39], a densidade conjunta do vetor de parâmetros $\theta = (\theta_1, \ldots, \theta_n)$ toma a forma:

$$\theta|(\tau_\theta, W, \rho) \sim N(0, (1 - \rho W^*)^{-1}\tau_\theta^{-1} D_{\tau_\theta}), \tag{4}$$

onde $\rho \in (0, 1)$, D_{τ_θ} é uma matriz diagonal com elementos $d_{ii} = 1/n_i$ e $n_i = \sum_{j=1}^{n} w_{ij}$. No caso do uso da matriz de adjacência, n_i representa o número de vizinhos da área i. A matriz W^* representa a matriz estocástica construída a partir de $\mathbf{W}$, onde $w_{ij}^* = w_{ij}/n_i$.

Como visto na equação 3, a distribuição dos θ's depende não somente de uma estrutura de vizinhança espacial mas também de um parâmetro desconhecido adicional τ_θ, inversamente relacionado com a variabilidade de $(\theta_1, \ldots, \theta_n)$.

3 Metodologia

A qualidade das análises dos modelos hierárquicos Bayesianos para mapeamento de dados de área é diretamente afetada pela seleção da matriz $\mathbf{W}$. A escolha da matriz de vizinhança $\mathbf{W}$ é feita de forma determinística e varia segundo os aspectos específicos do problema em análise. A maioria dos modelos para mapeamento de doença escolhe a matriz de vizinhança baseada na adjacência. Isso tem ocorrido devido a facilidade e conveniência trazida a partir dessa escolha. Podemos dizer que estabelecer uma matriz adequada torna-se uma tarefa difícil quando não se conhece a natureza dos

dados. Além disso, também existe a dificuldade ao ter que escolher uma única matriz a ser usada na análise com seus parâmetros e características. Estamos dizendo que essa matriz pode ser considerada um objeto aleatório e que temos que trata-lo como tal. Por esse motivo, nossa proposta é atribuir à matriz aleatória **W** uma classe de distribuição *a priori*. Dessa forma, através dos métodos Bayesianos, encontraremos as distribuições *a posteriori* e as distribuições condicionais completas.

3.1 Mapas como Grafos

Os grafos são estruturas matemáticas utilizadas para modelar pares de relacionamentos entre objetos de uma certa coleção e podem representar a estrutura de vizinhança de uma região.

Assim, cada área é um nó (ou um vértice) do grafo, geralmente localizado no centróide da área. Pares de áreas com $w_{ij} > 0$ são conectadas por uma aresta. Se $w_{ij} = 0$, nenhuma aresta é desenhada. A largura das arestas é proporcional ao valor de w_{ij} . Quando a matriz W é binária, somente são desenhadas as arestas diferentes de zero. Podemos ver na figura 2 um exemplo de grafo construido a partir de um mapa no qual somente ligamos áreas adjacentes (que compartilham fronteira).

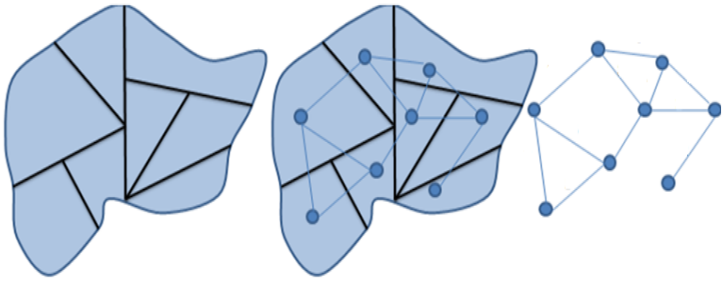


Figura 2: Representação de uma mapa a partir de um grafo.

O mapa em estudo é identificado por um grafo $\mathcal{G}$ que consiste em dois conjuntos V e E, também denotado por $\mathcal{G} = (V, E)$. O espaço de índices $V(\mathcal{G})$ é o conjuto de vértices (ou áreas) de $\mathcal{G}$ e $E(\mathcal{G})$ é um conjunto de pares não ordenados de vértices de $\mathcal{G}$, chamado de arestas: vértices v e v' são ligados por uma aresta se e somente se eles são vizinhos, isto é,$w_{vv'} > 0$.

Um subgrafo de um grafo $\mathcal{G}$ é um grafo $\mathcal{G}'$ tal que $V(\mathcal{G}') \subseteq V(\mathcal{G})$ e $E(\mathcal{G}') \subseteq E(\mathcal{G})$. Um grafo é dito não direcionado se suas arestas não tem direção definida.

Um grafo é dito conexo quando existe um caminho (sequência de vértices) ligando cada par de vértices distintos, caso contrário o grafo é dito desconexo. Um ciclo é um caminho $v_1, \cdots, v_k, v_{k+1}$, onde $v_1 = v_{k+1}$, $k \geq 3$. O grafo que não possui ciclos é dito acíclico.

Dado um grafo $\mathcal{G}$ conexo e não direcionado, uma árvore geradora $\mathcal{T}$ é um subgrafo acíclico que conecta todos os vértices. Assim, em uma árvore, quaisquer dois nós são unidos por um único caminho. Além disso, o número de arestas é igual ao número de nós menos 1. Isso implica que, se qualquer aresta for apagada, a árvore estará desmembrada em duas subárvores desconectadas. Um único grafo pode formar mais de uma árvore geradora.

Nós dizemos que $\mathcal{T}$ é compatível com $\mathcal{G}$, se $\mathcal{T}$ pode ser obtida a partir da poda de arestas de $\mathcal{G}$. Representamos essa definição como $\mathcal{T} \prec G$ quando $\mathcal{T}$ é compatível com $\mathcal{G}$ e $\mathcal{T} \not\prec G$ caso contrário.

3.2 Classe de distribuição de W

Por ser uma abordagem nova, vamos estabelecer algumas notações necessárias:

- A classe de matrizes $\mathbf{W}$ será denotada por $\mathcal{W}$;
- A matriz aleatória será denotada por $\mathbf{W}$;
- Os elementos da matriz $\mathbf{W}$ são representados por $\mathbf{W}_{ij}$;

Seguem alguns exemplos de classes de matrizes que podem ser utilizadas.

1. $\mathcal{W}_1 = \{\mathbf{W} : \mathbf{W} \in \{\mathcal{T} : \mathcal{T} \prec \mathbf{W}^{completo}\}\}$,

   onde $\mathcal{T}$ representa uma árvore geradora do grafo $\mathcal{G}$ e $\mathbf{W}^{completo}$ representa um grafo completo, no qual todas as áreas são vizinhas entre si. Nós dizemos que $\mathcal{T}$) é compatível com $\mathbf{W}^{completo}$, se $\mathcal{T}$ pode ser obtida a partir da poda de arestas de $\mathbf{W}^{completo}$. Como $\mathcal{T}$ é uma ávore geradora, devem sem podadas exatamente $n(n-1)/2-(n-1)$ arestas.

   O Teorema da Matriz-árvore, provado por Kirchhoff em 1847 [40], resolveu o problema de determinação do número de árvores geradoras de um grafo regular.

   Duas exemplos de árvores geradoras possívies a partir de um grafo com 20 áreas pode ser na figura 3.

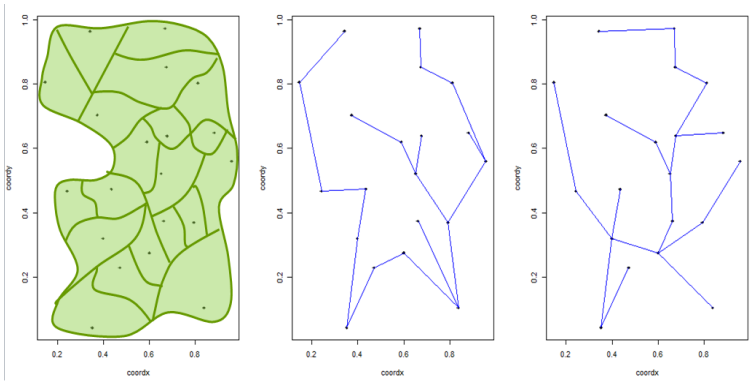


Figura 3: Duas ávores geradoras possíveis em um grafo com 20 nós.

2. $\mathcal{W}_2 = \{\mathbf{W}(k) : \mathbf{W}_{ij}(k) = 1 \text{ se } j \in \{l : d_j(i) \leq d_{(k)}(i)\}, \quad \mathbf{W}_{ij}(k) = 0 \quad c.c.\}$.

   onde $d_l(i) = ||i - l||$ com $l \neq i$ e $d_{(k)}(i)$ é a k-ésima estatística de ordem. Essa classe contém àqueles grafos que ligam somente os $k \in \{1, 2, \ldots, n-1\}$ vizinhos mais próximos de cada área. A figura 4 apresenta exemplos dessa classe em um mapa transformado em grafo no qual variamos os valores de k.

   A seguir, apresentamos duas alternativas para distribuições de k :

   - $k \sim U_d\big(1, (n-1)\big)$,
   - $k = 1 + k^*, \quad com \quad k^* \sim Bin\big((n-2), p\big) \quad e \quad p \in (0, 1)$,

     onde U_d representa a distribuição uniforme discreta. O hiperparâmetro p varia de acordo com o interesse do pesquisador. Se for mais plausível poucas arestas, ou seja, se não existir muitas conexões entre áreas do gráfico, o valor de p será próximo de zero. De outro modo, se na região em análise, as áreas estiverem muito próximas uma das outras, é intuitivo pensar que cada área se relacionará com muitas outras ao seu redor, fazendo com que existam muitas arestas no grafo. Nessa situação, valores de p não pequenos são mais adequados.

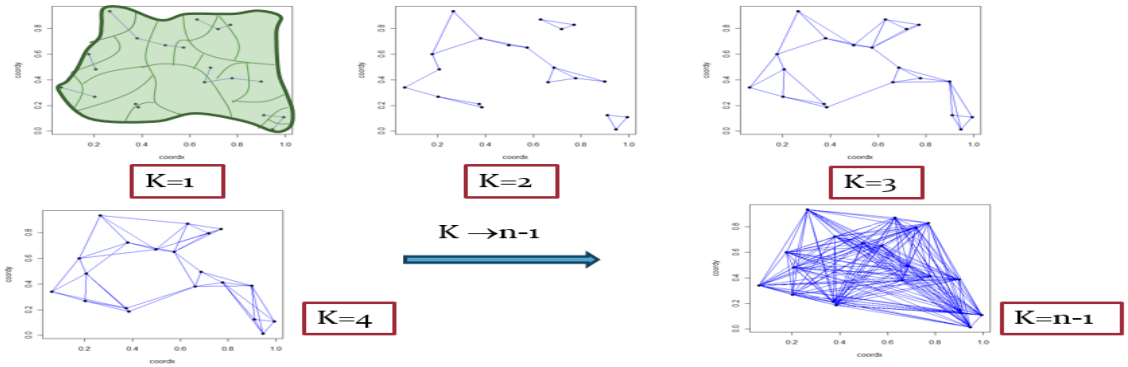


Figura 4: Possíveis grafos a partir da classe 2.

3. $\mathcal{W}_3 = \{\mathbf{W}(d) : \mathbf{W}_{ij}(d) = 1 \quad se \quad d_j(i) \leq d; \quad \mathbf{W}_{ij}(d) = 0 \quad c.c.\}$.

   Essa classe também incorpora um hiperâmetro desconhecido de distância $d \in \mathbb{R}^+$.

   A figura 5 apresenta exemplos dessa classe em um mapa transformado em grafo no qual variamos os valores de d. Observe que, mesmo variando o valor de d entre duas estatísticas de ordem consecutivas, o mapa não se modifica, ou seja, não é acrescentado nenhuma aresta. Dessa forma, o grafo apresenta somente uma aresta quando o valor de d permanece entre $d_{(1)}(\bullet)$ e $d_{(2)}(\bullet)$. Para que seja acrescentado exatamente uma outra aresta, o valore de d deve se localizar necessariamente no intervalo $[d_{(2)}(\bullet), d_{(3)}(\bullet)]$.

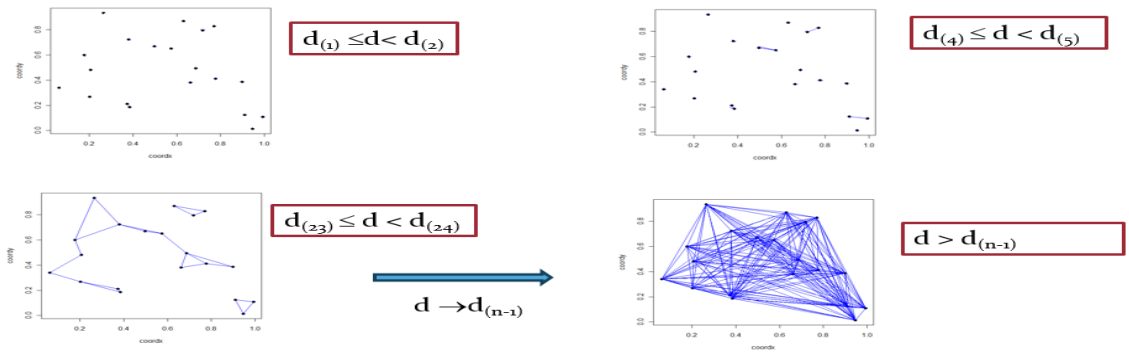


Figura 5: Possíveis grafos a partir da classe 3.

Pode-se, por exemplo, atribuir uma distribuição para o hiperparâmetro d da seguinte forma:

- $d \sim U(d_{(1)}(\bullet), d_{(n)}(\bullet))$, ou seja, a distância mínima e máxima entre todos os possíveis pares de vértices do grafo.
- $d = d_{(n)}(\bullet) \times d^*$, com $d^* \sim Beta(\alpha, \beta)$

  Os hiperparâmetros α e β permitem maior flexibilidade na escolha de d. Se α for igual a um, daremos mais peso para pequenos valores de d. Se também fixarmos β igual a um, estaremos no caso onde todos os valores possíveis de p tem o mesmo peso.

4. (a) $\mathcal{W}_4 = \{\mathbf{W}(\mathcal{C}) : \mathcal{T}_{\mathcal{AG}} \prec \mathbf{W}(\mathcal{C}) \prec \mathbf{W}^{adj}\}$,

   onde $\mathbf{W}(\mathcal{C})$ representa matriz compatível com $\mathbf{W}^{adj}$ na qual é obtida a partir da poda de $\mathcal{C}$ arestas de $\mathbf{W}^{adj}$. Pode ser atribuída uma distribuição para o hiperparâmetro $\mathcal{C}$.

- $\mathcal{C} \sim U_d\Big(0, \big[\mathcal{C}_{max} - (n-1)\big]\Big)$,
  onde $\mathcal{C}_{max}$ representa o número de arestas do grafo de adjacência. Devemos subtrair $(n-1)$, pois esse é o número de arestas que uma árvore possui.
- $\mathcal{C} \sim Binomial\Big(\big[\mathcal{C}_{max} - (n-1)\big], p\Big)$.
  O hiperparâmetro p permite maior flexibilidade na escolha do número de arestas $\mathcal{C}$, as quais serão cortadas da matriz de adjacência $\mathbf{W}^{adj}$.

(b) $\mathcal{W}_5 = \{\mathbf{W}(\mathcal{C}) : \mathcal{T}_{\mathcal{AG}} \prec \mathbf{W}(\mathcal{C}) \prec \mathbf{W}^{completo}\}$,
onde $\mathbf{W}^{completo}$ representa um grafo completo, no qual todas as áreas são vizinhas de toda as outras áreas. O grafo $\mathbf{W}(\mathcal{C})$ representa aquele grafo no qual foi acrescentado C arestas em uma árvore geradora do grafo $\mathcal{G}$. Um exemplo dessa classe pode ser observado na figura 6.

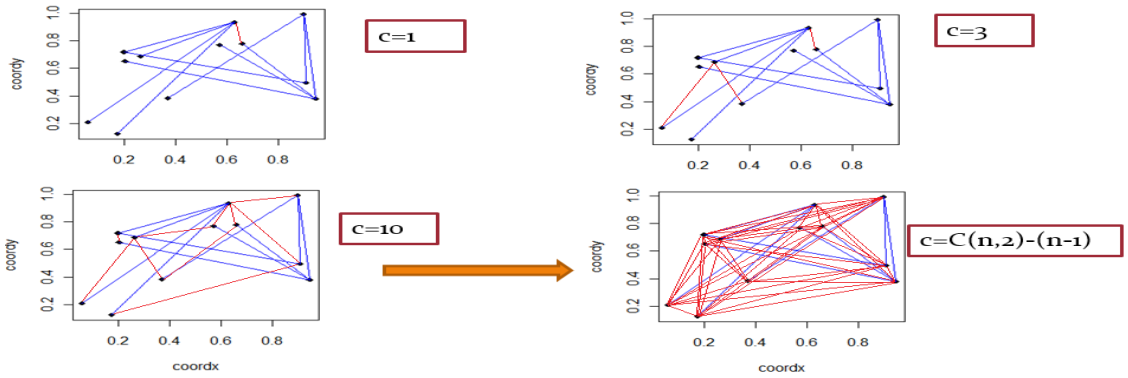


Figura 6: Possíveis grafos a partir da classe $\mathcal{W}_5$.

Podem ser sugeridas as seguintes distribuições para o hiperparâmetro $\mathcal{C}$:

- $\mathcal{C} \sim U_d\left(0, \left[\binom{n}{2} - (n-1)\right]\right)$. O número de arestas de uma grafo completo é igual a $\binom{n}{2}$ e $(n-1)$ é o número de arestas que uma árvore possui.
- $\mathcal{C} \sim Binomial\left(\left[\binom{n}{2} - (n-1)\right], p\right)$. O hiperparâmetro p permitem maior flexibilidade na escolha do número de arestas $\mathcal{C}$ que serão cortadas da matriz de adjacência $\mathbf{W}^{adj}$.

