

**UNIVERSIDADE FEDERAL DE MINAS GERAIS**  
**Spencer Barbosa da Silva**

**ANÁLISE DE FRONTEIRAS GEOGRÁFICAS E LIMITES NO  
PROCESSO TEMPORAL PARA AVALIAÇÃO DE  
CONGLOMERADOS ESPAÇO-TEMPORAIS**

BELO HORIZONTE  
2016

**Spencer Barbosa da Silva**

**ANÁLISE DE FRONTEIRAS GEOGRÁFICAS E LIMITES NO  
PROCESSO TEMPORAL PARA AVALIAÇÃO DE  
CONGLOMERADOS ESPAÇO-TEMPORAIS**

Tese apresentada ao Programa de Pós-Graduação em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial a obtenção do título de Doutor em Estatística.

Orientador: Professor Dr. Luiz H. Duczmal

Co-orientador: Professor Dr. Anderson R. Duarte

BELO HORIZONTE  
2016

## **DEDICATÓRIA**

Dedico este trabalho à minha esposa Daniela e filhas Gabriela e Alice, aos meus pais, meus fiéis seguidores, aos meus irmãos, que sempre me apoiaram, aos meus amigos, presentes em todos os momentos da minha vida, aos meus professores, que contribuíram muito com seus ensinamentos, aos meus colegas de trabalho, enfim, a todos que puderam de alguma forma serem facilitadores para a construção da minha formação.

**Amo Vocês!**

## **EPIGRAFE**

“A ciência é manca sem a religião; a religião é cega sem a ciência.”

**Albert Einstein**

“Apesar dos nossos defeitos, precisamos enxergar que somos pérolas únicas no teatro da vida e entender que não existem pessoas de sucesso ou pessoas fracassadas. O que existem são pessoas que lutam pelos seus sonhos ou desistem deles.”

**Augusto Cury**

## **AGRADECIMENTOS**

Início agradecendo a Deus por estar provendo este momento tão especial em minha vida, me concedendo saúde e força de vontade para buscar novos conhecimentos e qualidade profissional.

À minha esposa Daniela, companheira, amiga, em todos os momentos da minha jornada você é e sempre será o grande amor da minha vida. Te amo!

Minhas filhas, Gabriela, sempre carinhosa, companheira de estudos na grande maioria das vezes, e agora, a mais nova integrante da nossa família, Alice, que com certeza receberá muito amor e carinho! Amo vocês.

À minha família, por ter me apoiado e acolhido nos momentos alegres e tristes, apoio este concebido de forma carinhosa e incondicional. Em especial à meus pais, Geraldo e Francisca, pelo apoio e carinho recebido durante toda a minha vida, carinho este que foi transformado por mim em amor.

Ao meu amigo e co-orientador professor Anderson Duarte, pela paciência e dedicação durante todo este período em que estivemos juntos trabalhando, sempre motivador, exemplo de dedicação e profissionalismo.

Ao meu orientador Luiz Duczmal, uma pessoa atenciosa e humana, de coração puro e sempre disponível a ajudar com sua genialidade e paciência.

Quanto aos amigos, não posso deixar de citá-los, a amizade construída ao longo dos anos foi uma injeção de ânimo em todos os momentos difíceis. Sei que serei injusto ao citar nomes, provavelmente esquecerei de alguém, se isto ocorrer, por favor me perdoem. Não posso deixar de citar pessoas como: Flávio, Alexandre Celestino, Tiago, Diana, Marcelo, Fernando, Graziela, Ricardo, Juliano, Helgem, Rosielle, Victor, Érica, Ivair, Maria Cláudia e Eduardo pelo apoio incondicional nos corredores da UFOP.

Aos professores, funcionários e colegas do Departamento de Estatística da UFMG, em particular a Rogéria, Rose, José Luís e Paulo que sempre se mostraram solícitos nas minhas dificuldades. Muito obrigado!

Aos demais companheiros de profissão da UFOP que sempre buscaram me apoiar e foram facilitadores para que eu pudesse concluir esse trabalho.

Enfim, a todos aqueles que de uma forma ou de outra me ajudaram a chegar ao término desse trabalho.

## LISTA DE FIGURAS

1	Exemplo hipotético de um cilindro $\xi^{t_1, t_2}$ em um mapa $\mathcal{S}$ . . . . .	26
2	Taxas de mortalidade para câncer de tireoide . . . . .	39
3	Mapa do Novo México/EUA. . . . .	40
4	Função $\mathcal{F}$ para câncer de tireoide no Novo México/EUA. . . . .	41
5	Alarme através da função $\mathcal{F}$ para câncer de tireoide no Novo México/EUA. . . . .	42
6	Solução para câncer de tireoide através do Scan Prospectivo. . . . .	43
7	Alarme para câncer de tireoide através do Scan Prospectivo. . . . .	44
8	Dissipação espacial do efeito da função $\mathcal{F}$ espaço-temporal. . . . .	44
9	Dissipação espacial do efeito da Estatística Scan Prospectiva. . . . .	45
10	Solução para câncer de tireoide (Razão de verossimilhança adaptativa). . . . .	45
11	Alarme para câncer de tireoide (Razão de verossimilhança adaptativa). . . . .	46
12	Dissipação espacial do efeito da Razão de verossimilhança adaptativa. . . . .	46
13	Prevalência do Hanseníase (por 10 mil habitantes) no Amazonas/BRA. . . . .	48
14	Mapa do Amazonas/BRA. . . . .	49
15	Função $\mathcal{F}$ para Hanseníase no Amazonas/BRA. . . . .	51
16	Alarme através da função $\mathcal{F}$ para Hanseníase no Amazonas/BRA. . . . .	52
17	Dissipação espacial da função $\mathcal{F}$ no Amazonas/BRA. . . . .	53
18	Solução para Hanseníase através do Scan Prospectivo. . . . .	54
19	Alarme para Hanseníase através do Scan Prospectivo. . . . .	55
20	Dissipação para Estatística Scan Prospectiva no Amazonas/BRA. . . . .	56
21	<i>Cluster</i> artificial no mapa do Novo México/EUA. . . . .	60
22	<i>Cluster</i> artificial espacial no mapa do Novo México/EUA. . . . .	66
23	Comparação entre os delineamentos propostos. . . . .	68

## LISTA DE TABELAS

1	Sub-regiões do mapa do Novo México/EUA .....	40
2	Sub-regiões do mapa do Amazonas/BRA .....	50
3	Performance através de medidas de desempenho .....	62
4	Avaliação espacial através de medidas de desempenho .....	68
5	Comparação de performance .....	70

## RESUMO

A investigação de *clusters* espaço-temporais tem se tornado um assunto de vasto interesse nas pesquisas científicas. A velocidade da propagação de diversos fenômenos tem impulsionado esse interesse. Em particular, estudos associados à formação de *clusters* de doenças e propagação de surtos infecciosos são problemas práticos bastante relevantes nessa área. A função  $\mathcal{F}$  já se mostrou eficiente no delineamento de *clusters* puramente espaciais, por outro lado, a análise puramente espacial tem se mostrado ineficaz para contornar diversas situações. Na prática, a busca pelo delineamento tanto no espaço quanto no instante de início da formação do *cluster* são de extrema valia. Propõe-se a extensão da função  $\mathcal{F}$  para o problema espaço-temporal e a introdução de uma nova tecnologia capaz de avaliar a intensidade dos possíveis instantes de alteração efetiva no processo temporal, ou seja, delinear o possível instante de surgimento de *clusters* (alarme para *clusters* emergentes). As técnicas em avaliação serão aplicadas em bases de dados reais, com o interesse em aferir qualidade e comparar com mecanismos de detecção e inferência já existentes.

**Palavras-chave:** *Clusters* Espaço-Temporais; Função  $\mathcal{F}$ ; Estatística Scan; Delineamento de *Clusters*; Alarme para *Clusters* Emergentes.

## ABSTRACT

Research into space-time clusters has become a subject of broad interest in scientific research. The speed of propagation of several phenomena has encouraged this interest. Particularly, associated studies with formation of clusters and spread of infectious outbreaks are quite substantial practical problems in this area. The  $\mathcal{F}$ -function has proved efficient in designing purely spatial clusters, in other hand, the purely spatial analysis has proven ineffective to overcome various situations. In practice, tools for searching simultaneously in space and start time of cluster formation are extremely valuable. It is proposed an extension to the  $\mathcal{F}$ -function for the space-time problem and the introduction of a new technology capable of evaluating the intensity of possible moments of effective change in the temporal process, or to delineate a possible start time to clustering (alarm for emerging clusters). The techniques under evaluation are applied in real benchmark data, with the interest in assessing quality and compare with detection and inference mechanisms existing.

**Keywords:** Space-time clusters;  $\mathcal{F}$ -function; Scan statistic; Clusters design; Alarm for emerging clusters.

# SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b> .....	<b>12</b>
1.1	FORMATAÇÃO DE OBJETIVOS .....	14
<b>2</b>	<b>REFERÊNCIAL TEÓRICO</b> .....	<b>17</b>
<b>3</b>	<b>DETECÇÃO DE CLUSTERS-ESPAÇO-TEMPO</b> .....	<b>23</b>
3.1	ESTATÍSTICA SCAN ESPAÇO-TEMPO .....	23
3.2	FUNÇÃO INTENSIDADE.....	29
<b>4</b>	<b>VISUALIZAÇÃO DA INCERTEZA PARA CLUSTERS ESPAÇO-TEMPORAIS</b> <b>31</b>	
4.1	FUNÇÃO $\mathcal{F}$ ESPAÇO-TEMPORAL .....	31
4.2	ALARME PARA MUDANÇAS NO PROCESSO.....	33
4.3	ALARME ATRAVÉS DA FUNÇÃO $\mathcal{F}$ .....	35
<b>5</b>	<b>RESULTADOS EXPERIMENTAIS</b> .....	<b>38</b>
5.1	DADOS DE CÂNCER DE TIREOIDE - ESTADOS UNIDOS DA AMÉRICA ..	38
5.2	DADOS DE HANSENÍASE NO ESTADO DO AMAZONAS NO BRASIL ....	47
5.3	AVALIAÇÃO DE DADOS SIMULADOS .....	56
5.3.1	FUNÇÃO $\mathcal{F}$ PONDERADA .....	63
<b>6</b>	<b>CONSIDERAÇÕES FINAIS</b> .....	<b>72</b>
6.1	PROPOSTAS DE CONTINUIDADE .....	75
	<b>REFERÊNCIAS</b> .....	<b>76</b>

## 1 INTRODUÇÃO

O crescente desenvolvimento de tecnologias para análise de dados no espaço geográfico oferece possibilidades inovadoras de abordagem em diversas áreas aplicadas, em particular, nas áreas da saúde. Por exemplo, para os casos de epidemiologia e vigilância sindrômica, e de suas tendências, nesse cenário é propiciada uma melhor compreensão dos fatores socio-econômicos e ambientais, entre outros, que determinam as condições de vida e o estado de saúde da população.

Existe um grande volume de estudos através do uso de *softwares* de geoprocessamento. Esses estudos tornam disponíveis uma grande quantidade de ferramentas de identificação de conglomerados espaço-temporais. Diversos programas de treinamento de funcionários da área de saúde pública utilizam esse ferramental. A pesquisa nessa área é realizada através dos registros obtidos pelas agências de saúde pública. As informações do monitoramento, em geral, contém o local e o instante de ocorrência dos casos de doenças de interesse geral.

A referida instrumentação é bastante importante no Brasil, país em que acentuadas desigualdades se evidenciam, não apenas entre as grandes regiões, os estados e os municípios, mas também no interior dessas áreas, sobretudo nos espaços urbanos. Sobrepondo-se às situações diferenciadas que caracterizam o território brasileiro, as transformações demográficas observadas nas últimas décadas potencializam a dinâmica de distribuição de doenças e agravos, bem como de demanda e acesso aos serviços.