Apesar das classes $\mathcal{W}_4$ e $\mathcal{W}_5$ parecerem iguais, existe uma diferença muito importante entre elas. As duas classes representam o conjunto de matrizes que podem ser formadas a partir da poda de uma matriz específica. O número máximo de arestas tiradas é calculado de forma a garantir que o grafo resultante seja todo conexo, ou seja, obtenha-se uma árvore geradora. Na classe $\mathcal{W}_4$, as arestas são baseadas em fronteiras, ou seja, se existe alguma aresta entre duas áreas, significa que, obrigatoriamente elas são adjacentes. Já na classe $\mathcal{W}_5$, isso não é uma regra. Existe a possibilidade de duas áreas serem ligadas por uma aresta e elas não dividirem fronteira.

(c) $\mathcal{W}_6 = \{\mathbf{W}(\mathcal{C}) : \mathbf{W}(\mathcal{C}) \prec \mathbf{W}^{adj}\}$.
Onde o hiperparâmetro $\mathcal{C}$ pode ser distribuído das seguintes formas:

- $\mathcal{C} \sim U_d\Big(0, \mathcal{C}_{max}\Big)$, onde $\mathcal{C}_{max}$ representa o número de arestas do grafo de adjacência.
- $\mathcal{C} \sim Binomial\Big(\mathcal{C}_{max}, p\Big)$. O hiperparâmetro p permitem maior flexibilidade na escolha do número de arestas $\mathcal{C}$ que serão cortadas da matriz de adjacência $\mathbf{W}^{adj}$.

(d) $\mathcal{W}_7 = \{\mathbf{W}(\mathcal{C}) : \mathbf{W}(\mathcal{C}) \prec \mathbf{W}^{completo}\}$.

As seguintes distribuições para o hiperparâmetro $\mathcal{C}$ são sugeridas:

- $\mathcal{C} \sim U_d\left(0, \binom{n}{2}\right)$,
- $\mathcal{C} \sim Binomial\left(\binom{n}{2}, p\right)$.

As classes $\mathcal{W}_6$ e $\mathcal{W}_7$ são variações das classes $\mathcal{W}_4$ e $\mathcal{W}_5$, respectivamente. A modificação feita nessas classes permite que sejam retiradas todas as arestas dos grafos. Ressalta-se, contudo, que ao utilizar essas classes, os grafos resultantes podem não ser conexos.

5. $\mathcal{W}_8 = \{\mathbf{W}(r) : \mathcal{T}_{\mathcal{AG}} \prec \mathbf{W}(r) \prec \mathbf{W}^{completo}\}$,

onde $\mathbf{W}(r)$ corresponde a um grafo no qual todos os vértices estão ligados a r outros vértices. O hiperparâmetro r pode assumir:

- $r \sim U_d\left(0, \binom{n-1}{2}\right)$,
- $r \sim Binomial\left(\binom{n-1}{2}, p\right)$.

6. (a) $\mathcal{W}_9 = \{\mathbf{W}(h) : W_{ij}(h) = 1 \quad se \quad g(x_i, x_j) \geq h, \quad W_{ij}(h) = 0 \quad c.c\}$,

onde $x = x_1, \ldots, x_n$ é o vetor de observações da covariável X e $g(x_i, x_j) = |x_i - x_j|$. Essa classe torna possível adicionar outras informações importantes na construção de uma matriz de vizinhança.

Suponha três cidades equidistantes A, B e C. As cidades B e C são cidades pequenas, com pouca estrutura, enquanto a cidade A é de grande porte. Quando um indivíduo das cidades pequenas fica doente, é obvio que será levado para a cidade grande. Dessa forma, a falta de estrutura de uma cidade pequena faz com que exista uma "ligação"com uma cidade grande. Assim, a cidade A é vizinha de B e C e essas duas últimas não são vizinhas entre si. Nesse caso, a covariável X representa a população de cada cidade. Desse modo, quando a diferença entre os tamanhos for maior que um limite h, essas cidades serão interligadas.

(b) $\mathcal{W}_{10} = \{\mathbf{W}(h) : W_{ij}(h) = g(x_i, x_j)\}$,

onde $g(x_i, x_j) = I_{[x_i \geq h]} * I_{[x_j \geq h]}$ é o produto de funções indicadoras denotadas por I.

Essa é uma variação da classe $\mathcal{W}_9$, na qual as áreas somente serão consideradas vizinhas quando o valor observado de suas respectivas covariáveis ultrapassar um limite h. Como exemplo, podemos imaginar como ocorreu a disseminação da gripe suína (H1N1) em 2009. Devido ao trânsito de pessoas vindas dos mais variados lugares, cidades com aeroportos internacionais foram muito afetadas pelo vírus, em razão da facilidade na disseminação da doença [41], [42]. Por esse motivo, a Cidade do México rapidamente afetou várias cidades que possuem aeroportos internacionais [43]. Nesse caso, a Cidade do México e São Paulo, por exemplo, mesmo não sendo adjacentes ou próximas, são consideradas vizinhas. A covariável utilizada poderia ser o produto interno bruto (PIB) de cada cidade.

Um exemplo dessa classe pode ser visualizado na figura 7, no qual adotamos a covariável x como tamanho populacional das capitais do Brasil e o hiperparâmetro $h = 2,4$ milhões de habitantes.

Outros exemplos que podem ser utilizados como covariáveis seriam: área, tamanho, renda per capita, IDH, clima, entre outros.

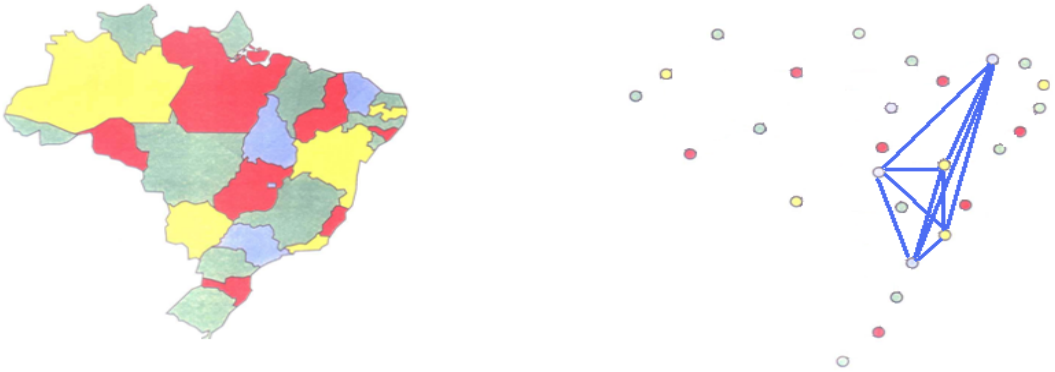


Figura 7: Possível grafos a partir da classe 10.

7. **Combinações de Classes:** Existem situações em que não é suficiente utilizar uma única classe. Podemos, por exemplo, estar analisando o tráfego de entregas (correio) em uma região. Temos nesse caso dois tipos de logísticas de distribuição de produtos. A primeira delas utiliza rotas rodoviárias, no qual as empresas tendem a escolher rotas curtas entre cidades. Esse tipo de distribuição pode ser expresso através da classe $\mathcal{W}_1$ de árvore geradora mínima, na qual o custo das arestas pode estar relacionado com distancias entre cidades. O segundo tipo de entregas é feito através do tráfego aéreo entre cidades de grande porte. Esse tipo de distribuição é indispensável, uma vez que o volume de entregas oriundas de cidades grandes é bastante superior quando comparado ao volume de cidades pequenas. A distribuição via aviões pode ser representada a partir da classe $\mathcal{W}_{10}$. Dessa forma, torna-se imprescindível a utilização de ambas as classes, veja na figura 8. A alternativa para esses casos é a combinação entre classes, representada a seguir:

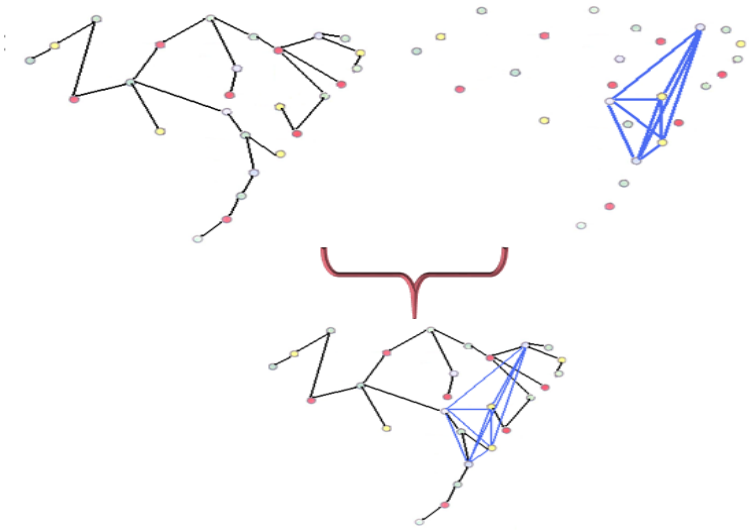


Figura 8: Possível grafos a partir da combinação das classe $\mathcal{W}_1$ e $\mathcal{W}_{10}$.

$\mathcal{W}_{11} = \{\mathbf{W} : \lambda_1\mathbf{W}_1 + \lambda_2\mathbf{W}_{10}\}$,

onde:

- $\mathbf{W}_1 \in \mathcal{W}_1$
- $\mathbf{W}_{10} \in \mathcal{W}_{10}$

- $\lambda_i \in (0,1)$ com $i = 1, 2$
- $\lambda_1 + \lambda_2 = 1$

Os parâmetros λ_i's controlam o peso que será dado a cada matriz de vizinhança. No nosso exemplo, a matriz $\mathbf{W}_1$ é conexa e a matrix $\mathbf{W}_{10}$ não é conexa. É possível acrescentar outras matrizes nessa combinação, desde que sejam mantidas as restrições $\sum_i \lambda_i = 1$ e $\lambda_i \geq 0$.

Depois de definir a classe $\mathcal{W}$, podemos reescrever o modelo hierárquico da seguinte forma:

$$y_i|(\phi_i, \theta_i, E_i) \sim Poisson(E_i e^{\phi_i + \theta_i}) \quad iid\ com\ i = 1, ..., n; \tag{5}$$

$$\phi_i|\tau_\phi \sim N(0, \tau_\phi^{-1}) \quad iid\ com\ i = 1, ..., n; \tag{6}$$

$$\tau_\phi|(a, b) \sim Gamma(a, b); \tag{7}$$

$$\theta_i|(\theta_{-i}, \mathbf{W}, \tau_\theta, \rho) \sim N\left(\frac{\rho \sum_j w_{ij}\theta_j}{\sum_j w_{ij}}, \frac{\tau_\theta^{-1}}{\sum_j w_{ij}}\right) \quad com\ i = 1, ..., n; \tag{8}$$

$$\tau_\theta|(c, d) \sim Gamma(c, d) \tag{9}$$

$$\mathbf{W} \in \mathcal{W} \tag{10}$$

Dessa forma, a distribuição *a posteriori* dos parâmetros e hiperparâmetros a menos de uma constante de integralização é dada por:

$$\begin{aligned} &f(\phi_1, \ldots, \phi_n, \theta_1, \ldots, \theta_n, \tau_\phi, \tau_\theta, \mathbf{W}|y_1, \ldots, y_n) \\ &\quad \propto \Big(\prod_{i=1}^{n} \frac{e^{-E_i e^{\phi_i + \theta_i}} (e^{\phi_i + \theta_i} E_i)^{y_i}}{y_i!}\Big) f(\phi_1, \ldots, \phi_n|\tau_\phi) f(\theta_1, \ldots, \theta_n|\tau_\theta, \mathbf{W}) f(\tau_\phi) f(\tau_\theta) f(\mathbf{W}). \end{aligned} \tag{11}$$

Para facilitar a notação faremos $f(\mathbf{W}|\phi_1, \ldots, \phi_n, \theta_1, \ldots, \theta_n, \tau_\phi, \tau_\theta, y_1, \ldots, y_n) = f(\mathbf{W}|\gamma_{-\mathbf{W}}, \mathbf{y},)$. Assim, a condicional completas de $\mathbf{W}$ é dada por:

$$f(\mathbf{W}|\gamma_{-\mathbf{W}}, \mathbf{y}) \quad \propto \quad \det\left(2\pi\ (1 - \rho\mathbf{W}^*)^{-1} \tau_\theta^{-1} D_{\tau_\theta}\right)^{-1/2} e^{\frac{-1}{2}\theta^t [\tau_\theta D_{\tau_\theta}^{-1}(1 - \rho\mathbf{W}^*)]\theta} f(\mathbf{W}). \tag{12}$$

3.3 Estimadores

Como parte do procedimento de estimação, é feita a amostragem de todos os parâmetros e hiperparâmetros apresentados na seção 2. Os parâmetros τ_ϕ e τ_θ, com condicionais completas e conhecidas, são estimados diretamente pelo amostrador de Gibbs. Já a inferência sobre os demais parâmetros, $\mathbf{W}$, $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$, é feita através do MCMC.

Ao utilizar o algoritmo de Metropolis-Hasting, obtivemos uma amostra de $\mathbf{W}$ segundo a distribuição *a posteriori* demonstrada na equação 12.

A partir dessa amostra de tamanho m, nosso objetivo é obter uma medida resumo para a amostra de matrizes, tais como média ou mediana *a posteriori*. Primeiramente, definimos $Q(\mathcal{G}_i, \mathcal{G}_j)$ como uma função que mede a dissimilaridade entre os grafos $\mathcal{G}_i$ e $\mathcal{G}_j$, tal que:

$$Q(\mathcal{G}_i, \mathcal{G}_j) = \frac{\left| E(\mathcal{G}_i) \backslash E(\mathcal{G}_j) \bigcup E(\mathcal{G}_j) \backslash E(\mathcal{G}_i) \right|}{|E(\mathcal{G}_i)| + |E(\mathcal{G}_j)|}, \tag{13}$$

onde $E(\mathcal{G}_i)$ e $E(\mathcal{G}_j)$ representam os conjuntos de arestas do grafo $\mathcal{G}_i$ e $\mathcal{G}_j$ respectivamente. A diferença entre os dois conjuntos $E(\mathcal{G}_i) \backslash E(\mathcal{G}_j)$ é o conjunto das arestas que pertencem a $\mathcal{G}_i$ e que não pertencem a $\mathcal{G}_j$.

O estimador a posteriori $\hat{\mathcal{G}}$ em uma amostra de tamanho m é dado por :

$$\hat{\mathcal{G}} = argmin_{\mathcal{G}_i} \sum_{j=1}^{m} Q(\mathcal{G}_i, \mathcal{G}_j). \tag{14}$$

Logo, o estimador *a posteriori* será aquele que apresenta a menor dissimilaridade entre todos os grafos amostrados via MCMC.

Para o cáculo do segundo estimador, vamos definir a matriz de caminhos $P(\mathcal{G}_i)_{nxn}$ da seguinte forma:

$$p_{kl} = \text{argmin}_r \sum_{t=1}^{n} w_{kt}^{*r} w_{tl}^{*r} > 0$$

onde $\mathbf{W}$ é a matriz de adjacência induzida pelo grafo $\mathcal{G}_i$, $\mathbf{W}^*$ é a matriz $\mathbf{W}$ padronizada por linhas e $\mathbf{W}^{*r}$ é a r-ésima potência da matriz estocastica $\mathbf{W}^*$. Isso implica que os elementos da matriz caminhos $P(\mathcal{G}_i)$ indicam a número mínimo de passos necessários para ir de um nó a outro em um grafo $\mathcal{G}_i$. Assim, vizinhos de primeira ordem (adjacentes) precisam de apenas um passo para se interligarem e vizinhos de segunda ordem (vizinho do vizinho) precisam de dois passos, $p_{kl} = 2$, e assim sucessivamente.

A partir da matriz de caminhos P_{nxn}, definimos uma segunda função para medir a dissimilariade entre grafos dada por:

$$Q_p(\mathcal{G}_i, \mathcal{G}_j) = \sum_{k=1}^{n} \sum_{l=1}^{n} (p(\mathcal{G}_i)_{kl} - p(\mathcal{G}_j)_{kl})^2 \tag{15}$$

O estimador a posteriori $\hat{\mathcal{G}}_p$ em uma amostra de tamanho m é dado por:

$$\hat{\mathcal{G}}_p = argmin_{\mathcal{G}_i} \sum_{j=1}^{m} Q_p(\mathcal{G}_i, \mathcal{G}_j). \tag{16}$$

Logo, o estimador *a posteriori* será aquele que apresenta a menor dissimilaridade entre todos os grafos amostrados via MCMC.

4 Exemplos - Preliminares

1. Em uma análise preliminar, iremos apresentar exemplos nos quais foram gerados vetores $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ que seguem uma distribuição CAR com os seguintes valores fixados:

$$\theta_i|(\theta_{-i}, \mathbf{W}, \tau_\theta, \rho) \sim N\left(\frac{\sum_j w_{ij}\theta_j\rho}{\sum_j w_{ij}}, \frac{\tau_\theta^{-1}}{\sum_j w_{ij}}\right), \tag{17}$$

onde:

- O número de áreas n é igual a 36,
- As áreas estão dispostas em forma regular em um lattice.
- $\rho = 0.7$,
- $\tau_\theta = 1$
- O grafo de vizinhança real é apresentado no primeiro grafo da figura 9,
- O tamanho da amostra $m = 1000$,
- A classe escolhida para fazer a amostra dos grafos de vizinhaça $\mathbf{W}$ foi a $\mathcal{W}_1$ com uma restrição de adjacência. Isso significa que os grafos candidatos são árvores geradoras. Esses grafos tem poucas arestas (35) que somente ligam duas áreas adjacentes.