A análise de saúde pública baseia-se na noção estatística de que cada pessoa tem uma determinada probabilidade (risco) de contrair uma determinada doença em estudo. O objetivo envolve identificar e quantificar quaisquer exposições, comportamentos e características que possam modificar o risco para o indivíduo. A incerteza para mensurar esse risco motiva o uso de métodos estatísticos, entre eles os métodos de Estatística espacial.

Ao longo do texto será utilizado o termo *cluster* que consiste de alguma coleção (conglomerado) de eventos que não tenha ocorrido de forma meramente casual, e ainda, cuja identificação é foco de pesquisas dentro da Estatística espacial. Esses conglomerados podem ser causados por diferentes fatores, tais como: agentes infecciosos; contaminação ambiental localizada; efeitos colaterais de tratamentos. Cada problema desses, com suas respectivas peculiaridades e particularidades, pode ser um dos efeitos relevantes da formação de um possível *cluster*.

Atualmente, existem diversas técnicas para análise de *clusters* com o intuito de buscar grupos de objetos semelhantes e dentro do contexto de *clusters* espaciais. O interesse está na detecção de um conjunto composto por uma ou mais regiões contíguas no mapa em estudo, regiões estas que contenham alguma característica comum. Neste texto, uma coleção de regiões, candidata potencial a *cluster*, será tratada por zona. Na epidemiologia, por exemplo, é possível pensar em um conjunto de municípios que apresente a incidência de uma determinada doença, significativamente maior do que no restante de um estado sob investigação.

Dentre os procedimentos de detecção de conglomerados, tem sido amplamente utilizada a técnica de Kulldorff (1997), denominada Estatística espacial Scan, baseada na Estatística Scan de Naus (1965). Entre suas características, a capacidade de delinear regiões críticas por meio de algoritmos computacionais gráficos, atribuindo significância a estatística associada através de simulações de Monte Carlo é relevante. Essa metodologia pode ser aplicada a dados espaciais, temporais ou espaço-temporais, para várias propostas de modelos de probabilidade quando considerada a distribuição dos casos do fenômeno de interesse.

Uma extensão importante da Estatística espacial Scan consiste na inclusão de uma avaliação temporal. Uma discussão detalhada acerca do estudo espaço-temporal pode ser obtida no trabalho de Tango (2010). Uma estatística de varredura utilizada no espaço-tempo é proposta por Kulldorff et al. (1998), definida com uma janela de varredura através de cilindros com base espacial circular e altura no intervalo de tempo

em análise.

Há duas estatísticas de verificação do espaço-tempo, uma prospectiva e outra retrospectiva. Para a Estatística Scan prospectiva, as análises são realizadas repetidamente em intervalos regulares, e a janela de busca é limitada, de modo que o ponto de extremidade temporal do cilindro é fixado no momento atual. É uma técnica capaz de detectar *clusters* que ainda estão ocorrendo (*clusters* vivos). Por essa razão, a Estatística Scan prospectiva é usada em sistemas de vigilância para detecção rápida de surtos da doença. Já a Estatística Scan retrospectiva relaxa a restrição sobre o ponto de extremidade temporal no cilindro, ou seja, realiza a procura por um *cluster* que ocorre ou ocorreu em qualquer momento durante o período de estudo.

As escolhas, por uma ou outra abordagem, são baseadas no interesse por avaliação de *clusters* atuais para a análise prospectiva e/ou *clusters* históricos para a análise retrospectiva. Para a abordagem espaço-tempo, a detecção de *cluster* consiste na detecção de forma rápida das mudanças nos padrões espacial e/ou temporal de eventos de interesse. Nesse trabalho, serão abordados prioritariamente sistemas de vigilância prospectivos espaço-temporais.

## 1.1 FORMATAÇÃO DE OBJETIVOS

Verifica-se na literatura uma escassez de métodos de detecção retrospectiva e prospectiva para *clusters* espaço-tempo que sejam capazes de combinar eficiência e rapidez no procedimento de identificação dos clusters. Obviamente o interesse pela detecção de clusters emergentes (detecção prospectiva) tende a parecer mais efetivo. Em outras palavras, o intuito maior parece focado em constatar, o mais rápido possível, os problemas ainda em curso. Entretanto, as técnicas eficientes tendem a se comportar bem nas duas abordagens (prospectiva e retrospectiva).

Não necessariamente a efetiva detecção é o objetivo mais relevante, a possibilidade de delinear as regiões mais e menos relevantes para o processo de formação de um *cluster*, assim como determinar os instantes de tempo mais e menos relevantes

nesse processo de formação da aglomeração espaço-temporal. Esse delineamento já se encontra detalhado para problemas puramente espaciais, como proposto por Oliveira et al. (2011). Será proposta aqui uma abordagem espaço-temporal para a utilização da função Intensidade.

O interesse efetivo está em mensurar a intensidade de cilindros espaço-temporais, e ainda em uma análise para o grau de incerteza no procedimento de detecção de clusters espaço-temporais. Em junção com esse mecanismo, a função Intensidade será utilizada para a determinação da relevância de possíveis instantes para o alarme temporal, com o interesse em aferir o momento que a possibilidade desse alarme é mais intensa. O mecanismo de alarme trabalha em conjunto com a análise da incerteza nas regiões limítrofes dentro e fora do *cluster* candidato. Esse conjunto pode tornar a avaliação para *clusters* emergentes mais abrangente que as técnicas usuais de detecção para *clusters* espaço-temporais emergentes.

Alguns métodos já estabelecidos para procedimentos de detecção em *clusters* espaço-temporais serão discutidos e comparados com a nova técnica proposta. Ressalta-se que a nova técnica não é um mecanismo de detecção, mas sim de delineamento. Em virtude desse fato, uma nova forma de análise comparativa será proposta e implementada; um dos principais interesses está em permitir quantificar a qualidade da estratégia de delineamento que será apresentada.

Este trabalho estará disposto da seguinte forma: inicialmente uma revisão bibliográfica abordando alguns métodos de detecção de *clusters* espaço-tempo será apresentada no capítulo 2. Posteriormente serão abordados os métodos de detecção de *clusters* espaço-tempo através da Estatística Scan espaço-tempo e através dos métodos baseados na Razão de verossimilhança adaptativa no capítulo 3. Em seguida, a construção da nova proposta de função Intensidade espaço-temporal com interesse em mecanizar a estratégia de visualização de áreas e instantes temporais de maior relevância na constituição de possíveis *clusters* espaço-temporais será discutida no capítulo 4. Um estudo através de dados reais para câncer de tireoide no estado do Novo

México/EUA e também com dados reais de ocorrências de Hanseníase no estado do Amazonas/BRA serão utilizados para verificação das respostas fornecidas pela nova técnica, e ainda um estudo comparativo baseado em dados simulados será mostrado no capítulo 5. Por fim, conclusões, considerações e propostas de continuidade são discutidas no capítulo 6.

## 2 REFERÊNCIAL TEÓRICO

O conjunto metodológico referente às técnicas para detecção e inferência em *clusters* espaço-temporais contempla diversas abordagens. As técnicas já têm sido estudadas há algum tempo, e em virtude disso, muitos resultados podem ser encontrados na literatura. Por outro lado, esse grande volume de resultados não leva a um desinteresse pela área; a aplicabilidade e o relevante número de pesquisadores envolvidos mostra que esse tema de pesquisa ainda é bastante atual e de suma importância.

No estudo das técnicas para detecção de *clusters* espaço-temporais se faz necessária a fixação de um período de tempo (discretização no tempo) com o intuito de agregar os casos ocorridos dentro de cada unidade de tempo. Escolhas inadequadas podem acarretar problemas significativos. Ao incluir poucos períodos, o teste pode não ter poder suficiente para detectar alguma doença de risco baixo ou moderado que ocorre há um tempo considerável. Por outro lado, o teste também pode não ter poder suficiente para detectar uma doença de risco elevado com ocorrências em um período curto.

É importante destacar a existência de um conflito significativo na relação entre espaço e tempo no processo de detecção. Ao minimizar o atraso temporal no procedimento de detecção, os métodos tendem a perder precisão na identificação espacial. Já ao apurar a precisão espacial no processo de detecção, ocorre uma tendência de atraso na detecção temporal.

O intuito central é detectar padrões anômalos que não sejam visuais, ou seja, que não seriam notados facilmente sem qualquer metodologia mais sofisticada. Se faz necessária uma suficiente sensibilidade ao método para identificar esses padrões, sem que isso acarrete na redução da possibilidade de que meras flutuações estocásticas sejam interpretadas como efetivos padrões de anomalia. De outra forma, os métodos devem apresentar capacidade de discernir diferenças entre sinal e ruído.

Uma das primeiras abordagens para detecção de *clusters* espaço-temporais foi uma abordagem retrospectiva proposta por Knox e Bartlett (1964), em que foram estabelecidas distâncias de corte para o espaço e para o tempo. As aproximações entre pares de observações para cada uma das dimensões são analisadas. É obtida uma tabela  $2 \times 2$  na qual os totais marginais possibilitaram o cálculo do valor esperado para cada uma das caselas, e em especial para a casela contendo observações próximas em relação ao espaço e ao tempo. Esse valor esperado era considerado como parâmetro para uma distribuição Poisson, assim viabilizando a obtenção da probabilidade de se observar um número de casos igual ou superior ao observado nos dados em estudo. Essa probabilidade era utilizada com o intuito de mensurar a significância estatística.

Com um foco semelhante, Mantel (1967) define um formato para as relações de aproximação entre pares de observações, além de uma outra forma para avaliar a significância estatística. Tanto para Knox e Bartlett (1964) quanto para Mantel (1967), os métodos são construídos sob o princípio de que, se existe efetivamente um *cluster*, as observações estarão próximas tanto no espaço quanto no tempo. Analogamente se as observações não são relacionadas, estas possuirão uma maior separação no tempo e no espaço.

Ao considerar apenas as medidas de distâncias entre observações pareadas, é possível verificar deficiências referentes às propostas anteriormente mencionadas. Isso pode ser verificado ao lidar com situações puramente espaciais ou puramente temporais, como observado por Whittemore et al. (1987).

Uma extensão para a Estatística Scan puramente espacial (KULLDORFF; NAGARWALLA, 1995; KULLDORFF, 1997) com uma avaliação espaço-temporal é abordada por Kulldorff et al. (1998). Posteriormente, Kulldorff (2001) utiliza a estatística espaço-temporal proposta anteriormente com o interesse em monitorar casos de doenças emergentes.

Ainda focado na vigilância prospectiva, Neill et al. (2005) apresentam uma abordagem Bayesiana com interesse em detectar *clusters* espaço-temporais de casos de doenças emergentes. O estudo é feito através da análise de séries temporais com interesse em determinar a quantidade de casos que se espera observar para alguma determinada região espacial e para algum intervalo de tempo pré-fixado. São propostas novas estatísticas para a verificação de “*clusters* emergentes” no espaço-tempo. Ainda é apresentado um critério de decisão para verificar se um aumento constatado para o número de casos observados em uma região é realmente significativo ou não.

Neill et al. (2005) verificam que a Estatística Scan espaço-temporal de Kulldorff et al. (1998) apresenta uma perda de poder. Segundo os autores, isso seria decorrente da utilização no seu formato de um valor esperado condicional com os dados do tempo atual, para identificar *clusters* formados com diversas localidades.

Outra proposta relevante é apresentada por Rogerson (2001), no qual é utilizada a estatística de Knox local (KNOX, 1963) combinada com métodos de somas cumulativas (CUSUM), para monitoramento de *clusters* pontuais, com o intuito de detectar algum conglomerado existente, através da visualização da posição de formação.

Sonesson (2007) utiliza como ferramenta a estatística CUSUM para detecção de *clusters* no espaço-tempo em dados de área. Outra forma de abordagem para este problema foi apresentada por Porter e Brown (2007) que utilizam versões adaptativas dos métodos CUSUM e Shiryaev-Roberts (SHIRYAEV, 1963; ROBERTS, 1966; KENETT; POLLAK, 1996) para detecção de mudanças espaço-temporais em processos de Poisson Homogêneos. Nesse mesmo senso, é possível mencionar o trabalho de Assunção e Corrêa (2009), no qual é proposto o monitoramento de uma função associada com um processo de Poisson Heterogêneo espaço-temporal através da estatística de Shiryaev-Roberts.