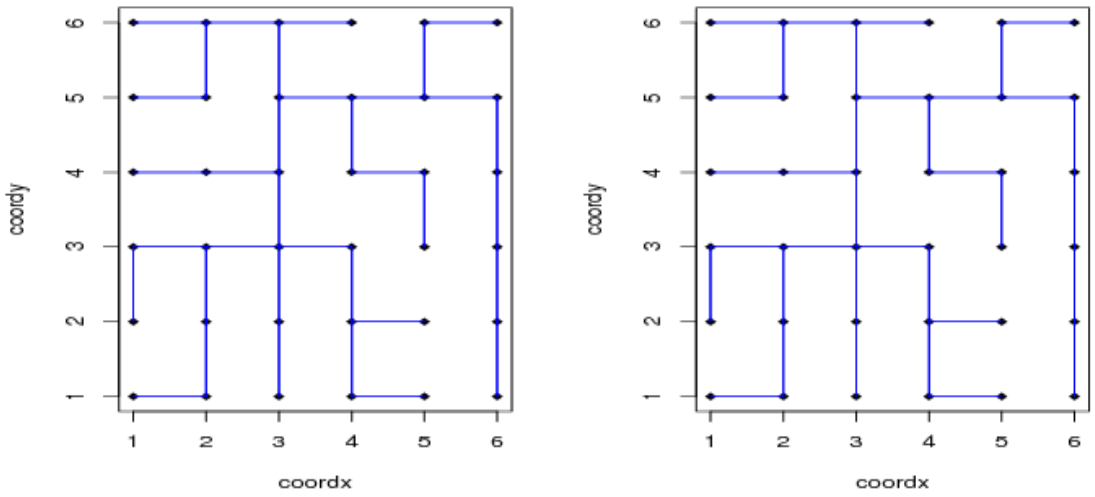


Figura 9: Simulação 1 - Grafo real e grafo estimado $\hat{\mathcal{G}}$.

A partir da figura 9 que o nosso modelo conseguiu recuperar a correta estrutura de vizinhança.

2. No segundo cenário acrescentamos uma covariável que representa a população em cada área. Nesse caso, implementamos um cenário em que uma cidade grande exerce uma influência em pequenas cidades vizinhas. Essa influência seria expressa através do grafo, no qual existiria uma aresta entre a cidade grande e as cidades vizinhas. Acrescentamos os seguintes parâmetros: pop_i representa o tamanho da população da área i e foi obtido através de uma distribuição de poisson com média igual a 50. A cidade polo, fixada, foi escolhida aleatoriamente e o valor de sua população foi multiplicado por 100. Já para o procedimento MCMC, denotamos por $\mathcal{P}$

o conjunto de cidades que serão amostradas como possíveis cidades polo. Para esse exemplo, fixamos a cardinalidade $|\mathcal{P}| = 1$. A classe escolhida será a união da classe $\mathcal{W}_1$ (com uma restrição de adjacência) com uma classe nova $\mathcal{W}'$, definida como:

$$\mathcal{W}' = \begin{cases} \mathbf{W}(k) : w_{ij} = 1 & \text{se } i \in \mathcal{P} \text{ e } j \in \{l : d_l(i) \leq d_{(k)}(i)\} \\ \mathbf{W}(k) : w_{ij} = 0 & c.c \end{cases}$$

onde $d_l(i)$ e $d_{(k)}(i)$ foram definidos em 3.2, $P(i \in \mathcal{P}) = pop_i / \sum_{i=1}^{n} pop_i$ e $k \sim Binomal((n-2), 0.35)$.

O grafo estimado é apresentado na figura 10. Embora não seja idêntico, apresenta grandes semelhanças. Ressalta-se que as informações trazidas a partir do vetor $\boldsymbol{\theta}$ foram suficientes para que o modelo proposto nesse trabalho conseguisse recuperar a estrutura do grafo que está por trás desse vetor.

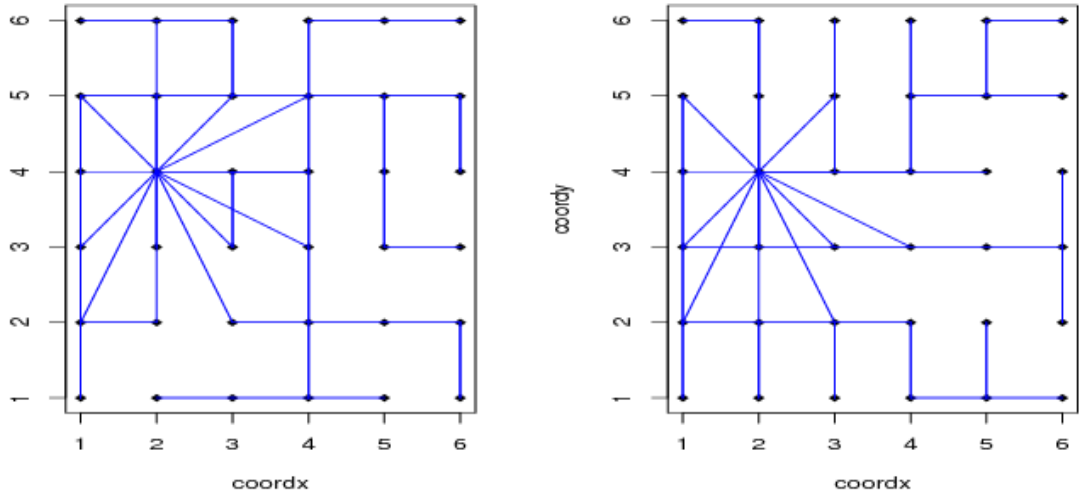


Figura 10: Simulação 2 - Grafo real e grafo estimado $\hat{\mathcal{G}}$.

3. A terceira ilustração possue um cenário com três cidades grandes, mas apenas duas delas serão consideradas polos (influenciam a vizinhança). O parâmetro $\rho = 0.9$, $|| = 2$ e a classe escolhida foi mesma fixada na simulação anterior. O resultado obtido encontra-se representado na figura 11. Frise-se que os círculos são proporcionais ao tamanho da população de cada cidade. Desse modo, podemos observar que duas cidades polos influenciam mais cidades que o restante do mapa.

   O grafo estimado é apresentado na figura 11. Como no exemplo anterior, podemos observar que os grafos são parecidos e que as informações trazidas a partir do vetor de efeitos espaciais estruturados ($\boldsymbol{\theta}$) foram suficientes para que o modelo conseguisse recuperar a estrutura do grafo que está por trás desse vetor.

4. Nesse quarto cenário, acrescentamos os valores das contagens de doenças $y_i|(\theta_i, \phi_i) \sim Poisson(0.1 * pop_i * e^{\phi_i + \theta_i})$ com $\phi_i = 1$. Os demais parâmetros e classe, foram escolhidas como no exemplo anterior. Podemos notar, através da figura 12, que o grafo estimado pelo segundo tipo de dissimilaridade recuperou, quase totalmente, a verdadeira estrutura espacial. A escolha do estimador *a posteriori* é opcional uma vez que, em geral, os dois apresentaram bons resultados.

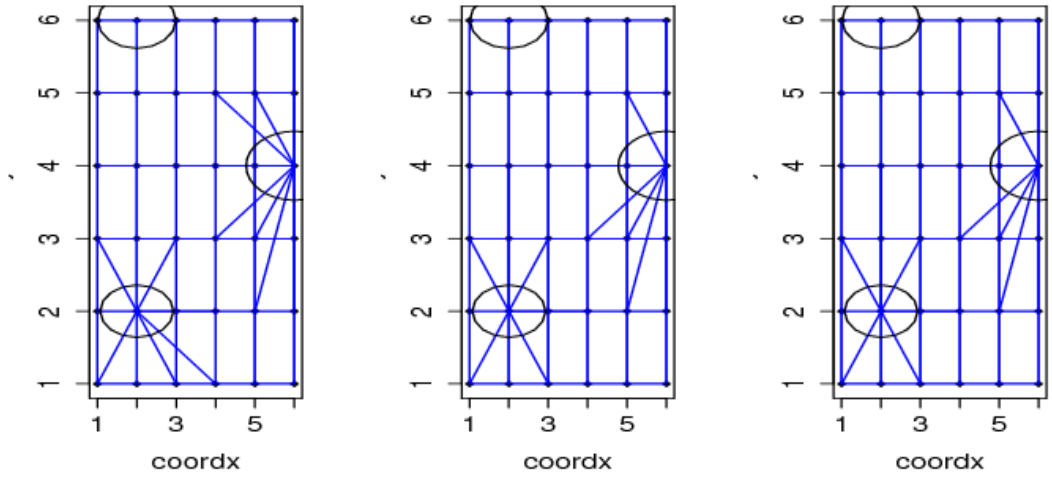


Figura 11: Simulação 3 - Grafo real, grafo estimado $\hat{\mathcal{G}}$ e grafo estimado $\hat{\mathcal{G}}_p$.

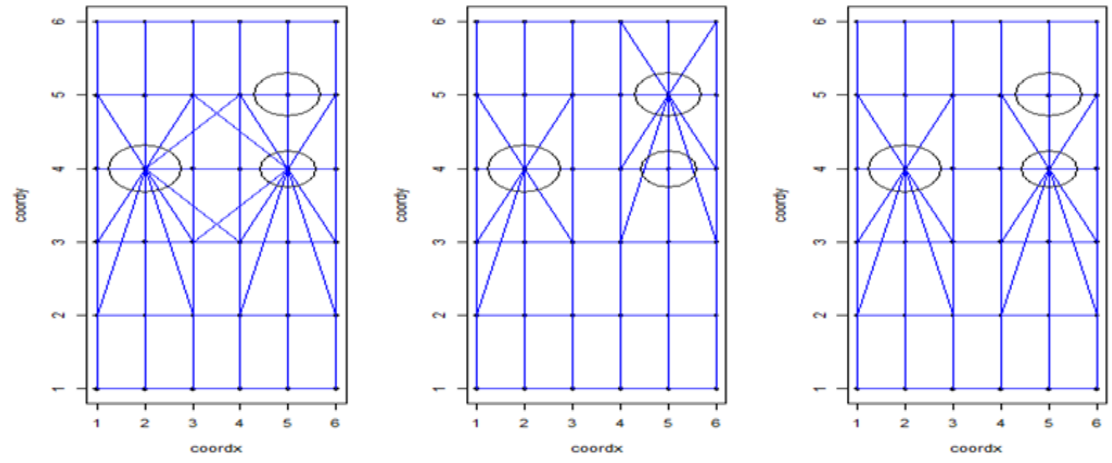


Figura 12: Simulação 4 - Grafo real, grafo estimado $\hat{\mathcal{G}}$ e grafo estimado $\hat{\mathcal{G}}_p$.

5 Exemplos - Mapas reais

Nos exemplos anteriores, utilizamos um mapa simplificado, com poucas áreas e de fácil visualização. O mapa era reticulado e todas as informações foram escolhidas ou geradas sem nenhuma fonte real de dados.

A partir desse ponto, nossos cenários apresentam diferentes características tais como tamanhos de áreas e população. Nosso novo foco será o de comparar nossos resultados com os obtidos através de modelos existentes. Em específico, usaremos o modelo BYM, no qual se supõem os efeitos aleatórios espacialmente estruturados segundo o modelo CAR e a estrutura de vizinhança como adjacência.

O número observado de casos y_i, foi simulado tomando-se como base o esquema hierárquico descrito na seção 3. Ressalta-se que para o cálculo da taxa de incidência por faixa etária, $r_j = \frac{\sum_{i=1}^{n} y_{ij}}{\sum_{i=1}^{n} pop_{ij}}$, usaremos como base os dados de bronquite e de população descritos na seção 1. O restante dos parâmetros e hiperparâmetros foram gerados segundos as seguintes distribuições de probabilidades:

$$\phi_i|\tau_\phi \sim N(0, \tau_\phi^{-1}) \quad iid\ com\ i = 1, ..., n;$$

$$\theta_i|(\theta_{-i}, \mathbf{W}, \tau_\theta, \rho) \sim N\left(\frac{\rho \sum_j w_{ij}\theta_j}{\sum_j w_{ij}}, \frac{\tau_\theta^{-1}}{\sum_j w_{ij}}\right) \quad com\ i = 1, ..., n;$$

$$\tau_\phi \sim \Gamma(0.01, 0.01);$$

$$\tau_\theta \sim \Gamma(0.01, 0.01);$$

$$\rho \sim U(0, 1).$$

O método foi implementado na linguagem R de acordo com a seção 3.3: método MCMC com 10.000 replicações com um burn in de 1000.

A qualidade das estimações foi medida a partir do Deviance Information Criterion (DIC) [44], dada por:

$$DIC = p_d + \bar{D} \tag{18}$$

Onde $\bar{D}$ representa a soma da deviance média a posteriori e p_d é o termo que penaliza a complexidade do modelo. Sob essa ótica, menores valores representam melhores modelos.

Além do DIC, foram calculado as medidas de Root Average of Mean Square Error ($RAMSE$) e Root Average of Mean Square Error Logarithm ($RAMSEL$). A medida $RAMSE$ representa a raiz quadrada do erro quadrático médio, dada por :

$$RAMSE = \sqrt{1/n \sum_{i=2}^{n} E\big((\psi_i - \hat{\psi}_i)^2|y\big)}.$$

Já a $RAMSEL$ é dada por :

$$RAMSEL = \sqrt{1/n \sum_{i=2}^{n} E\big((log(\psi_i) - \hat{log(\psi_i)})^2 | y\big)}.$$

Delineadas essas premissas básicas para a compreensão do modelo, passemos à análise de cada cenário.

A seguir é apresentado todos os detalhes de cada cenário.

1. No primeiro cenário , a estrutura espacial fixada é igual à vizinhança de adjacência. Isso significa que duas áreas somente serão consideradas vizinhas se dividirem fronteira. A distribuição a priori dos grafos pertence à seguinte classe:

$$\mathcal{W} = \begin{cases} \mathbf{W}(k) : w_{ij} = 1 & \text{se } w_{ij}^{adj} = 1 \\ \mathbf{W}(k) : w_{ij} = 1 & \text{se } i \in \mathcal{P} \text{ e } j \in \{l : d_l(i) \leq d_{(k_2)}(i)\} \\ \mathbf{W}(k) : w_{ij} = 0 & c.c \end{cases}$$

   onde w^{adj} representa a matriz de adjacência, $P(i \in \mathcal{P}) = \log(pop_i)/\sum_{i=1}^{n} \log(pop_i)$, $|\mathcal{P}| = k_1$, $k_1 \sim Binomal(n, 0.03)$ e $k_2 \sim Binomal((n-2), 0.05)$.

2. No segundo cenário, a estrutura de vizinhaça fixada é aquela em que todas as áreas são vizinhas entre si. Utilizaremos as mesmas *prioris* definidos no primeiro cenário.

3. No terceiro cenário, fixamos a mesma estrutura de vizinhança que a do cenário anterior e adicionamos ligações entre $k_1 \sim Binomal(n, 0.03)$ cidades grandes, as quais foram escolhidas com probabilidade proporcional ao logaritmo de sua população.

   Esse cenário reflete doenças que se disseminam por meio de contato físico ou das vias aéreas, razão pela qual se espalham rapidamente em locais em que há grande fluxo de pessoas. As cidades próximas (adjacentes) apresentam um tráfego intenso de pessoas e, por esse motivo, encontram-se interligadas. Desse modo, mesmo que distantes, as cidades grandes são interligadas devido ao tráfego de pessoas que existe entre elas através de malha aérea.

   A classe de distribuição a priori de $\mathbf{W}$ é dada por:

$$\mathcal{W} = \begin{cases} \mathbf{W} : w_{ij} = 1 & \text{se } w_{ij}^{adj} = 1 \\ \mathbf{W} w_{ij} = 1 & \text{se } \{i, j\} \subset \mathcal{P} \\ \mathbf{W} : w_{ij} = 0 & c.c \end{cases}$$

   onde $P(i \in \mathcal{P}) = \log(pop_i)/\sum_{i=1}^{n} \log(pop_i)$, $|\mathcal{P}| = k_1$ e $k_1 \sim Binomal(n, 0.03)$.

4. No cenário quatro, sorteamos $k_1 \sim Binomal(n, 0.03)$ cidades polos com probabilidade proporcional ao logaritmo de sua população. Além da usual adjacência, as cidades polos foram ligadas a algumas cidades próximas segundo uma distribuição $Binomial((n-2), 0.15)$. Esse cenário pode ser exemplificado em situações onde existem dois tipos de relações entre as áreas. A primeira delas é a proximidade, na qual áreas próximas tendem a ser mais parecidas entre si do que áreas distantes. Além da proximidade, também é possível detectar a influência que as cidades grandes exercem em seus arredores. Dessa forma, as cidades pequenas são influenciadas não só por suas

vizinhas geográficas, como tamém pela cidade grande mais próxima. A classe de distribuição *a priori* de $\mathbf{W}$ é a mesma que a classe do primeiro cenário.

5. No quinto cenário, $k_1 \sim Binomal(n, 0.03)$ cidades polos foram escolhidas com probabilidade proporcional ao logaritmo de sua população. Essas cidades grandes são ligadas entre si, formando um grafo completo. Além disso, cada cidade pequena é ligada á cidade polo mais próxima. Esse cenário pode ser observado em doenças que se transmitem mais facilmente em locais com grande concentração de pessoas e dependem de fatores sócio-ambientais (ex: poluição). Dessa forma, somente se contrai a doença se alguem reside ou permanece por um período de tempo em uma cidade grande. Assim, acreditamos que haverá uma ligação entre as cidades pequenas e a cidade polo mais próximo, bem como entre as cidades polos. Neste último caso, as malhas áreas e rodoviárias explicam os altos índices de incidência da doença.