Neill (2009) propõe uma Estatística Scan espaço-tempo baseada em uma janela temporal limitada e ajustada conforme uma cota máxima para a altura do *cluster*.

Segundo o autor, a vantagem deste método é ser possível ajustar o valor esperado para o número de casos conforme possíveis tendências ou sazonalidades observadas na série, tendo então um maior controle sobre a estatística de teste. Considera-se o monitoramento simultâneo de um grande número de séries temporais espacialmente localizadas com interesse em detectar padrões espaciais emergentes. Propõe-se uma abordagem em duas etapas, inicialmente calcula-se a contagem esperada de casos para cada unidade de tempo recente para cada localização espacial; em seguida, encontram-se regiões espaciais, nas quais as contagens recentes são significativamente mais elevadas do que o esperado.

Outra possibilidade de formato para o monitoramento e a detecção de *clusters* espaço-temporais proposta por Lima e Duczmal (2014), define uma ampla classe de possíveis candidatos a *cluster* espacial e aplica os métodos de identificação de *clusters* temporais em cada um desses candidatos. Com isso, são propostos métodos adaptativos para detecção de *clusters* espaço-temporais. É apresentada uma metodologia baseada em razões de probabilidade adaptativa para a detecção de *clusters* emergentes. O limite superior para a taxa de falsos alarmes do método proposto depende apenas da quantidade de candidatos avaliados. Simulações de Monte Carlo não são necessárias tornando o custo computacional reduzido. O desempenho é avaliado por meio de simulações para mensurar o atraso de detecção média e a probabilidade de detecção do *cluster* correto.

Em conjunto com diversos métodos de detecção já mencionados, técnicas de delineamento e visualização são também preponderantes para o estudo que se desenvolverá. O conceito de delineamento de áreas de intensidade, em clusters puramente espaciais, tem sido discutido a algum tempo; entretanto tratam-se de discussões secundárias, abordando o assunto de forma marginal. O assunto é bem posto e de forma específica pela primeira vez por Oliveira et al. (2011).

A abordagem é feita através da utilização da função Intensidade e algumas discussões estratégicas sobre o assunto são mencionadas. Existe alguma incerteza

na estimativa para as taxas de ocorrência de fenômenos em estudo, especialmente para áreas pouco populosas. O delineamento dos limites de borda espacial na formação de possíveis *clusters* é sujeito a um efeito de variação, inerente às flutuações nas taxas de ocorrência nessas regiões fronteiriças. Suponha um candidato a *cluster* produzido por algum método dado, o que se sabe sobre as áreas externas, porém, limítrofes a esse candidato a *cluster*? Existe informação suficiente para garantir que tais áreas não possuem poder algum de influência no próprio *cluster*, ou até mesmo se pertencem ao verdadeiro *cluster* se esse existir?

Oliveira et al. (2012) discutem e apresentam resultados de aplicação da metodologia transcendendo a utilização apenas através da metodologia do Scan Circular. A utilização de métodos baseados em algoritmos genéticos (DUCZMAL et al., 2007; DUCZMAL; CANÇADO; TAKAHASHI, 2008; DUARTE et al., 2010; CANÇADO et al., 2010) mostra que a eficácia da análise continua consistente para outros mecanismos de detecção.

A função Intensidade é capaz de captar informações sobre a incerteza estatística inerente ao processo de identificação de *clusters*. Oliveira et al. (2015) discutem uma adaptação da função Intensidade para processos de Poisson. Essa adaptação tem interesse em fornecer uma maneira mais eficiente e precisa para determinar e visualizar o grau de incerteza correspondente de cada região específica pertencente a um possível verdadeiro *cluster*.

Procedimentos para discussão do delineamento das regiões ao redor de um possível *cluster* verdadeiro, assim como a relevância por importância das próprias regiões pertencentes ao possível *cluster* são na prática discussões a respeito do procedimento inferencial utilizado. Discussões profundas sobre o procedimento de inferência são apresentadas por Almeida et al. (2010) para o caso puramente espacial e são mencionadas possibilidades para o caso espaço-temporal.

Uma técnica de visualização dos resultados obtidos através da Estatística

Scan, proposta por Boscoe et al. (2003), amplia o conteúdo informativo das saídas. A razão de verossimilhança e o risco relativo são considerados simultaneamente, é possível identificar sub-blocos de maior (ou menor) risco relativo entre excessos ou deficit regionais.

Yasui et al. (2000) discutem uma estimativa empírica bayesiana para pequenos riscos de doenças. É proposta uma distribuição a priori sobre os riscos específicos a cada zona. O critério para a escolha de uma priori sobre outras, em geral, segue um raciocínio lógico. Os autores avaliam empiricamente várias priores utilizadas na estimativa bayesiana para pequenos riscos. As estimativas obtidas apresentam erro quadrático substancialmente menor que estimativas baseadas em verossimilhança Poisson.

Em particular, a estatística Scan espaço-temporal e a função Intensidade serão abordadas com maior profundidade e serão de extrema importância para a construção das propostas discutidas nesse texto.

### 3 DETECÇÃO DE CLUSTERS-ESPAÇO-TEMPO

A extensão da Estatística Scan de Kulldorff (1997) do espaço para o espaço-tempo ocorreu através da ampliação da estatística de varredura com formato circular para uma estatística de varredura agora com formato cilíndrico. A base circular na estratégia espacial corresponde à dimensão geográfica e à altura do cilindro corresponde ao intervalo de tempo sob investigação. A fim de detectar *clusters* espaço-temporais, Kulldorff et al. (1998) utilizam uma versão estendida da estatística Scan espacial (KULLDORFF, 1997) propondo a estatística espaço-tempo, em que o interesse é a detecção de *cluster* em um determinado local espacial e em algum intervalo de tempo específico.