   A classe de distribuição *a priori* de $\mathbf{W}$ será dada por:

$$\mathcal{W} = \begin{cases} \mathbf{W} : w_{ij} = 1 & \text{se } \{i,j\} \subset \mathcal{P} \\ \mathbf{W} : w_{ij} = 1 & \text{se } i \in \mathcal{P} \text{ e } d_i(j) = d_{(1)}(j) \\ \mathbf{W} : w_{ij} = 0 & c.c \end{cases}$$

   onde $P(i \in \mathcal{P}) = \log(pop_i)/\sum_{i=1}^{n} \log(pop_i)$, $|\mathcal{P}| = k_1$ e $k_1 \sim Binomal(n, 0.03)$.

5.1 Resultado-Mapas Reais

Chamaremos o modelo proposto de SN - Stochastic Neighborhood.

A partir da análise da tabela 1, não restam dúvidas de que nosso modelo apresentou valores de DIC e $RAMSEL$ inferiores aos obtidos pelo modelo CAR, o que demontra que as estimativas a posteriori são mais próximas dos valores fixados.

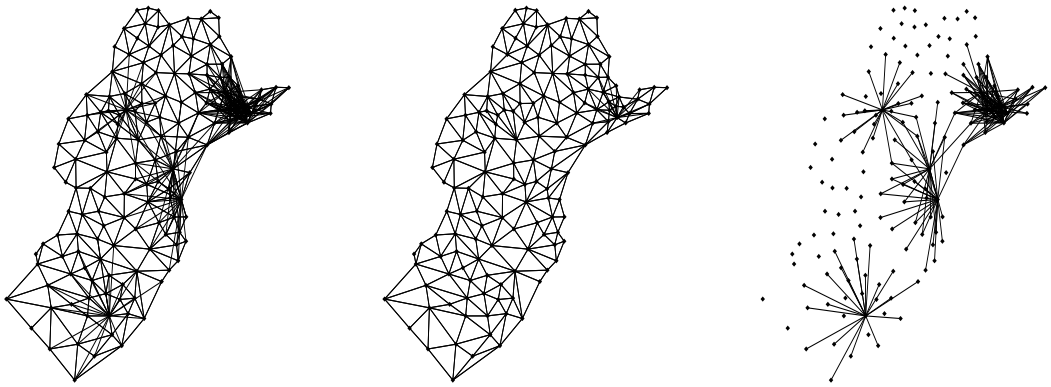
No exemplo 1, o grafo estimado foi exatamente igual ao fixado o que dispensa uma figura para sua visualização. É importante salientar que isso só foi possível devido a *piori* estabelecida para esse exemplo. Nota-se que, ao atribuir uma distribuição Binomial$(n, 0.03)$ para o números de cidades polos, permitimos que esse número seja igual a zero com a probabilidade 0.00062. Isso significa que em média, em nossa cadeia com tamanho 10 mil, esperamos encontrar seis matriz iguais a de adjacência. Como resultado do procedimento de estimação, uma dessas matrizes de adjacência foi escolhida como estimador, pois apresentou menor dissimilaridade entre todas as matrizes amostradas.

No exemplo 2, o grafo fixado foi aquele onde todas as áreas são vizinhas entre si. Devido ao maior número de arestas, as estimativas do modelo SN se aproximam mais do grafo real. Por ser um grafo completo, torna-se inviável a comparação visual do grafo real e do grafo estimado. Dessa forma, devemos analisar o desempenho do método através da 1.

As figuras 13(a),14(a) e15(a) apresentam os resultados obtidos dos exemplos 3, 4 e 5, respectivamente. Em cada uma delas, o grafo da esquerda representa o grafo real e o grafo do meio representa o grafo estimado por nosso modelo. Já o grafo da direita, representa o grafo da diferença entre os dois primeiros. Dessa forma, podemos observar em preto as arestas subestimadas: aquelas que estão faltando para que o grafo estimado seja igual ao real. As arestas superestimadas, ou seja, arestas que estão no grafo estimado mas não existem no verdadeiro grafo, são representadas pela cor cinza. A partir dos resultados, concluímos que, apesar de não possuir todas as arestas do grafo real, o modelo SN é mais próximo da realidade quando comparado com o modelo CAR.

Tabela 1: Resultados mapas reais - Comparação entre o modelo SN e CAR

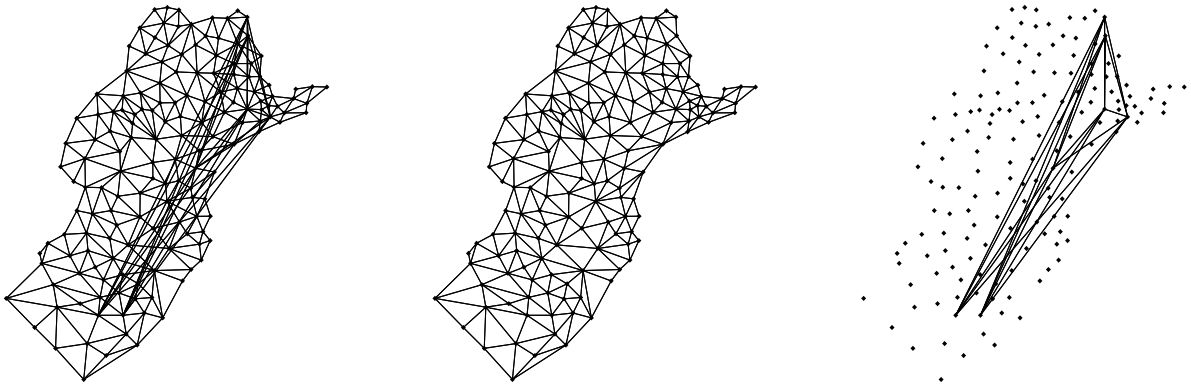
	Modelos	DIC	RAMSE	RAMSEL
Exemplo 1	SN	44626.436	0.030	0.568
	CAR	46469.601	0.030	0.579
Exemplo 2	SN	32820.729	0.0327	0.581
	CAR	35297.956	0.0328	0.597
Exemplo 3	SN	53618.842	0.062	0.573
	CAR	57233.713	0.062	0.592
Exemplo 4	SN	30190.273	0.026	0.528
	CAR	31664.623	0.026	0.538
Exemplo 5	SN	33335.892	0.013	0.765
	CAR	53375.402	0.048	1.209



(a)

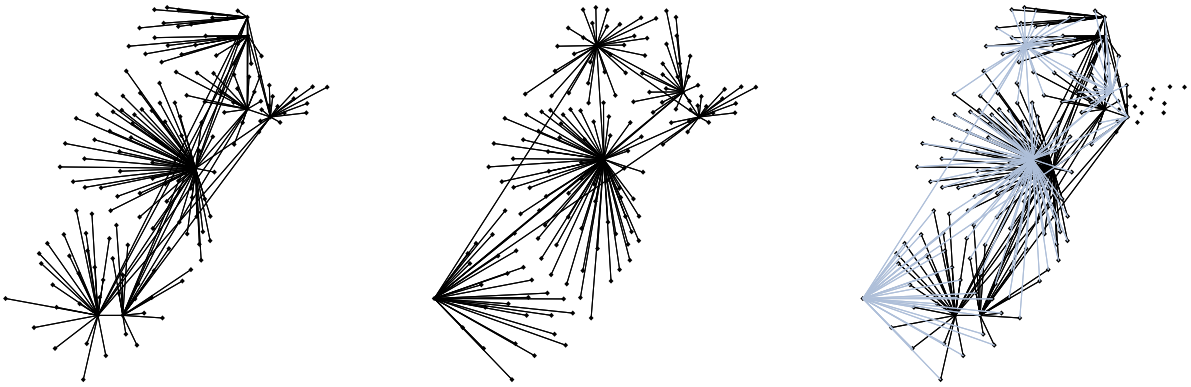
Figura 13: Grafo a esquerda representa o grafo fixado, o grafo do centro representa o grafo encontrado pelo método SN e o último grafo representa o primeiro grafo menos o segundo do exemplo 3

Além disso, observamos ainda que, os grafos obtidos a partir do nosso modelo se aproximam muito mais dos grafos fixados, o que implica em um resultado geral superior ao do modelo CAR.



(a)

Figura 14: Grafo a esquerda representa o grafo fixado, o grafo do centro representa o grafo encontrado pelo método SN e o último grafo representa o primeiro grafo menos o segundo do exemplo 4



(a)

Figura 15: Grafo a esquerda representa o grafo fixado, o Grafo do centro representa o grafo encontrado pelo método SN e o último grafo representa o primeiro grafo menos o segundo do exemplo 5. As arestas pretas do último grafo representam arestas subestimadas e as arestas cinzas representam as arestas superestimadas.

6 Simulação

Como já visto nas seções anteriores, no estudo de casos com diferentes estruturas espaciais nosso modelo apresenta melhores resultados quando comparado ao modelo CAR.

Com o intuito de melhor avaliar o desempenho do método, realizamos, nesta seção, um estudo baseado em simulações. Para tanto, adotamos o procedimento para o cálculo do número esperado de eventos e a estrutura hierárquica descritos na seção 5. Nesse estudo, usamos a matriz a priori

especificada no terceiro exemplo da seção 5 e realizamos 80 simulações do cenário selecionado. Em cada simulação, verificamos o valor do DIC, o desvio da estimativa do risco relativo e a distribuição a posteriori do ρ.

Os gráficos dos desvios das estimativas dos riscos relativos podem ser vistos na figura 16, no qual o boxplot do nosso modelo se encontra do lado esquerdo e o boxplot do modelo CAR se encontra do lado direito.

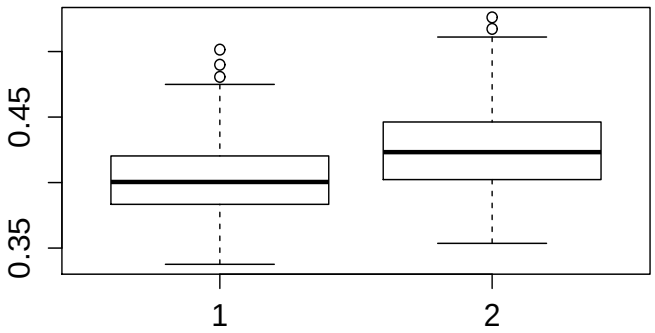


Figura 16: Boxplot do desvio das estimativas do risco relativo. O gráfico do lado esquerdo representa o modelo SN e o gráfico do lado direito representa o modelo CAR.

Analisando os resultados, fica claro que nosso modelo apresentou menor variabilidade das estimativas.

Podemos também analisar as estimativas do parâmetro ρ no ajuste dos modelos. No lado esquerdo da figura 17 encontra-se representado o boxplot das estimativas do ρ do modelo SN e, do lado direito, visualizamos o boxplot do modelo CAR. Nos dois modelo as estimativas estão concentradas no verdadeiro valor, ou seja, 0.99. No entanto, nosso modelo apresenta menor variabilidade. Em outras palavras, no modelo SN as estimativas de ρ estão mais concentradas em torno do verdadeiro valor.

Por fim, podemos observar, por meio da tabela 2, que em 76.2% das simulações o valor do DIC foi menor em nosso modelo e, em mais da metade das simulações, o mesmo fenômeno também ocorreu com o RAMSEL.

Tabela 2: Resultados da simulação comparando o modelo SN e CAR

Modelos	% menor DIC	% menor RMSE	% menor RMSEL
SN	76.2	32.5	55
CAR	23.8	67.5	45

Dessa forma, podemos concluir que nossas estimativas encontram-se mais próximas do verdadeiro valor atribuído ao risco relativo.

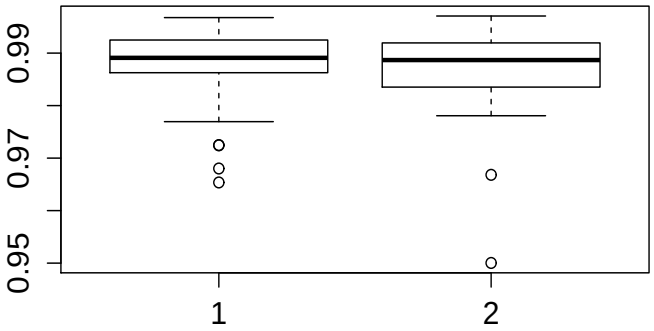


Figura 17: Boxplot das estimativas do ρ. O gráfico do lado esquerdo representa o modelo SN e o gráfico do lado direito representa o modelo CAR.

7 Aplicação

Nesta seção, apresentamos os resultados obtidos com a aplicação do modelo à duas doencas. Na primeira aplicação foi utilizado os dados de bronquite e bronquilite agudas, especialmente no que concerne à população feminina de 127 microrregiões dos estados do Paraná, do Rio Grande do Sul, de Santa Catarina e de São Paulo, no período de agosto de 2010 a agosto de 2011. Já na segunda aplicação foi utilizado os dados de óbitos de Meningite infecciosa no mesmo período de tempo.

Optamos por utilizar duas matrizes de vizinhança a priori. A primeira foi utilizada na primeria aplicação e é definida por:

$$\mathcal{W} = \begin{cases} \mathbf{W}(k) : w_{ij} = d_i(j)^{\beta} & \text{se} \quad i \in \mathcal{P} \quad \text{e} \quad d_i(j) \leq 200km \\ \mathbf{W}(k) : w_{ij} = 0 & c.c \end{cases}$$

onde $P(i \in \mathcal{P}) = \log(pop_i)/\sum_{i=1}^{n} \log(pop_i)$, $|\mathcal{P}| = k_1$, $\beta = \left(\frac{\log 0.5}{\log 50}\right)$ e $k_1 \sim Binomal(n, 0.03)$.

O fato de β ser negativo propociona um decaimento do valor de w_{ij} quando a distância entre as áreas aumenta. O valor de β foi escolhido de tal forma que $w_{ij} = 0.5$, quando duas cidades distam 50km. Delimitamos esse valor em 200 km por acreditar que esse limiar é razoável quando considerada a influência ambiental de áreas. Uma vez que nossas coordenadas geográficas são dadas por graus de latitude e longitude, usamos a conversão de Haversine para encontrar a distância entre duas áreas.

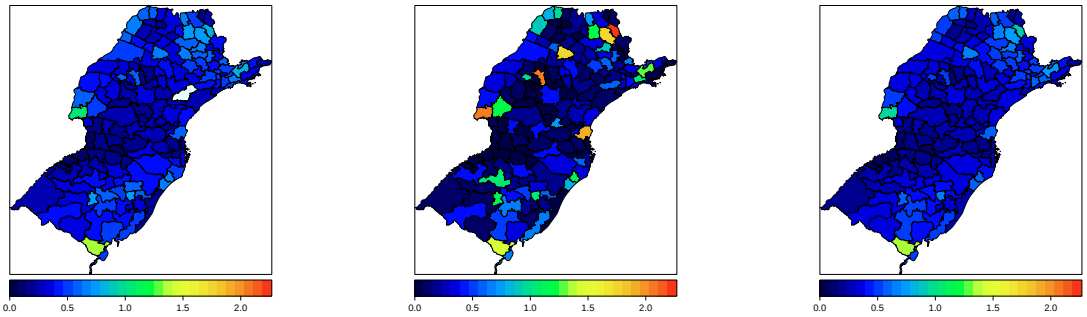
A segunda matrizes de vizinhança *a priori* foi aplicada no segundo conjunto de dados e foi definida como no terceiro exemplo da seção 5.

A partir da tabela 3, podemos observar que nosso modelo apresenta um valor de DIC bem menor do que aquele encontrado pelo modelo CAR nas duas aplicações.

Tabela 3: Resultados da aplicação dos modelos SN e CAR

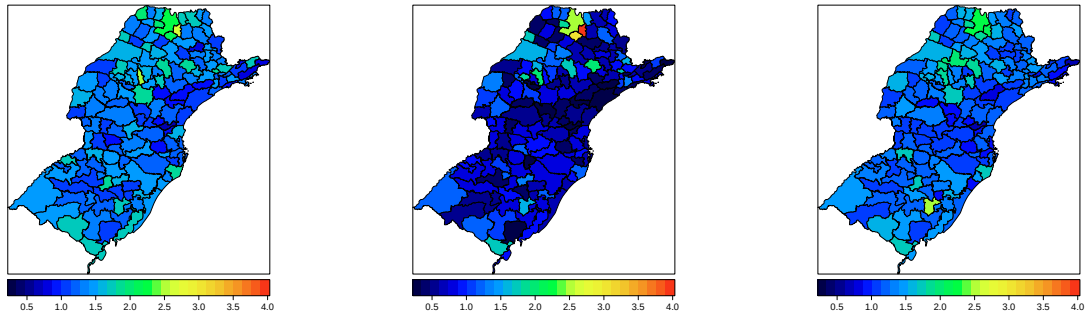
Modelos	DIC - Aplicação1	DIC - Aplicação2
SN	33605.29	49859.76
CAR	40130.87	60353.66

As figuras 18 e 19 apresentam os mapas de estimativas dos riscos relativos encontrados pelo nosso modelo, pela estimativa SMR e pelo modelo CAR, respectivamente, para os dados de bronquite e meningite, Podemos concluir que nosso modelo suaviza as estimativas dos riscos de forma mais nítida do que o modelo CAR, uma vez que acrescenta mais arestas entre as áreas.



(a) Risco Estimado CAR (b) Risco SMR (c) Risco Estimado SN

Figura 18: Mapa de risco dos dados de Bronquite



(a) Risco Estimado CAR (b) Risco SMR (c) Risco Estimado SN

Figura 19: Mapa de risco dos dados Meningite

Os gráficos dos desvios das estimativas dos riscos relativos e das estimativas de ρ podem ser vistos na figuras 20 e 21, respectivamente. Em ambas figuras, o boxplot do nosso modelo se encontra do lado esquerdo e o boxplot do modelo CAR se encontra do lado direito. Assim como nos resultados obtidos nas simulações, observa-se que nosso modelo apresenta menor variabilidade das estimativas.