#### 3.1 ESTATÍSTICA SCAN ESPAÇO-TEMPO

Para definir formalmente o procedimento de detecção de *clusters* espaço-temporais considere um mapa dividido em  $L$  áreas, com  $n$  instantes de tempo em investigação. Seja  $c_{t,\ell}$ ,  $\ell \in \{1, 2, \dots, L\}$  e  $t \in \{1, 2, \dots, n\}$  o número casos observados de certa anomalia no tempo  $t$  na área  $\ell$  do mapa original em estudo. Da mesma forma que a proposta para a identificação de *clusters* espaciais, sob  $H_0$  o número de casos em cada área e instante de tempo segue uma distribuição Poisson, agora com risco relativo constante no espaço e no intervalo de tempo e sob  $H_a$  o risco relativo é distinto em pelo menos uma região e algum intervalo temporal na parte exterior do cilindro sob investigação.

O uso do risco relativo se deve ao fato de ser possível fazer comparações entre dois mapas de uma mesma região, sendo os mapas de doenças diferentes e possuindo incidências muito diferentes. Até mesmo mapas referentes a uma mesma doença mas com dados de períodos de tempos muito distantes se tornam comparáveis. Para o risco relativo como uma razão (de mortalidade, morbidade, de casos, etc) padronizada, considere um mapa em estudo e selecione uma região  $z$  independentemente dos dados observados, o estimador natural do risco relativo na região  $z$  é dado

pela razão:

$$RR(z) = \frac{I(z)}{O(z)} = \frac{\left(\frac{N_z}{\mathbb{E}(c_z)}\right)}{\left(\frac{C - N_z}{C - \mathbb{E}(c_z)}\right)}$$

em que  $C$  é o número total de casos,  $N_z$  é o número de casos observados na região  $z$  e  $\mathbb{E}(c_z)$  é o número esperado de casos na área  $z$  sob a hipótese nula.

Este é um estimador imparcial quando a região  $z$  é escolhida independentemente dos dados observados. Assim, claramente se o risco estimado nas regiões  $z$ ,  $I(z) = \frac{N_z}{\mathbb{E}(c_z)}$ , está perto do risco estimado fora da região  $z$ ,  $O(z) = \frac{C - N_z}{C - \mathbb{E}(c_z)}$ , ou seja, se a razão entre eles está suficientemente próxima de 1 existe uma forte evidência da não existência de *cluster* no mapa em estudo. Se o aumento de  $I(z)$  com respeito a  $O(z)$  é representativamente grande, indica-se a possível existência de um *cluster* na região  $z$ .

Denote por  $S$  uma região em estudo particionada em  $L$  áreas  $A_\ell$  contíguas, com  $\ell \in 1, \dots, L$ , identificadas por um ponto  $s_\ell$ . Em que, por exemplo,  $S$  é um mapa e  $s_\ell$  representa um ponto arbitrário no interior do polígono que delinea uma área  $A_\ell$ , esse ponto é usualmente denotado por centróide da área  $A_\ell$ .

Considere sobre  $S$  o monitoramento estatístico de um processo estocástico  $\mathcal{X} = \{X_t(s_\ell), t = 1, 2, \dots \text{ e } \ell = 1, 2, \dots, L\}$ . A cada tempo discreto  $t \geq 1$  observa-se um vetor  $L$ -variado  $X_t = (X_t(s_1), X_t(s_2), \dots, X_t(s_L))'$ , em que  $X_t(s_\ell)$  representa algum valor de interesse, que tenha sido observado na  $\ell$ -ésima área. Seja ainda,  $\xi$  algum subconjunto conexo de localizações espaciais em  $S$  e denomine  $\xi$  por alguma zona pertencente ao mapa  $S$  em estudo.

Existem diversas estratégias para a construção das zonas  $\xi$  que serão monitoradas. Em particular, o interesse aqui está em discutir procedimentos mais simples como os utilizados no Scan Circular puramente espacial (KULLDORFF; NAGARWALLA, 1995). Por outro lado, a definição que será apresen-

tada é bastante abrangente podendo incluir diversos outros procedimentos de construção das zonas sob investigação.

Seja  $\mathcal{D}_n = \{X_1, \dots, X_n\}$  o conjunto de dados acumulados até o instante  $n$  para o valor de interesse, considere o conhecimento das populações nas  $L$  áreas em estudo e  $n$  instantes em investigação, dados por  $p_{t,\ell}; t \in \{1, \dots, n\}$  e  $\ell \in \{1, \dots, L\}$  e em particular seja  $X_t(s_\ell)$  o número de casos ocorridos para o fenômeno de interesse (por exemplo, casos de dengue, casos de homicídios, ou outro qualquer) na área  $\ell$  no tempo  $t$ .

Busca-se responder a seguinte questão: Os casos estão distribuídos de forma aleatória nessas áreas nesse intervalo de tempo? Existe alguma região do mapa nesse intervalo de tempo que apresenta valores discrepantes das demais áreas de estudo? Em outras palavras, objetiva-se testar então:

$$\begin{cases} H_0 : \text{Não existe } cluster \text{ no mapa no intervalo de tempo entre } 1 \text{ e } n \\ H_a : \text{Existe algum } cluster \text{ no mapa no intervalo de tempo entre } 1 \text{ e } n \end{cases}$$

O objetivo do monitoramento é verificar se existe algum *cluster* em  $\mathcal{S}$  e se existir, detectar o *cluster* que tenha se iniciado em um tempo desconhecido  $k \leq n$ . Um *cluster* no espaço-tempo é expresso em função de uma mudança no padrão da distribuição do sub-processo  $\{X_t(s_\ell) \in \mathcal{X} : s_\ell \in \xi; t \geq k\}$  tal que  $\xi^{k_a, k_b} = \{s_\ell \in \mathcal{S} : s_\ell \in \xi\} \times [k_a, k_b]$  representa um *cluster* no espaço-tempo.

Em analogia com a proposta de Kulldorff (1997) para detecção de *clusters* espaciais, considere o intervalo de tempo  $[k_a, k_b]; k_a \geq 1$  e  $k_b \leq n$ . A estatística espaço-temporal será construída a partir da razão de verossimilhanças  $LR(\xi^{k_a, k_b})$  dada da seguinte forma:

$$LR(\xi^{k_a, k_b}) = \begin{cases} I(\xi^{k_a, k_b})^{c(\xi^{k_a, k_b})} O(\xi^{k_a, k_b})^{C^{k_a, k_b} - c(\xi^{k_a, k_b})}, & \text{se } I(\xi^{k_a, k_b}) > 1; \\ 1, & \text{caso contrário.} \end{cases}$$

em que  $c(\xi^{k_a, k_b})$  representa a soma dos casos na dentro de uma zona  $\xi$  no intervalo de tempo  $[k_a, k_b]$ , ou seja, o número de casos no cilindro  $\xi^{k_a, k_b}$  que é definido pela expressão  $c(\xi^{k_a, k_b}) = \sum_{t=k_a}^{k_b} \sum_{\ell; A_\ell \in \xi} X_t(s_\ell)$ . Desse modo,  $C^{k_a, k_b} - c(\xi^{k_a, k_b})$  representa o total de casos ocorridos fora do cilindro  $\xi^{k_a, k_b}$  no intervalo de tempo  $[k_a, k_b]$ , ou seja, será representado por  $C^{k_a, k_b} - c(\xi^{k_a, k_b}) = \sum_{t=k_a}^{k_b} \sum_{\ell; A_\ell \notin \xi} X_t(s_\ell)$ .

A Figura 1 ilustra um possível cilindro em um mapa em estudo.

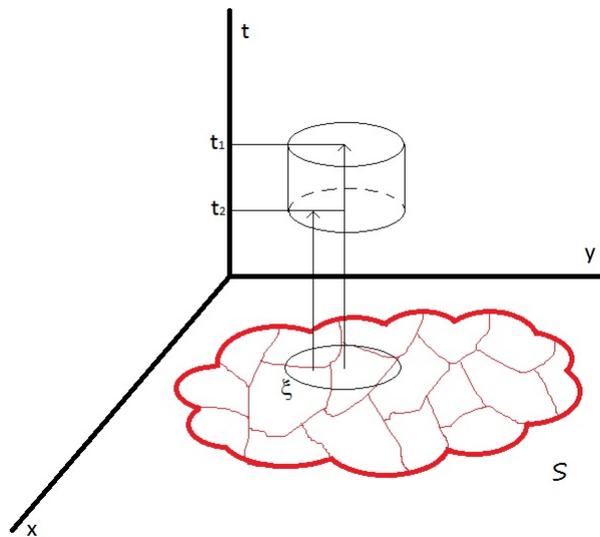


Figura 1: Exemplo hipotético de um cilindro  $\xi^{t_1, t_2}$  em um mapa  $S$ .

Os termos  $I(\xi^{k_a, k_b})$  e  $O(\xi^{k_a, k_b})$  representam o risco relativo dentro e fora, respectivamente, do cilindro formado pela zona  $\xi$  no intervalo de tempo  $[k_a, k_b]$ . O *cluster* mais verossímil, denotado por  $\hat{\xi}^{k_a^*, k_b^*}$ , será algum cilindro com  $k_a^* \geq 1$  e  $k_b^* \leq n$ , tal que  $LR(\hat{\xi}^{k_a^*, k_b^*})$  é máximo da função  $LR(\cdot)$ . Defina  $\mathcal{C}$  o conjunto de todos os cilindros a serem investigados, dessa forma,  $LR(\hat{\xi}^{k_a^*, k_b^*}) \geq L(\xi^{k_a, k_b}), \forall \xi$  entre as zonas em estudo em qualquer intervalo de tempo  $[k_a, k_b]$  em estudo. Fixadas, as zonas  $\xi$  e os intervalos de tempo  $[k_a, k_b]$ , o conjunto  $\mathcal{C}$  fica bem definido sendo seus elementos do tipo  $\xi^{k_a, k_b}$ , dessa forma, a estatística de teste fica definida por:

$$\Lambda = \max_{\xi^{k_a, k_b} \in \mathcal{C}} LR(\xi^{k_a, k_b})$$

e ainda,  $\hat{\xi}^{k_a^*, k_b^*}$  fica bem definido por  $\arg \max LR(\cdot)$ , restrito aos elementos do conjunto  $\mathcal{C}$ .

Com foco em reduzir o atraso na detecção de possíveis *clusters* espaço-temporais, ou seja, a identificação de um *cluster* no menor tempo possível, (KULLDORFF, 2001) discute a estatística Scan espaço-temporal para a detecção de *clusters* emergentes, ou seja, detectando locais que apresentam elevados fatores de risco tão previamente a fim de detectar concentrações para doenças desconhecidas ou doenças conhecidas em locais desconhecidos.

No propósito de detectar *clusters* emergentes, a proposta de Kulldorff (2001) consiste em utilizar a estatística proposta por Kulldorff et al. (1998) com foco apenas nos *clusters* que ainda existam no último período de tempo disponível para a investigação. Logo, busca-se identificar o *cluster* logo após seu surgimento através de uma varredura com a Estatística Scan espaço-temporal. Em outras palavras, o intervalo  $[k_a, k_b]$  a ser investigado é um intervalo do tipo  $[k, n]$  com  $k \leq n$  para o intervalo completo de investigação definido por  $[1, n]$ .

Se os possíveis intervalos  $[k_a, k_b]$  são quaisquer intervalos contidos em  $[1, n]$ , a investigação é dita retrospectiva. Por outro lado, se os intervalos a serem investigados sempre possuem seu limite superior em  $n$  a busca em questão é denominada prospectiva.

Conforme novos casos são adicionados, seguindo a entrada de novas informações ao longo do andamento temporal, é possível repetir o procedimento, porém é necessário realizar um ajuste no grau de significância estatística calculado, de tal forma a permitir a comparabilidade entre os *clusters*. Para isso deve-se realizar a maximização da verossimilhança sob  $H_0$  levando em conta todos os cilindros utilizados nas análises anteriores e na presente.