Os grafos estimados *a posteriori* podem ser observados no lado esquerdo das figuras 22 e 23. Com o intuíto de facilitar a diferenciação do grafo estimado no modelo SN e do grafo de adjacência usado no modelo CAR, plotamos no lado direito das figuras as arestas que foram acrescentadas de um grafo para outro.

Na aplicação de bronquite, podemos notar que foram acrescentadas mais arestas perto da região de São Paulo e Campinas, quando comparadas a matriz de adjacência. Além disso, o grau de confiança médio dessas novas arestas é de 0.92 e variância igual a 0.07. Esses valores foram calculados a partir de suas probabilidades a posteriori, ou seja, o número de vezes em que foram observadas essas arestas em cada amostra gerada. Isso significa que, no âmbito de saúde pública, é bastante razoável analisar essa região com mais cuidado, uma vez que essas cidades se influenciam. Já na aplicação de meningite, podemos observar a existência de quatro cidades polos: Curitiba, Porto Alegre, São Paulo e Campinas.

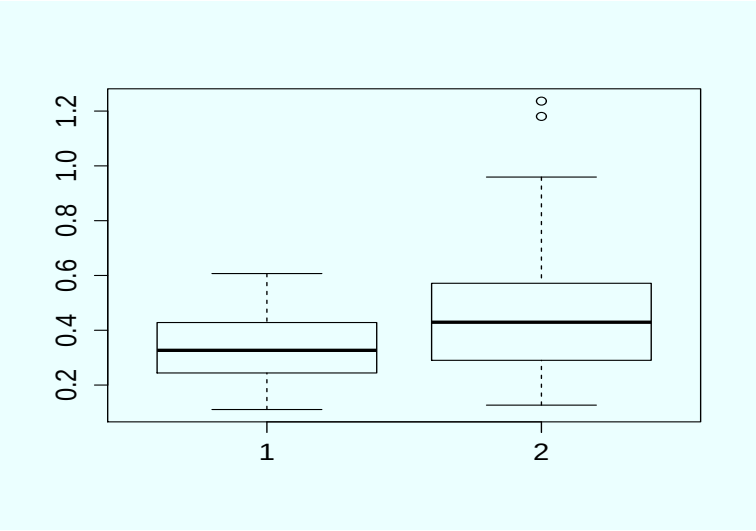


Figura 20: Boxplot dos desvios das estimativas do ρ para os dados de Bronquite. O gráfico do lado esquerdo representa o modelo SN e o gráfico do lado direito representa o modelo CAR

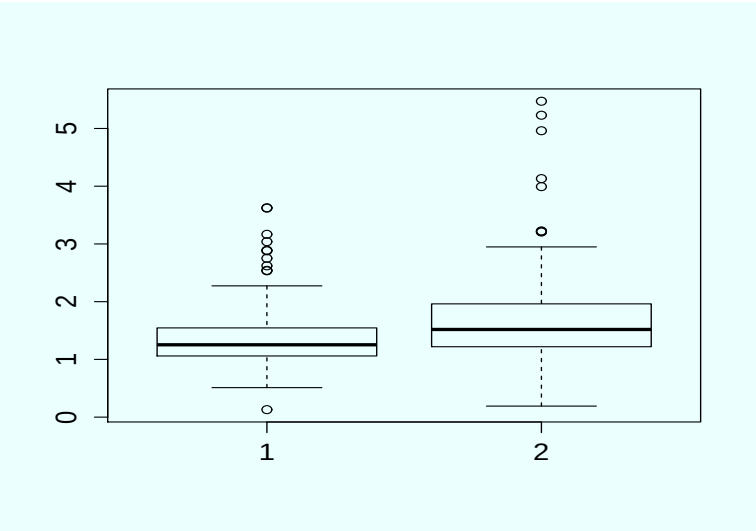


Figura 21: Boxplot das estimativas do ρ para os dados de Meningite. O gráfico do lado esquerdo representa o modelo SN e o gráfico do lado direito representa o modelo CAR

O grau de confiança médio das arestas acrescentadas (além das de adjacência) é de 0.43 com variância igual a 0.42, apresentando uma menor grau de confiançao quando comparado ao da primeira aplicação.

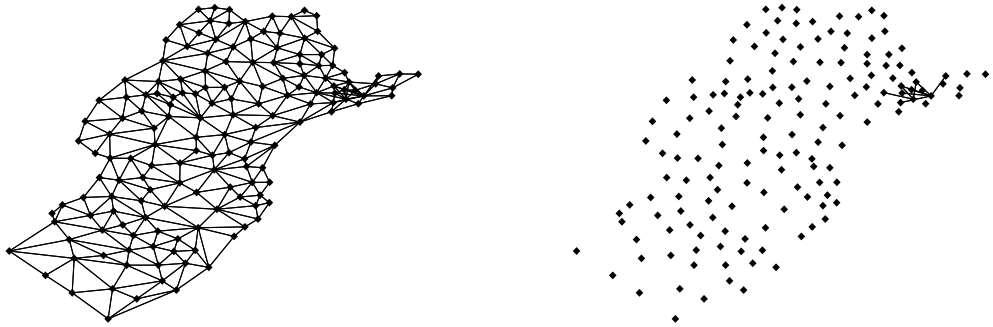


Figura 22: Grafo à esquerda representa o grafo estimado e o grafo à direta representa o grafo estimado menos o grafo de adjacência nos dados de bronquite.

Figura 23: Grafo à esquerda representa o grafo estimado e o grafo à direta representa o grafo estimado menos o grafo de adjacência nos dados de meningite.

8 Conclusão

Os estudos sobre mapeamento de doenças em dados de áreas continuam gerando grande interesse científico. As contagens são geralmente assumidas como variáveis de Poisson, nas quais o parâmetro é formado pelo número esperado de observações e pelo risco relativo, objetivando obter estimativas realistas do agravo em análise.

Neste trabalho, dispensamos especial atenção ao método CAR, por ser amplamente utilizado no mapeamento de doenças. Este método hierárquico bayesiano considera que o risco relativo de uma dada doença pode ser explicado por dois efeitos aleatórios. O primeiro efeito é considerado não estruturado segundo uma distribuição normal multivariada independente. De outro modo, o segundo efeito é um Campo de Markov com distribuição normal. Ressalte-se que este efeito traz consigo, ainda, a ideia de dependência espacial entre áreas (a distribuição de uma área, dada toda a região, depende apenas de suas áreas vizinhas), devido ao fato de que as áreas podem ser influenciadas por outras que

se encontram em seu derredor. Em outras palavras, se uma área possui grande risco relativo, as áreas sob sua influência também apresentarão riscos elevados. Assim, a correlação espacial é capturada através da matriz de vizinhança utilizada nas distribuições normais dos efeitos aleatórios.

No método CAR a matriz de vizinhança é fixada antes da análise e geralmente é baseada na adjacência por conveniência. Em assim sendo, nosso trabalho atenta para duas grandes questões.

Em primeiro lugar, a vizinhança em adjacência sempre é uma boa opção? Existem diversos cenários nos quais a estrutura de adjacência não é a melhor escolha. Nesses casos, pode-se dizer que localização espacial não é suficiente para estimar os riscos relativos e a relação entre as áreas, o que ficou demonstrado por meio da correlação existente entre o tamanho das cidades e a incidência de doenças respiratórias ou de crimes. Daí porque não se apresenta como melhor solução a fixação da matriz de vizinhança com base na simples adjacência.

Partindo dessa conclusão, existe a difícil tarefa de se estabelecer uma matriz de vizinhança "razoável". Qual seria a melhor escolha? Conseguiríamos identificar essa única matriz? Objetivando responder a estes questionamentos, fomos levados a criar classes de matrizes razoáveis para solucionar o problema, o que possibilitou a utilização de métodos (e.g., o MCMC) para fazer inferências sobre as matrizes. Para encontrar estimativas *a posteriori*, propusemos dois estimadores para a amostra de matrizes de vizinhança obtidas durante o procedimento de amostragem das cadeias.

Partindo dessa proposta, passamos a submeter nosso modelo a diversas situações, com o intuito de testar o seu desempenho.

A primeira parte dos exemplos nos leva a concluir que os resultados obtidos na recuperação da estrutura espacial por trás dos dados na forma de um lattice foram satisfatórios.

Na segunda parte dos exemplos, simulamos as contagens da uma dada doença a partir do número esperado de casos em cada área. Para gerar as contagens, utilizamos dados da bronquite e bronquilite agudas na população feminina, dividida em faixas etárias, no ano de 2012. Os resultados encontrados atestaram uma melhor adequação do nosso modelo quando comparado ao CAR, o que confirmado, ainda, por meio das simulações e da sua efetiva aplicação.

Em suma, podemos concluir que nossa proposta é mais adequada no mapeamento de doenças, uma vez que torna desnecessária a escolha de uma única matriz de vizinhança e os resultados são mais precisos do que os apresentados pelos modelos existentes.

Referências

[1] Bithell, J. A classification of disease mapping methods. *Statistics in Medicine* 2000; **19**(17-18):2203–2215. URL http://onlinelibrary.wiley.com/doi/10.1002/1097-0258(20000915/30)19:17/18%3C2203::AID-SIM5

[2] Diggle, PJ. Overview of statistical methods for disease mapping and its relationship to cluster detection. *Spatial Epidemiology: Methods and Applications* 2000; :87–103.

[3] Lawson, AB. Disease map reconstruction. *Statistics in Medicine* 2001; **20**(14):2183–2204. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.933/full.

[4] MacNab, YC, Kmetic, A, Gustafson, P, Sheps, S. An innovative application of bayesian disease mapping methods to patient safety research: A canadian adverse medical event study. *Statistics in medicine* 2006; **25**(23):3960–3980. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.2507/abstract.

[5] Yu, HL, Chiang, CT, Lin, SD, Chang, TK. Spatiotemporal analysis and mapping of oral cancer risk in changhua county (taiwan): an application of generalized bayesian maximum entropy method. *Annals of epidemiology* 2010; **20**(2):99–107. URL http://www.sciencedirect.com/science/article/pii/S1047279709003421.

[6] MacNab, YC, Gustafson, P. Regression b-spline smoothing in bayesian disease mapping: with an application to patient safety surveillance. *Statistics in medicine* 2007; **26**(24):4455–4474. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.2868/abstract.

[7] Cressie, N, Read, TR. Spatial data analysis of regional counts. *Biometrical Journal* 1989; **31**(6):699–719. URL http://onlinelibrary.wiley.com/doi/10.1002/bimj.4710310607/abstract.

[8] Clayton, D, Kaldor, J. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987; **43**:671–681. URL http://www.jstor.org/stable/10.2307/2532003.

[9] Marshall, RJ. Mapping disease and mortality rates using empirical bayes estimators. *Applied Statistics* 1991; :283–294URL http://www.jstor.org/stable/10.2307/2347593.

[10] Tsutakawa, RK, Shoop, GL, Marienfeld, CJ. Empirical bayes estimation of cancer mortality rates. *Statistics in Medicine* 1985; **4**(2):201–212. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.4780040210/abstract.

[11] Mollie, A, Richardson, S. Empirical bayes estimates of cancer mortality rates using spatial models. *Statistics in Medicine* 1991; **10**(1):95–112. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.4780100114/abstract.

[12] Besag, J, York, J, Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 1991; **43**(1):1–20. URL http://link.springer.com/article/10.1007/BF00116466.

[13] Carlin, BP, Banerjee, S. Hierarchical multivariate car models for spatio-temporally correlated survival data. *Bayesian statistics* 2003; **7**:45–63.

[14] Jin, X, Carlin, BP. Multivariate parametric spatiotemporal models for county level breast cancer survival data. *Lifetime Data Analysis* 2005; **11**(1):5–27. URL http://link.springer.com/article/10.1007/s10985-004-5637-1.

[15] Gelfand, AE, Vounatsou, P. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* 2003; **4**(1):11–15. URL http://biostatistics.oxfordjournals.org/content/4/1/11.short.

[16] Held, L, Natário, I, Fenton, SE, Rue, H, Becker, N. Towards joint disease mapping. *Statistical methods in medical research* 2005; **14**(1):61–82. URL http://smm.sagepub.com/content/14/1/61.short.

[17] Held, L, Graziano, G, Frank, C, Rue, H. Joint spatial analysis of gastrointestinal infectious diseases. *Statistical methods in medical research* 2006; **15**(5):465–480. URL http://smm.sagepub.com/content/15/5/465.short.

[18] Martínez-Beneito, M, López-Quilez, A, Botella-Rocamora, P. An autoregressive approach to spatio-temporal disease mapping. *Statistics in medicine* 2008; **27**(15):2874–2889. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.3103/abstract.

[19] Knorr-Held, L, Best, NG. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2001; **164**(1):73–85. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-985X.00187/abstract.

[20] Sun, D, Tsutakawa, RK, Kim, H, He, Z, et al. Spatio-temporal interaction with disease mapping. *Statistics in Medicine* 2000; **19**(15):2015–2035.

[21] Silva, GL, Dean, C, Niyonsenga, T, Vanasse, A. Hierarchical bayesian spatiotemporal analysis of revascularization odds using smoothing splines. *Statistics in medicine* 2008; **27**(13):2381–2401. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.3094/abstract.

[22] Kafadar, K, Freedman, LS, Goodall, CR, Tukey, JW. Urbanicity-related trends in lung cancer mortality in us counties: white females and white males, 1970–1987. *International journal of epidemiology* 1996; **25**(5):918–932. URL http://ije.oxfordjournals.org/content/25/5/918.short.

[23] Goodall, CR, Kafadar, K, Tukey, JW. Competing and using moral versus urban measures in statistical applications. *The American Statistician* 1998; **52**(2):101–111. URL http://amstat.tandfonline.com/doi/abs/10.1080/00031305.1998.10480548.

[24] Kafadar, K. Geographic trends in prostate cancer mortality: an application of spatial smoothers and the need for adjustment. *Annals of epidemiology* 1997; **7**(1):35–45. URL http://www.sciencedirect.com/science/article/pii/S1047279796001019.

[25] Jackson, MC, Huang, L, Xie, Q, Tiwari, RC. A modified version of moran's i. *International Journal of Health Geographics* 2010; **9**:33.

[26] Sunyer, J, Jarvis, D, Gotschi, T, Garcia-Esteban, R, Jacquemin, B, Aguilera, I, Ackerman, U, de Marco, R, Forsberg, B, Gislason, T, et al. Chronic bronchitis and urban air pollution in an international study. *Occupational and environmental medicine* 2006; **63**(12):836–843. URL http://oem.bmj.com/content/63/12/836.short.

[27] Holland, W, Reid, D. The urban factor in chronic bronchitis. *The Lancet* 1965; **285**(7383):445–448.

[28] Lindgren, A, Stroh, E, Montnémery, P, Nihlén, U, Jakobsson, K, Axmon, A. Traffic-related air pollution associated with prevalence of asthma and copd/chronic bronchitis. a cross-sectional study in southern sweden. *International journal of health geographics* 2009; **8**(1):2.

[29] Friedman, J, Hastie, T, Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 2008; **9**(3):432–441. URL http://biostatistics.oxfordjournals.org/content/9/3/432.short.

[30] Teyssier, M, Koller, D. Ordering-based search: A simple and effective algorithm for learning bayesian networks. *arXiv preprint arXiv:1207.1429* 2012; URL http://arxiv.org/abs/1207.1429.

[31] Tsamardinos, I, Brown, LE, Aliferis, CF. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning* 2006; **65**(1):31–78. URL http://link.springer.com/article/10.1007/s10994-006-6889-7.

[32] Chickering, DM. Optimal structure identification with greedy search. *The Journal of Machine Learning Research* 2003; **3**:507–554. URL http://dl.acm.org/citation.cfm?id=944933.

[33] Heckerman, D, Geiger, D, Chickering, DM. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning* 1995; **20**(3):197–243. URL http://link.springer.com/article/10.1007/BF00994016.

[34] Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausble Inference*. Morgan Kaufmann Pub, 1988.

[35] Bornn, F, Luke; Caron. Bayesian clustering in decomposable graphs. *Bayesian Analysis* 2010; **6**:829–846.

[36] White, G, Ghosh, SK. A stochastic neighborhood conditional autoregressive model for spatial data. *Computational statistics & data analysis* 2009; **53**(8):3033–3046. URL http://www.sciencedirect.com/science/article/pii/S0167947308003885.

[37] Mollié, A. Bayesian mapping of disease. *Markov chain Monte Carlo in practice* 1996; **1**:359–379.

[38] Besag, J, Newell, J. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 1991; **154**:143–155. URL http://www.jstor.org/stable/10.2307/2982708.

[39] Brook, D. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika* 1964; **51**(3/4):481–483. URL http://www.jstor.org/stable/10.2307/2334154.

[40] Kirchhoff, G. Über die auflösung der gleichungen, auf welche man bei der untersuchung der linearen verteilung galvanischer ströme geführt wird. *Ann. Phys. Chem* 1847; **72**:497–508.

[41] Khan, K, Arino, J, Hu, W, Raposo, P, Sears, J, Calderon, F, Heidebrecht, C, Macdonald, M, Liauw, J, Chan, A, et al. Spread of a novel influenza a (h1n1) virus via global airline transportation. *New England Journal of Medicine* 2009; **361**(2):212–214. URL http://www.nejm.org/doi/full/10.1056/NEJMc0904559.

[42] Warren, A, Bell, M, Budd, L. Airports, localities and disease: representations of global travel during the h1n1 pandemic. *Health & place* 2010; **16**(4):727–735. URL http://www.sciencedirect.com/science/article/pii/S1353829210000328.

[43] Hsu, CI, Shih, HH. Transmission and control of an emerging influenza pandemic in a small-world airline network. *Accident Analysis & Prevention* 2010; **42**(1):93–100. URL http://www.sciencedirect.com/science/article/pii/S0001457509001717.

[44] Speigelhalter, D, Best, N, Carlin, B, Van der Linde, A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* 2003; **64**(4):583–616.

Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Space-time prospective surveillance based on Knox local statistics

Aline Piroutek[a], Renato Assunção[*] and Thaís Paiva[a]

We propose a surveillance system to prospectively monitor the emergence of space-time clusters in point pattern of disease events. Its aim is to detect a cluster as soon as possible after its emergence and it is also desired to keep the rate of false alarms at a controlled level. It is an easily understood and easily implemented system, requiring very little input from the user. This makes it a promising candidate to practical use by public health official agencies. Our method is a modification from a previous proposal made by Rogerson [1], who examined a retrospective surveillance scenario, looking for the earliest time in the past that change could have been deemed to occur. We modify his method to take into account the prospective case. We evaluated our surveillance system in several scenarios, including without and with emerging clusters, checking distributional assumptions and assessing performance impacts of different emergence times, shapes, extent and intensity of the emerging clusters. Our conclusion is that our space-time surveillance system based on local Knox statistics is very efficient in its statistical properties and it is appealing to epidemiologists and public health officials due to its simplicity of use and easy understanding.

Keywords: Spatial statistics; Disease mapping; Disease surveillance; Prospective space-time surveillance; Prospective surveillance, Space-time clustering.

1. Introduction

The increasing availability of on-time recording of disease events with high resolution spatial and temporal coordinates has sparked the interest on prospective surveillance systems to detect as soon as possible the emergence of localized space-time disease clusters. All methods must deal with the trade-off of either detecting emerging clusters very early with a high rate of false positives or reduce the false positive rates at the cost of delaying for too long the detection of true clusters emerging. Most statistical methods routinely in use for the early detection of disease outbreaks are purely temporal in nature but that are now some methods using also the spatial information (see the excellent reviews in [2] ; [3]; [4]).

The space-time surveillance methods can be divided into two broad categories according to their use of events locations recorded either at point level data or aggregated into disjoint areas that partition the region of interest. One early work using area based data is [5], who described analytical methods and necessary procedures to implement these methods in epidemiology and public health. Another early attempt is [6], using a dynamic Ising model parameter to monitor patterns over time in areas. A more promising approach was proposed by Kulldorff [7] who used a space-time scan statistic to continuously monitor disease cases in areas. A modification of this scan statistic was suggested in [8] by introducing time series methods to model the baseline rates of events for each area. An adaptation of scan statistics methods to deal with point-based events rather than area-based was suggested in [9] and [10]. The use of scan statistics based on statistical hypothesis testing ideas in this prospective surveillance context has been criticized by Woodall *et al* [11]. As they convincingly argue, scan statistics resort to statistical hypothesis concepts such as error type I probability and power, which are appropriate for a retrospective cluster detection problem but not for the prospective surveillance perspective.

It was proposed in [12] a method for space-time point process data that does not require the functional specification of the purely spatial or the purely temporal functions. They adapted the Shiryayev-Roberts control chart method introduced

[a] *Departamento de Estatística, Universidade Federal de Minas Gerais, 31270-901. Belo Horizonte, MG, Brazil*
[*] *Correspondence to: Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 31270-010. Belo Horizonte, MG, Brazil. E-mail: assuncao@dcc.ufmg.br*

Contract/grant sponsor: The second author acknowledges support received from FAPEMIG, CNPq, and CAPES, Brazilian research supporting agencies.

in [13] for the space-time situation and used the Optional Stopping Theorem to derive the values for the tuning parameters of their method. This method is implemented in the `surveillance` R package and it is very simple to use, requiring little statistical expertise and fine-tuning from the user. In this R package, there are also space-time surveillance methods for point process data and areal data proposed by [14] and [15]. In contrast with the method proposed by [12], these latter methods require rather extensive modeling expertise from the user to take into account the marginal spatial and temporal patterns. This restricts their use to large public health agencies where these resources are more readily available.

This same requirement of extensive preliminary statistical analysis is one aspect of the Bayesian approach for the prospective detection of space-time clusters proposed by [16], [17], and by [18]. These methods require great deal of sophisticated statistical modeling, expertise in Bayesian posterior distribution simulation, and are subject to constant review of the marginal spatial and temporal patterns to track overall changes in these marginal distributions. They are hence of little use for agencies that intend to use methods in a semi-automatic way.

Extensive research on multivariate surveillance methods have been carried out by the group at the University of Gothenburg (see [19] and [20]). They considered the simultaneous space-time surveillance of more than one disease and used methods based on quality control ideas, such as the Shewhart, EWMA, and CUSUM methods. These are promising methods for widespread use in local areas as they try to make automatic most of the modeling decisions the user must take.

Rogerson in [21] suggested the use of CUSUM methods to monitor the spatial dynamics of disease rates. Later, he improved his proposal by allowing for individually point-referenced events by means of Knox local statistics [1]. It is important to note that Rogerson examined a retrospective surveillance scenario, asking in effect when the earliest time that change could have been deemed to occur, based on known locations to the point in time where the question was being asked. His approach did not include the prospective case, the main focus of this paper. Nevertheless, his space-time surveillance method has several advantages. It is based on the Knox statistics, a very intuitive statistic. It does not require extensive previous knowledge from the user to use the method. There is no need to model the marginal spatial and temporal patterns in advance as the method automatically takes them into account. Its tuning parameters are kept to a minimum and it is not difficult to set reasonable values for them as they are readily interpretable. The method is entirely based on quality control type of ideas in which the repeated testing implied by the prospective surveillance is naturally accommodated. More over, the method is easily implemented in computer code as it does not require simulation or hard numerical calculation.

Using Rogerson approach in the prospective surveillance situation, [22] found serious problems with this local Knox statistics method. Their paper evaluated extensively its performance by simulation under several scenarios, examining the average run length (ARL), the effect of population density, region shape, among other issues. They showed that the nominal fundamental parameters, such as the ARL, are misleading and that the performance of the method is highly influenced by several factors such as the population density and region shape. In general terms, they did not recommend the use of this method for prospective purposes in any practical application.

In this paper, we redefined the local Knox statistic in Rogerson proposal to make it a proper space-time prospective surveillance method. The definition proposed by Rogerson was based on a retrospective use of the Knox statistics rather than considering its longitudinal use, when future events are not available yet. The change we introduce implies a different probability distribution for the local Knox statistics and, as a consequence, changes in the surveillance system behavior. Our new definition retains the good properties of the Rogerson original proposal, namely, its intuitive interpretation, easy implementation, and adherence to quality control rather than statistical hypothesis testing principles. At the same time, we were able to eliminate the problems identified by [22]. The resulting surveillance method has very good statistical performance and the nominal tuning parameters can be trusted. Our conclusion is that this new method should be seen as an important candidate for the space-time surveillance of point-referenced events.

In the next section, we briefly review the local Knox statistics and summarize the problems found by [22] in its performance evaluation. We also discuss how it should be modified to be useful in a prospective surveillance method. We obtain the distribution of this statistic when there is no emerging cluster and discuss how to establish tuning thresholds. In section 3, we describe the extensive simulation scenarios used to evaluate the method performance. This includes scenarios without and with emerging clusters, scenarios to check distributional assumptions, and to assess the impacts of different emergence times, shapes, extent, and intensity of the emerging clusters. Results are presented in section 4. We close the paper with final remarks and conclusions in section 5.

2. Review of The Local Knox Monitoring Method

Suppose that n disease events are represented by (x_i, y_i, t_i) where the x_i and y_i are the spatial coordinates and t_i is the occurrence time of the i-th event, with $t_1 < t_2 < \ldots < t_n$. With the aim of quantifying the evidence for an infectious etiology for the disease, Knox proposed a test in [23] to assess if pairs of events geographically close tend to be also close in time. It requires two threshold tuning parameters. To distinguish them from another threshold required by the surveillance method, we will refer to these tuning parameters as bound parameters. The first bound parameter is a critical distance ρ such that two events are considered geographical neighbors, or spatially close to each other, if their distance is less than ρ. Analogously, two events are considered temporal neighbors if their time difference is less than a temporal bound τ selected by the user. The Knox statistic is the total number X of pairs of events that are simultaneously close in space and time. Although there are analytical approximations for the Knox test statistic distribution under space and time independence, nowadays the best practice is to run a randomization test conditioned on the observed spatial and temporal locations. That is, holding the

spatial locations fixed, permute the times $t_1, \ldots, t_n$ among the events to generate pseudo point patterns $(x_i, y_i, t_{\pi(i)})$, with $(\pi(1), \ldots, \pi(n))$ being a random permutation of the vector $(1, 2, \ldots, n)$. The Knox test statistic is calculated and saved for each random permutation generated. After a large number B of random permutations, an exact p-value is calculated based on the empirical distribution function of the test statistic using the simulated values (see [24] for details).

The Knox test is a retrospective method that globally verify the presence of space time clustering throughout the data, without identifying specific localized clusters. This is appropriate when the test is aimed at finding evidence of disease contagion or infection. However, when the interest is in spatially localized episodic or epidemic outbreak, the identification of clusters is important as the space-time interaction appears in the form of raised incidence on localized regions over a short time period [25]. We focus on this problem of identifying as early as possible the emergence of a space-temporal disease cluster when point events are under monitoring. One of the reasons for increased interest in this problem is the need for the earliest possible detection of bioterrorism attacks. Another reason is the utility of enhanced detection of localized surging of both endemic and new diseases incidence rates. Another possibility for the emergence of a localized cluster is the surging of an environmental hazard (e.g. radiation leak).

For the detection of emergent localized clusters, the second situation is clearly the most appropriate. This prompted Rogerson to define in [1] a local Knox statistic that evaluated the amount of excess events around each given event. If few highly clustered events are responsible for the rejection of the independence between time and space, this local Knox statistics would be able to identify them. He decomposed the global Knox statistic X by writing $X = \sum_{i=1}^{n} n(i)/2$, where $n(i)$ is the number of events that are within distance ρ and did occur within τ time units away from the i-th event. The number 2 in the decomposition is due to the double counting of each pair of events. Conditioning on the observed spatial and temporal coordinates and assuming that the assignment of the times to the spatial locations are equally likely, the distribution of X is a weighted sum of hypergeometric distributions. In [1], its first two moments were derived. The second moment was corrected in [26] and [22].

The surveillance method proposed in [1] proceeds by calculating a standardized local Knox score,

$$z_i = \frac{n(i) - E\{N(i)\} - 0.5}{\sqrt{\mathrm{Var}\{N(i)\}}} \tag{1}$$

where $N(i)$ is the random variable associated with the observed value $n(i)$. These scores are cumulatively summed to generate

$$S_i = \max\{0, S_{i-1} + z_i - k\} \tag{2}$$

with $S_0 = 0$ and k usually set equal to $1/2$. The value of S_i is monitored and an alarm sounds off when it exceeds a critical value h . The choice of h is determined by the desired average run length (ARL_0), defined as the mean number of events until the alarm sounds off when there is no emerging clusters, or the process is in-control state. The higher the parameter h, the longer the average run lengths, implying in few false alarms. As a trade-off, there will be a long delay to detect a truly emerging cluster, when the process is in the out-of-control state. The relationship between ARL_0 and h is given by Rogerson [1] when the successive z_i's are i.i.d standard normally distributed and $k = 1/2$:

$$ARL_0 \approx 2\{\exp(h + 1.166) - h - 2.166\} . \tag{3}$$

An indirect way to establish the ARL_0 (and therefore h) has been suggested in [1]. This requires an additional stochastic assumption. Let $RL = \min_i\{S_i > h\}$ be the random run length until the alarm system sounds off and assume that RL follows approximately a discrete geometric probability distribution with expected value ARL_0. We can then fix a desired probability α of a false alarm during a study period of d successive observations and relate this to the ARL_0:

$$\alpha = \mathrm{P}(RL > d) = 1 - \exp(-d/ARL_0) . \tag{4}$$

Hence, either the user starts directly with the ARL_0 value and finds h by the approximation (3), or he establishes a false alarm rate α in d consecutive events and finds the implied ARL_0. The expression (3) was inverted in [27] to give approximately the value of h as a function of ARL_0:

$$h \approx \left(\frac{ARL_0 + 4}{ARL_0 + 2}\right) \log\left(\frac{ARL_0}{2} + 1\right) - 1.166 .$$

In conclusion, the method is extremely simple to implement, intuitively appealing, and require little previous knowledge from the user apart from the critical bounds. Using the method in a prospective perspective, Marshall *et al.* [22] carried out an extensive simulation study to evaluate its performance and they found disappointing results. More specifically, they examined the in-control ARL_0 performance for varying values of the space and time bounds and the CUSUM control limit h. They also studied the effect of population density and region shape on ARL_0 performance. They showed that the real ARL_0, in contrast to the nominal value established by the user, is highly influenced by the bound values, population density, and region shape. They concluded that this method cannot be recommended in any practical application for prospective surveillance.

In this paper, we show that this bad performance disappears if we redefine the local Knox score with the prospective use in mind. The statistic $n(i)$ is the number of events within within a cylinder parallel to the time axis, centered at the i-th

event and with radius ρ and height 2τ. That is, $n(i)$ counts the number of other nearby events occurring immediately before *and after* the i-th event. This is is not appropriate if we are dealing with a prospective surveillance situation. When the newest i-th event comes in, only the previous events will be available and any evidence of local excessive incidence should be based only on the previously observed events, not on the yet to be recorded events.

Consider the i-th event as the most recently observed so there are no other events after its occurrence. Denote by $N_{\rho\tau}(i)$ the random variable counting the events within a geographical distance ρ and occurring previously by at most τ time units from the i-th event (x_i, y_i, t_i). Let $n_\rho(i)$ be the number of events spatially close to the i-th event, considering only those events j with $t_j < t_i$. Similarly, $n_\tau(i)$ denote the number of events close in time to the i-event, considering only those occurring previously to t_i. Conditioning on the marginal spatial locations and the marginal times of the previous events, the in-control distribution of $N_{\rho\tau}(i)$ is simply a hypergeometric distribution with parameters $(i-1)$, $n_\tau(i)$, and $n_\rho(i)$. That is,

$$P\left(N_{\rho\tau}(i) = k\right) = \frac{\binom{n_\tau(i)}{k}\binom{i-1-n_\tau(i)}{n_\rho(i)-k}}{\binom{i-1}{n_\rho(i)}}$$

In Figure 1 we show the space-time events in a three-dimensional plot where the vertical axis indicates the time. The i-event (x_i, y_i, t_i) is at the center of a circle lying in plane parallel to the $x-y$ plane and delineated by dashed lines. This circle has radius equal to ρ. Two additional parallel planes laid at times $t_i - \tau$ and $t_i + \tau$ are also drawn. The count $n(i)$ as defined by Rogerson would be equal to 3, corresponding to the events within the cylinder with center at (x_i, y_i, t_i), radius ρ, and delimited by the planes at $t_i - \tau$ and $t_i + \tau$. This implies on a height equal to 2τ. In contrast, our count $N_{\rho\tau}(i)$ would be equal to 2, as considers only those events (other than the i-th event) in the cylinder with height τ between the planes at $t_i - \tau$ and t_i. The value of $n_\tau(i)$ is equal to 3 and corresponds to all events between the planes at at $t_i - \tau$ and t_i, irrespective of their spatial location. The count $n_\rho(i)$ is also equal to 3 and corresponds to all events within the cylinder centered at the location (x_i, y_i) with radius ρ and height t_i, with bases at $t = 0$ and $t = t_i$.

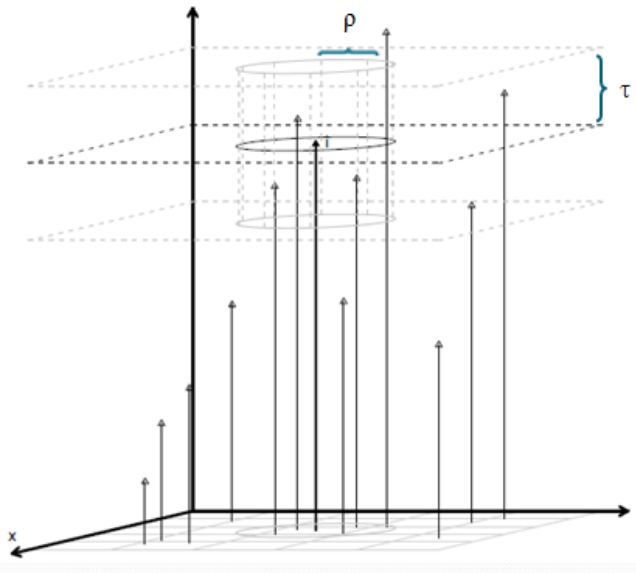


Figure 1. Illustration of three dimensional events with the i-th event (x_i, y_i, t_i) being the center of a circle parallel to the horizontal axes and with radius ρ. Two additional planes are at $t = t_i - \tau$ and $t_i + \tau$.

The mean and the variance of $N_{\rho\tau}(i)$ are given by:

$$E\left(N_{\rho\tau}(i)\right) = \frac{n_\rho(i)n_\tau(i)}{i-1} \tag{5}$$

and

$$\mathrm{Var}\left(N_{\rho\tau}(i)\right) = \frac{n_\rho(i)n_\tau(i)(i-1-n_\rho(i))(i-1-n_\tau(i))}{(i-2)(i-1)^2}. \tag{6}$$

With these moments, we can define the standardized local Knox score z_i and the cumulative sum S_i as in (1) and (2), respectively. The threshold for S_i can be found in the same way as in [1]. However, this threshold choice is based on stochastic assumptions (such as normality) that may not hold true, specially if the expected number of cases is small. We decided to evaluate another practical approach to find the threshold h for the cumulative sum S_i, one based on simulation using a

training sample dataset. This new method requires a set of $m+n$ previous events. Let $\pi_1, \ldots, \pi_B$ represent a large number B of random permutations of the times among the spatial locations. This generate B pseudo datasets $(x_i, y_i, t_{\pi_j(i)})$ where $(\pi_j(1), \ldots, \pi_j(n+m))$ is the j-th random permutation of the vector $(1, 2, \ldots, n+m)$. For each one of these permutations, we calculate the cumulative sum $S_i^j = \max\{0, S_{i-1}^j + z_i^j - k\}$ where z_i^j is the local Knox score evaluated at the i-th event $(x_i, y_i, t_{\pi_j(i)})$ with the j-th permuted dataset.

To the i-th event in the j-th permutation, with $i = m, m+1, \ldots, m+n$, we associate the value $MS_i^j = \max_t\{S_t^j, t = i+1, \ldots, i+d\}$, the maximum of the d successive values after the i-th event of the cumulative sum process based on the j-th permutation. The tuning parameter d plays the same role as in equation (4). It is the reference number of successive events one associates with the unmotivated alarm rate α. We consider only $i > m$ in order to stabilize the local score z_i. The reason is that the counts $n_{\rho\tau}(i)$ will be typically zero when we have few previous events, not reflecting its true distribution after entering a stationary regime. Our experience has shown that, when the spatial and temporal bounds are such that $E(N_{\rho\tau}(i)) \geq 3$, setting $m = 200$ events is enough for that. We obtain the empirical $1-\alpha$ percentile h_i of the values $MS_i^1, \ldots, MS_i^B$ for each $i > m$. Next, we average these percentiles obtaining $h = \sum_i h_i/n$. This provides a value for the threshold h. This threshold means that we expect that, in the absence of emerging localized clusters, a sequence of d events will trigger a false alarm with probability α.

Since we are not willing to assume independence between successive z_i's, we prefer not to use (4) to relate the false alarm rate α in d events with the implied ARL_0. Hence, we found the ARL_0 implied by this h threshold as we explain in section 3.

We illustrate our method with residence place and onset date for 1001 Meningitis cases that occurred between 2001 and 2005 in a 2 million inhabitants Brazilian city, Belo Horizonte. Figure 2 shows the spatial distribution of the events within Belo Horizonte boundaries is the left hand side plot. We can see the large differences in intensities, mostly due to the widely different population density in this town. In 60% of the days no cases were recorded and only one case was recorded in 72% of the remaining days with a maximum number of 6 cases. Exploratory data analysis did not indicate trend or periodicity in these data. We selected $\rho = 2000$ meters, $\tau = 30$ days, and $\alpha = 0.1$ in 100 events. The right hand side plot in Figure 2 shows the cumulative sums S_i versus the i-th event date t_i with the continuous and dashed horizontal lines representing the Rogerson and empirical thresholds. The alarm would have sounded off using either threshold at events $i = 249$ and $i = 608$ while the empirical threshold would also lead to detection at $i = 434$ and $i = 696$.

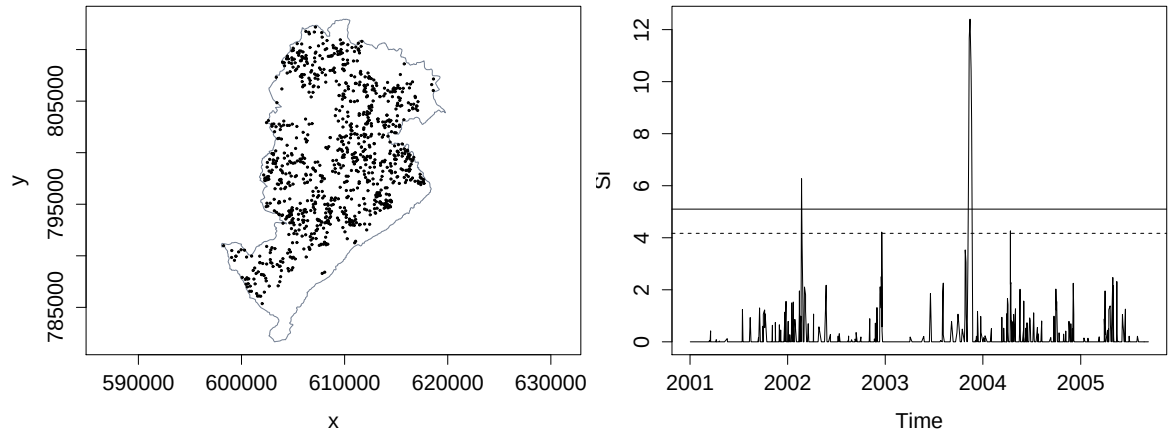


Figure 2. Right: Map of Belo Horizonte boundary with the residential locations of 1001 meningitis cases occurred between 2001 and 2005. Left: Cumulative sum S_i versus i. Two thresholds are shown as horizontal lines.

3. Performance: evaluation measures and assumptions tested

Our work analyzes the performance of our prospective space-time local Knox surveillance method using scenarios under control (without the emergence of clusters) and out-of-control (with the emergence of clusters). Each scenario was independently simulated 1000 times to collect the performance statistics. The events were generated in a cube formed by the spatial coordinates in $[0, 100]^2$ and the times were spread along the time interval $[0, 400]$. The under control scenarios generated events according to a Poisson point process in two different ways. The spatial coordinates are either uniformly distributed or follow an intensity function given by $\lambda(x, y) = 60(2 + \sin(2\pi x/35))$ which produced a non-homogeneous geographical pattern of events. This function was chosen to generate three zones with higher incidence than the rest of the region. The time coordinates are uniformly distributed. A total number of 4000 events were generated implying an average intensity of 0.001 events per space and time units.

For the out-of-control scenarios, we generated the same 4000 events according to the in control situation We added other events on top of these baseline events to create the emerging localized clusters. These new added events are generated from uniform distributions in smaller three-dimensional regions within the study region. We varied the shape, extent, intensity and the emergence time of the clusters. More specifically, the scenarios are the following:

- Variations in the cluster shape: Two scenarios with 400 additional events each. One of them is located in the square $[50, 60]^2$ and the other is quite thin and elongated, covering the band $[50, 52] \times [50, 100]$. Both of them start at time $t = 50$ and have intensity equal to 0.011. The reason for this scenarios is a limitation on the local Knox statistic definition. The geometric structure used to monitor the point pattern is a cylinder centered at each observation. When the cluster shape is different from a cylinder, the method will have more difficulty to detect it. For example, if the cluster has a linear shape along a river or an avenue, we can anticipate a longer waiting time for the alarm to go off.
- Variations in the cluster intensity: three scenarios with the emerging cluster located in the square $[50, 60]^2$ and starting at time $t = 50$. We generated additional events within the clusters producing cluster intensities in excess of 0.007, 0.0110, 0.018 with respect to the baseline intensity equal to 0.001.
- Variations in the emergence time: Two scenarios, both with 400 additional events in the square $[50, 60]^2$ and with different onset times. In one, the cluster starts emerging early at $t = 50$ and, in the other, the cluster starts at $t = 90$. In both cases, the within cluster intensity is equal to 0.101. We expect a larger ARL when the cluster starts early due to the scarce number of previous events for the method to learn the marginal spatial and temporal patterns.
- Variations in the cluster extent: three scenarios, all of them having the added events' intensity approximately equal to 0.01 within the clusters. Since the baseline intensity is equal to 0.001, this gives clusters of different extent but having all of them an intensity about 100 times larger than the baseline intensity. The relatively small geographical cluster is located in the region $[52, 60]^2$ and has 204 additional events uniformly distributed between times $[52, 400]$. The moderately extent cluster has 400 added events sized in the region $[50, 60]^2 \times [50, 400]$, while the relatively widespread cluster has 1600 added events in the region $[40, 60]^2 \times [40, 400]$. These scenarios will be denoted by 1, 2, and 3, in increasing order of extent.

As pointed out by a referee, our cluster scenarios show an increased intensity in a fixed and localized geographical shape. The cluster does not change as times passes. This might be a good model for point-source outbreaks, but not for some epidemic diseases. Such type of spread is currently not considered.

To run the prospective surveillance method, the user needs to select the spatial and temporal bounds ρ and τ. We selected three spatial values ($\rho = 5, 10$, and 20), and three time values ($\tau = 5, 10$, and 20). We crossed all the combinations of these critical limits, in a total of 9 pairs of spatial and time bounds. Besides these bounds, the user also needs to establish the threshold h. For this, we show results for two different choices of h. One of them obtains h empirically simulating the under control scenarios as explained previously. We ran B=1000 simulations to determine the threshold h and used $\alpha = 0.1$ and $d = 100$ producing then a 10% chance of false alarm on each block of 100 consecutive events. The other option is to use the threshold determination suggested by [1]. To compare with the empirically obtained threshold, we found the nominal ARL_0 using (4) with the same values of $\alpha = 0.1$ and $d = 100$ and then we used the approximation (3) to establish the value of h. This is called Rogerson Threshold in the results.

Let rl_i be the realized run length in the i-th simulation, $i = 1, \ldots, B$. The number of simulations in which the alarm went off is equal to $\sum_i I[rl_i < t_{4000}]$, where $I[A]$ is the indicator function evaluated at the event A. Let ϕ be the emergence time of the cluster, with $\phi = \infty$ when the process is under control. The number of false alarms in a set of simulations is given by $\sum_i I[rl_i < \phi]$. For out-of-control scenarios, we have one additional quantity of interest. Conditionally on the alarm going off in the i-th out-of-control simulation, we define by c_i the number of events belonging to the cluster occurring up to the time the alarm goes off.

Using these definitions, we calculated simple statistics to evaluate the performance of the method. We calculate a false alarm rate defined as

$$FAR = \frac{\sum_i I[rl_i < \phi]}{\sum_i I[rl_i < t_{4000}]} .$$

In the under control scenario, all the alarms are false and FAR is equal to 1.

In the out-of-control situation, the alarm went off in virtually all simulations carried out. In the in control situation, however, there some scenarios in which there was up to 45% of censoring. That is, in several simulations, the alarm did not went off after 4000 events being observed. To track this censoring of the run length, we calculate the statistic

$$PA = \frac{\sum_i I[rl_i < t_{4000}]}{B} .$$

Everything else equal, the smaller PA, the longer the average run length of the alarm system.

To distinguish the average run length under control, ARL_0, from this average run length when the system is out-of-control we use ARL to denote the latter. This measure is influenced by the emergence time of the real cluster. The longer it takes to appear, the longer will be the ARL. To avoid this influence, we calculate the conditional expected delay CED. Conditional on the alarm going off after the cluster emergence time, the CED is the average number of events from the cluster one needs to wait and it was estimated as:

$$CED = \frac{\sum_i c_i}{\sum_i I[rl_i < t_{4000}]} .$$

Table 1. Results for the under control situation, homogeneous intensity. The empirical threshold is denoted by ET. The Rogerson threshold is constant and equal to $h = 5.01$. The associated ARL_0 based on formula (3) is equal to 949.12.

ρ	τ	PA	ET	$ARL_0 - EMV$	$ARL_0 - Reg$
5	5	0.66	3.89	3168.69	3238.55
5	10	0.60	3.83	3325.63	3560.02
5	20	0.55	3.79	3518.60	4270.23
10	5	0.82	4.18	2605.16	2360.18
10	10	0.82	4.18	2496.66	2162.70
10	20	0.81	4.31	2525.44	2124.43
20	5	0.97	4.94	1695.19	1377.79
20	10	0.98	5.01	1304.42	1018.36
20	20	0.95	5.03	1540.70	1176.56

We used the in control situation to study the appropriateness of the geometric distribution assumption for the run length RL that is behind the expression between the ARL_0 and α in the formula (4) proposed by [1]. This analysis was based on QQ-plots and statistical tests of the observed run lengths in each in control scenario and, as we will show, that distribution provides an excellent fit to the data and it is warranted.

Based on the geometric distribution for the RL, we estimated the ARL_0 of the method in each scenario. To deal with the censoring of the run lengths, specially in the in control situation, we used two procedures. One is based simply in the least squares regression estimate of the intercept in the QQ-plots of the RL's.

Another approach is to use the maximum likelihood estimate of the ARL_0 and is based on a censored data model assuming the geometric distribution for the run lengths. In our simulations, we found empirical evidence that this assumption seems reasonable (see Figure 3). Namely, we have k observed run lengths $rl_1, \ldots, rl_k$ and $n - k$ censored run lengths at the maximum value c. If we assume that each run length RL follows a geometric distribution with parameter θ then $E(RL) = 1/\theta$ and the maximum likelihood estimate of the average run length ARL_0 is given by $1/\hat{\theta}$. That is,

$$\hat{ARL_0} = \frac{\sum_{i=1}^{k} rl_i + (n-k)(c-1)}{k} \tag{7}$$

Note that we work in a different order than that used by Rogerson [1]. He starts with α and d, finds ARL_0 by means of (4) and then uses (3) to obtain the threshold h. In contrast, we set α and d, find the threshold empirically with no need to calculate an intermediate ARL_0. The formula (3) should not be used since it underestimates the true ARL_0, as we will see in the next section.

4. Simulations

4.1. Results Under Control

Table 1 shows the results when the baseline intensity is geographically constant and we are in the under control situation. Only the homogeneous results will be shown here, the non-homogeneous results being essentially the same. The first two columns give the bounds ρ and τ for the neighborhood determination between pairs of events. The Rogerson threshold produced $h = 5.01$. This value was calculated using $\alpha = 0.1$ and $d = 100$ in (4) and inputting the found $ARL_0 = 949.12$ in (3). Note that this Rogerson threshold does not depend on the values of ρ and τ. The alarm triggering threshold used in the resulting statistics are based on the empirical thresholds, found by permutation. The Rogerson threshold is given only for comparison purposes with the empirical one.

The first three scenarios, with a relatively small spatial radius $\rho = 5$ present a sizable percentage of no triggering alarms at the end of the simulations. This is somewhat surprising given the relatively high false alarm probability equal to 0.1 in $d = 100$ consecutive events. We would expect a much larger number of simulations with alarms going off eventually during the 4000 events generated in each simulation. We conjecture that this result may be due to the auto-correlation in the cumulative sum series S_i. The most relevant column in table 1 is the one showing the empirical threshold. In contrast with the constant value of 5.01 obtained by Rogerson procedure, we found a clear trend as the neighborhood search cylinder changes its volume. The parameter τ has very little effect but changes in ρ alter significantly the empirical threshold with the larger cylinders giving approximately the same value as Rogerson threshold. Therefore, when ρ and τ are large enough to provide about 6 expected events in the search cylinder of the local Knox statistics, the simple Rogerson calculation for h can be used. The empirically estimated threshold h passes from around 3.8 to 5.0 as ρ increases and so it is always smaller than the Rogerson threshold. This implies that, using the simple analytical method proposed by Rogerson to determine the threshold for S_i, we will be on the conservative side, taking in fact longer to trigger a false alarm than the nominal value obtained by those methods when the number of events in the cylinder is small.

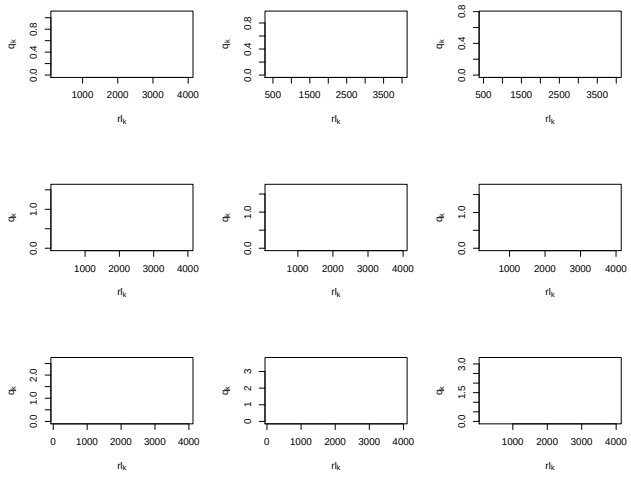


Figure 3. QQ-plots to verify the assumption of a geometric distribution for the run length RL random variable. The columns show the results for $\tau = 5$, $\tau = 10$, and $\tau = 20$, respectively. The rows correspond to $\rho = 5$, $\rho = 10$, and $\rho = 20$, respectively. The horizontal axis shows the order statistics $rl_{(k)}$ and the vertical axis shows the values of $q_k = -\log(1 - k/4000)$.

To obtain an estimate of these ARL_0, we verified the goodness of fit of the geometric distribution for the RL random variable under control. Figure 3 shows the QQ-plot of the RL observed values for each one of the nine possible combinations of ρ and τ. The censored values in each case are all larger than $t = 4000$ and they are not shown in the plots. Based on these observations, it is clear that the geometric distribution provides an excellent fit to the run length distribution in the in-control situation.

We estimated the ARL_0 in two different ways. In the first one, we used the maximum likelihood method assuming a geometric distribution for the censored observed RL sample. The results are in the fifth column and they show a substantial decrease as either ρ and τ enlarge, with the latter having much less impact. The ARL_0 based in the regression line fitted to the QQ-plots give similar results to the EMV estimates and they are in the last column of Table 1. Using ARL_0 equal to 949.12, the threshold h based on formula (3) is equal to 5.01, similar to the last rows in Table 1. The empirically estimated ARL_0 are rather larger than this nominal mean value. Given the very good fit of the exponential distribution to the data, we prefer the EMV estimates for the ARL_0.

4.2. Non-homogeneous Poisson process

The effect of population distribution is an issue often discussed. The performance of the global Knox statistic is not affected by temporal variation of events intensity if this temporal variation is constant in space. It is not affected either by the spatially heterogeneous distribution of the risk population if this spatial variation is constant in time. It is not clear if this good behavior is also present in the surveillance method based on the local Knox statistics.

Comparing Tables 1 and 2, we see that the percentage of alarms that went off during the simulations is very similar in both situations. The ARL_0 estimated by either the maximum likelihood method or the simple regression method applied to the QQ-plot are also very similar to the homogeneous case. The geometric distribution assumption for the run length has an excellent fit to the data, producing plots similar to those in Figure 3. The most relevant tuning parameter is the threshold h. Table 2 shows that the empirical thresholds are very close to those values found in the homogeneous case in Table 1. Therefore, the threshold h of the local Knox statistic is not substantially affected when the population density varies in space. Note that the h threshold proposed by Rogerson is constant and equal to 5.01 for all spatial and temporal bounds in both cases, homogeneous and heterogeneous. As we found previously, this h threshold is appropriate if the tuning parameters ρ and τ are large enough to contain about six events, in average.

The CUSUM method is usually adopted to monitor i.i.d. observations that follow a normal distribution. Through their simulations, Marshall *et al.* (2007) concluded that the normal approximation is not valid for the local Knox standardized statistics Z_i. We found this same result in our own simulations in both situations, the homogeneous and the non-homogeneous cases. As the formula (3) associating the ARL_0 and the threshold h is based on this normality assumption, this may be one explanation for its bad fit to the empirical thresholds and the ARL_0 we found in our simulations.

4.3. Results out-of-control

In this subsection, we present the out-of-control situation performance results of the local Knox surveillance method using the empirical threshold. Results using the threshold (3) are available in the Supplementary Web Materials. Although the

Table 2. Results for the under control situation, non-homogeneous Poisson process intensity. The Rogerson threshold is constant and equal to $h = 5.01$ and its associated ARL_0 based on formula (3) is equal to 949.12.

ρ	τ	PA	ET	ARL_0-EMV	$ARL_0 - Reg$
5	5	0.67	3.94	3172.72	3221.75
5	10	0.57	3.85	3355.41	3649.80
5	20	0.55	3.77	3491.60	4180.61
10	5	0.83	4.23	2438.89	2153.87
10	10	0.84	4.25	2536.76	2182.46
10	20	0.82	4.33	2575.64	2150.31
20	5	0.98	4.96	1304.00	1036.54
20	10	0.97	4.97	1342.07	1050.27
20	20	0.95	5.03	2088.05	1891.51

numbers are different, the qualitative conclusions are the same as those using our empirical threshold. The range of our simulation scenarios allows us to understand which factors cause larger impacts on the performance of the method and how they occur. The factors considered were the geometrical shape, intensity, time of emergence, and extent of the cluster. We present results only for the homogeneous spatial pattern, with the non-homogeneous case having similar results.

4.3.1. Impact of Cluster Shape Table 3 shows the results of running the space-time local Knox statistics when two different cluster shapes are considered, a cube shaped cluster and an elongated cluster shaped. Let us focus initially only on the general pattern of the performance measures when the cluster is cubic shaped. When $\rho = 5$, the spatial search area around each newly arriving event is equal to $\pi 5^2 \approx 79$, smaller than the true cluster area, which is equal to $10 \times 10 = 100$. The ARL and CED in this situation is substantially smaller than that when the spatial bound ρ increases to 10, turning the search area of each moving cylinder equal to 314, larger than the true cluster area. By increasing ρ further, taking a search area equal to $\pi 20^2 = 1257$, much larger than the real cluster area, the CED and ARL start to increase and hence we have an undesirable situation. There is also a cost in an increased value for FAR, the false alarm rate. From virtually zero when $\rho \leq 10$, it increases substantially to values above 13%. That is, with a very large radius ρ one starts to have too many false alarms. To reach a balance between a small CED and a low value for the FAR, we recommend that the user selects ρ about the size of the anticipated diameter of the true circular cluster. This suggestion is in keeping with the matched filter theorem, which concludes that the size of the filter used to maximize the signal-noise ratio should be equal to the size of the feature to be extracted [28].

When the true cluster is linear shaped, the performance deteriorates substantially as compared with the true cluster being a cube. This is expected as the search area of the local Knox statistic is quite different from the true cluster and hence, much of its area intersects a non-cluster region.

4.3.2. Impact of the cluster intensity As one can expect, increasing the point pattern intensity within the cluster speeds up the detection of the emergent cluster. We can see in Table 4 that, for all choices of ρ and τ, the CED and ARL decreases as we pass from low intensity to high intensity. As we saw previously, selecting too large values for the bounds ρ and τ is not adequate. Indeed, the CED and FAR increase as ρ increases beyond the extent of the true cluster.

4.3.3. Impact of the emergence time Table 5 shows the results for two scenarios differing with respect to the emergence time of the cluster. One has the cluster emerging early during the surveillance procedure, at $t = 50$, while the other starts emerging later, at $t = 90$. We can see that the ARL increases for the late emergent cluster. This happens because the surveillance method must wait until the cluster emerges to start detecting something. The more interesting measure is the CED which shows that a late emergence pays off in terms of a quicker detection time. This is so because the marginal spatial and temporal patterns are more reliably estimated in the calculations of the local scores z_i. We can also see that the FAR increases substantially when the cluster starts emerging late. This is explained by the longer risk exposure of false alarms when the cluster emerges late.

4.3.4. Impact of cluster extent Table 6 shows the results for three increasing cluster extents, denoted by extent 1,2, and 3. The ARL shows that, increasing the cluster extent, the time needed for the alarm to go off decreased. The same decreasing pattern can be found in the CED column. Once again, the CED has a reverse behavior when the tuning parameter ρ is equal to 20 and the cluster extent is equal to 1 or 2. With the largest extent 3 cluster, we have a monotone decrease of the CED because the search area associated with the local Knox statistic with radius equal to 10 has the smallest difference with respect to the cluster extent.

Table 3. Results for the out-of-control situation with different cluster shapes.

Scenario with cube shaped cluster					
ρ	τ	ARL	CED	PA	FAR
5	5	662.02	17.50	1.00	0.06
5	10	626.58	13.23	1.00	0.02
5	20	649.97	15.50	1.00	0.00
10	5	556.51	8.88	1.00	0.18
10	10	568.08	8.66	1.00	0.13
10	20	616.29	12.39	1.00	0.04
20	5	610.68	16.84	1.00	0.27
20	10	617.89	15.27	1.00	0.19
20	20	647.35	16.53	1.00	0.13
Scenario with line shaped cluster					
ρ	τ	ARL	CED	PA	FAR
5	5	1023.13	54.51	1.00	0.07
5	10	939.43	45.29	1.00	0.03
5	20	859.06	36.86	1.00	0.00
10	5	811.69	34.83	1.00	0.18
10	10	721.11	24.23	1.00	0.13
10	20	725.43	23.41	1.00	0.06
20	5	709.03	26.51	1.00	0.25
20	10	679.22	21.48	1.00	0.19
20	20	702.39	22.06	1.00	0.12

Table 4. Results for the out-of-control situation with different cluster intensities.

Scenario with low intensity					
ρ	τ	ARL	CED	PA	FAR
5	5	1135.42	35.18	0.99	0.08
5	10	1064.03	30.84	1.00	0.03
5	20	1023.34	28.43	1.00	0.00
10	5	910.88	23.85	0.99	0.18
10	10	810.42	17.80	1.00	0.13
10	20	824.76	17.95	1.00	0.05
20	5	971.30	28.15	0.99	0.25
20	10	966.80	26.75	1.00	0.20
20	20	942.30	24.89	0.99	0.13
Scenario with average intensity					
ρ	τ	ARL	CED	PA	FAR
5	5	662.02	17.50	1.00	0.06
5	10	626.58	13.23	1.00	0.02
5	20	649.97	15.50	1.00	0.00
10	5	556.51	8.88	1.00	0.18
10	10	568.08	8.66	1.00	0.13
10	20	616.29	12.39	1.00	0.04
20	5	610.68	16.84	1.00	0.27
20	10	617.89	15.27	1.00	0.19
20	20	647.35	16.53	1.00	0.13
Scenario with high intensity					
ρ	τ	ARL	CED	PA	FAR
5	5	552.86	8.99	1.00	0.09
5	10	569.71	10.40	1.00	0.04
5	20	606.05	15.39	1.00	0.01
10	5	514.39	6.52	1.00	0.20
10	10	539.83	8.04	1.00	0.14
10	20	580.90	12.21	1.00	0.06
20	5	528.09	10.74	1.00	0.25
20	10	549.24	11.70	1.00	0.19
20	20	599.76	15.96	1.00	0.11

Table 5. Results for the out-of-control situation with different cluster emergence times.

Early emergence time					
ρ	τ	ARL	CED	PA	FAR
5	5	662.02	17.50	1.00	0.06
5	10	626.58	13.23	1.00	0.02
5	20	649.97	15.50	1.00	0.00
10	5	556.51	8.88	1.00	0.18
10	10	568.08	8.66	1.00	0.13
10	20	616.29	12.39	1.00	0.04
20	5	610.68	16.84	1.00	0.27
20	10	617.89	15.27	1.00	0.19
20	20	647.35	16.53	1.00	0.13
Late emergence time					
ρ	τ	ARL	CED	PA	FAR
5	5	878.91	6.56	1.00	0.31
5	10	918.79	7.38	1.00	0.22
5	20	975.61	10.84	1.00	0.10
10	5	800.64	5.20	1.00	0.41
10	10	828.96	5.28	1.00	0.39
10	20	914.48	8.34	1.00	0.25
20	5	784.32	8.77	1.00	0.47
20	10	806.14	8.11	1.00	0.43
20	20	887.38	10.45	1.00	0.35

Table 6. Results for the out-of-control situation with different cluster extents.

Scenario with cluster extent 1					
ρ	τ	ARL	CED	PA	FAR
5	5	1000.89	27.25	0.99	0.08
5	10	892.02	21.10	0.99	0.04
5	20	832.59	17.57	1.00	0.01
10	5	883.39	21.90	0.99	0.19
10	10	770.44	15.11	1.00	0.16
10	20	791.38	15.54	1.00	0.06
20	5	967.65	27.69	1.00	0.25
20	10	938.83	25.24	1.00	0.21
20	20	931.67	23.85	1.00	0.14
Scenario with cluster extent 2					
ρ	τ	ARL	CED	PA	FAR
5	5	662.02	17.50	1.00	0.06
5	10	626.58	13.23	1.00	0.02
5	20	649.97	15.50	1.00	0.00
10	5	556.51	8.88	1.00	0.18
10	10	568.08	8.66	1.00	0.13
10	20	616.29	12.39	1.00	0.04
20	5	610.68	16.84	1.00	0.27
20	10	617.89	15.27	1.00	0.19
20	20	647.35	16.53	1.00	0.13
Scenario with cluster extent 3					
ρ	τ	ARL	CED	PA	FAR
5	5	429.57	15.06	1.00	0.05
5	10	444.35	20.59	1.00	0.01
5	20	470.60	32.75	1.00	0.00
10	5	409.71	9.13	1.00	0.10
10	10	421.00	12.01	1.00	0.07
10	20	441.25	19.22	1.00	0.01
20	5	392.80	8.09	1.00	0.18
20	10	408.62	10.67	1.00	0.14
20	20	431.18	16.63	1.00	0.07

5. Discussion and conclusions

The space-time prospective surveillance is a difficult problem with few proposals available for public health agencies. In this paper, we propose a method that is simple to use and requires very little statistical modeling from the user. Its tuning parameters are interpretable and can be set by consideration of the specific aspects of the disease under surveillance. Our method is based on previous proposal made by Rogerson [1] which was concerned to the detection of retrospective space-time clusters. This method has been criticized as inadequate in [22] but we showed that their negative results are due to the retrospective definition of the monitoring local Knox statistic. Generalizing to take into account the prospective view and after obtaining the new distribution under control, the method based on local Knox statistics has a very good performance, as we verified by extensive simulations.

Under control, we found that the ARL_0 follows an exponential distribution and that, although not explicitly stated, the Rogerson approach to determine h works only for situations where the search areas contain at least six events. Indeed, the functional relationship between ARL_0 and the tuning parameter h given by equation (3) is not valid unless ρ and τ are large enough to contain approximately six events. These conclusions are valid either under homogeneous or inhomogeneous population density. Under out-of-control, the results were as expected. The emergence time of the cluster is an important factor in this method. The later the cluster appears, the quicker it is detected. The shape of the cluster also affects its detection. Since the search region is a cylinder, highly non-circular shaped clusters tend to take longer to be detected. Higher intensity and larger clusters are detected more easily.

There are some limitations associated with our method. The user must specify the tuning parameters ρ, τ and d. The specification of d is a desired aspect of the method since this is a parameter the user wants to have control. As with ρ and τ, the user must select what seems reasonable in each specific application. These parameters should be chosen with respect to the anticipated extent of the clusters. This is the main limitation of the method but we found in our simulations that it only affects its performance, not its validity. Another limitation is the requirement of a training set in order to establish a threshold h. However, we think that is most practical cases this set will be available in the form of historical events previously recorded. Furthermore, in the cases where the user chooses large values for ρ and τ, the threshold h might be set by equation (3). We conclude that a surveillance method with properly defined local Knox statistics is a viable method for a problem that has few methods available and that is too important to be ignored by statisticians.

Acknowledgements

The authors thank Marcos Prates for helpful suggestions and help with the C code to run the simulations. We also want to thank the reviewers for their careful reading. Their suggestions improved substantially this paper.

References

[1] Rogerson, PA. Monitoring point patterns for the development of space–time clusters. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2001; **164**(1):87–96.

[2] Sonesson, C, Bock, D. A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society, Series A* 2003; **166**:5–21.

[3] Lawson, AB, Kleinman, K, Wiley, J. *Spatial and syndromic surveillance for public health.* Wiley Online Library, 2005.

[4] Unkel, S, Farrington, CP, Garthwaite, PH, Robertson, C, Andrews, N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A* 2012; **175**:49–82.

[5] Raubertas, RF. An analysis of disease surveillance data that uses the geographic locations of the reporting units. *Statistics in Medicine* 1989; **267**:267–271.

[6] Järpe, E. Surveillance of the interaction parameter of the ising model. *Communications in Statistics – Theory and Methods* 1999; **28**:3009 – 3027.

[7] Kulldorff, M. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A* 2001; **164**:61–72.

[8] Neill, DB, Moore, AW, Sabhnani, M, Daniel, K. Detection of emerging space-time clusters. In *KDD'05: Proceedings of the Eleventh ACM SIGKDD international Conference on Knowledge Discovery in Data Mining (Chicago, Illinois, USA, August 21 - 24, 2005)*. 218–227.

[9] Kulldorff, M, Heffernan, R, Hartman, J, Assunçao, R, Mostashari, F. A space–time permutation scan statistic for disease outbreak detection. *PLoS medicine* 2005; **2**(3):216–224.

[10] Assunção, R, Maia, A. A note on testing separability in spatial-temporal marked point processes. *Biometrics* 2007; **63**(1):290–294.

[11] Woodall, WH, Brooke Marshall, J, Joner Jr, MD, Fraker, SE, Abdel-Salam, ASG. On the use and evaluation of prospective scan methods for health-related surveillance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2008; **171**(1):223–237.

[12] Assunção, R, Correa, T. Surveillance to detect emerging space–time clusters. *Computational Statistics & Data Analysis* 2009; **53**(8):2817–2830. URL http://www.sciencedirect.com/science/article/pii/S0167947308004921.

[13] Shiryayev, AN. On the detection of disorder in a manufacturing process. i. *Theory of Probability & Its Applications* 1963; **8**(3):247–265. URL http://epubs.siam.org/doi/abs/10.1137/1108029.

[14] Höhle, M. Additive-multiplicative regression models for spatio-temporal epidemics. *Biometrical journal* 2009; **51**(6):961–978. URL http://onlinelibrary.wiley.com/doi/10.1002/bimj.200900050/full.

[15] Meyer, S, Elias, J, Höhle, M. A space–time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics* 2012; **68**(2):607–616.

[16] Diggle, P, Rowlingson, B, Su, Tl. Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics* 2005; **16**(5):423–434.

[17] Vidal Rodeiro, CL, Lawson, AB. Monitoring changes in spatio-temporal maps of disease. *Biometrical Journal* 2006; **48**(3):463–480. URL http://onlinelibrary.wiley.com/doi/10.1002/bimj.200510176/abstract.

[18] Corberán-Vallet, A, Lawson, AB. Conditional predictive inference for online surveillance of spatial disease incidence. *Statistics in medicine* 2011; **30**(26):3095–3116.

[19] Frisén, M, Andersson, E, Schiöler, L. Evaluation of multivariate surveillance. *Journal of Applied Statistics* 2010; **37**(12):2089–2100.

[20] Frisén, M, Andersson, E, Schiöler, L. Sufficient reduction in multivariate surveillance. *Communications in Statistics-Theory and Methods* 2011; **40**(10):1821–1838.

[21] Rogerson, PA. Surveillance systems for monitoring the development of spatial patterns. *Statistics in Medicine* 1997; **16**(18):2081–2093.

[22] Marshall, JB, Spitzner, DJ, Woodall, WH. Use of the local knox statistic for the prospective monitoring of disease occurrences in space and time. *Statistics in medicine* 2007; **26**(7):1579–1593.

[23] Knox, E. The detection of space-time interactions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1964; **13**(1):25–30. URL http://www.jstor.org/stable/10.2307/2985220.

[24] Kulldorff, M, Hjalmars, U. The knox method and other tests for space-time interaction. *Biometrics* 1999; **55**(2):544–552.

[25] Assunção, R, Tavares, A, Correa, T, Kulldorff, M. Space-time cluster identification in point processes. *Canadian Journal of Statistics* 2007; **35**(1):9–25.

[26] Simões, TC, Assunção, RM. Sistema de vigilância para detecção de interações espaço-tempo de eventos pontuais. In *GEOINFO 2005 - Proocedings of the VII Simpósio de Geoinformática Fonseca F. and Casanova M. A., eds., São José dos Campos: Instituto Nacional de Pesquisas Espaciais, INPE.* 381–393.

[27] Rogerson, PA. Formulas for the design of cusum quality control charts. *Communications in Statistics - Theory and Methods* 2006; **35**(1):373–383.

[28] Rosenfeld, A, Kac, AC. *Digital Picture Processing, volume 2.* Academic Press, 1982.