Supondo o conhecimento para a distribuição de probabilidade da estatística de teste  $\Lambda$  sob a validade da hipótese nula de não existência de *cluster*, um valor crítico

$crit_\alpha$  poderia ser escolhido para um nível de significância pré-estabelecido  $\alpha$ . Se o valor  $\Lambda$  efetivamente verificado no conjunto de dados observados foi obtido através do cilindro  $\hat{\xi}^{k_a^*, k_b^*}$  e é tal que  $\Lambda > crit_\alpha$  então o conglomerado definido pelo cilindro  $\hat{\xi}^{k_a^*, k_b^*}$  é significativo ao nível  $\alpha$  ou seja,  $\hat{\xi}^{k_a^*, k_b^*}$  poderá ter seu status alterado de conglomerado candidato para *cluster* detectado. De uma outra forma, essa constatação significa dizer que o valor da estatística de teste é suficientemente discrepante a ponto de ser estatisticamente significativa a discrepância do volume de ocorrências no cilindro  $\hat{\xi}^{k_a^*, k_b^*}$  em relação ao restante dos cilindros em estudo.

Entretanto, o não conhecimento da distribuição de probabilidade da estatística de teste  $\Lambda$  sob a validade da hipótese nula, pode ser visto como um entrave para busca através da Estatística Scan espaço-temporal. Para amenizar este problema, a distribuição de probabilidades para a estatística de teste pode ser estimada através de uma distribuição empírica obtida por procedimentos de Monte Carlo (DWASS, 1957).

Para os conjuntos aleatórios formados, casos são simulados de forma que espaço e tempo sejam independentes e a verossimilhança desses conjuntos é maximizada sob todos os cilindros construídos a partir do tempo de início da vigilância. Diversas repetições desse procedimento são capazes de gerar um volume de pontos necessários para estimar de forma eficaz a distribuição para a estatística de teste. Para o caso puramente espacial existem estudos comprovando boa qualidade de ajuste através de uma aproximação com a distribuição Gumbel (ABRAMS; KULLDORFF; KLEINMAN, 2010). Para o casos espaço-temporal, não existem estudos definitivos, mas é minimamente razoável manter a suposição de aproximação Gumbel. De posse da distribuição, seja ela obtida por métodos de Monte Carlo, ou por aproximação Gumbel, dada uma solução  $\hat{\xi}^{k_a^*, k_b^*}$  significativa ao nível  $\alpha$  é possível obter estimativas para o seu  $p$ -valor.

### 3.2 FUNÇÃO INTENSIDADE

A função Intensidade (OLIVEIRA et al., 2011) é proposta sobre o aspecto de análises puramente espaciais. Os métodos de detecção de clusters espaciais, em geral, classificam uma determinada região do mapa em duas modalidades: (1) pertence ao *cluster*; (2) não pertence ao *cluster*. Partindo do pressuposto de buscar identificar a plausibilidade de que uma determinada região de um mapa esteja sob a influência de uma anomalia de interesse, o objetivo sobrepõe a ideia básica de simplesmente identificar a existência um *cluster*. Trata-se de uma importante ferramenta de visualização, para auxiliar profissionais de diversas áreas na tomada de decisões.

Oliveira et al. (2015) redefinem de forma mais simples a função Intensidade, renomeando a mesma para função  $\mathcal{F}$ . Considere novamente um mapa dividido em  $L$  áreas, mas agora sem a distinção temporal. Seja  $c_\ell, \ell \in \{1, 2, \dots, L\}$  o número casos observados de certa anomalia na área  $\ell$  do mapa original em estudo. São executadas simulações de Monte Carlo, o número de casos simulados segue uma distribuição Poisson com a utilização da informação dos casos observados em cada uma das áreas. Seja  $c_\ell$  o número de casos observados na área  $\ell$ , o processo de simulação será tal que, para cada região, o número de casos seguirá uma distribuição  $\text{Poisson}(c_\ell)$ . Para cada simulação de Monte Carlo, é utilizado algum algoritmo de detecção de *clusters*, como por exemplo, o Scan Espacial Circular.

O interesse é obter a proporção de vezes que cada uma das  $L$  regiões apareceu em alguma das soluções mais verossímeis obtidas. Seja  $\mathcal{F}_\ell$  a proporção de vezes que a área  $\ell$  pertence a alguma das soluções mais verossímeis dentre todas as novas simulações de Monte Carlo executadas,  $\mathcal{F}_\ell$  é dita função  $\mathcal{F}$  da área  $\ell$ , é possível descrever a função  $\mathcal{F}$  por:

$$\mathcal{F}_\ell = \frac{\sum_{q=1}^Q \mathbb{1}_{\{\ell; \ell \in \hat{\xi}_q\}}}{Q}$$

em que  $q \in \{1, \dots, Q\}$  representa a simulação corrente para um total de  $Q$  simulações de Monte Carlo,  $\hat{\xi}_q$  é a zona mais verossímil na  $q$ -ésima execução,  $\mathbb{1}_A$  é função indicadora para o conjunto  $A$ .

## 4 VISUALIZAÇÃO DA INCERTEZA PARA CLUSTERS ESPAÇO-TEMPORAIS

O interesse dos métodos de detecção espaço-temporais está em constatar a existência de padrões de anomalia não visuais (ou seja, que não seriam notados facilmente sem qualquer metodologia) no menor tempo possível. É necessária sensibilidade suficiente ao método para identificar esses padrões, entretanto com a capacidade de reduzir a possibilidade de que meras flutuações estocásticas sejam interpretadas como efetivos padrões.

Para utilizar essas técnicas, realiza-se um monitoramento estatístico de um processo estocástico  $\{X_t; t = 1, 2, \dots\}$  com o objetivo de detectar uma mudança relevante no processo, em algum tempo desconhecido  $k$ , tão rápida e precisa quanto possível. O monitoramento que será proposto aqui se baseia na reformulação da função  $\mathcal{F}$  para se ajustar ao problema espaço-temporal em conjunto com uma técnica de alarme para a constatação do instante de detecção de mudança no processo temporal.

### 4.1 FUNÇÃO $\mathcal{F}$ ESPAÇO-TEMPORAL

O procedimento para a definição da função  $\mathcal{F}$  espaço-temporal será análogo ao procedimento puramente espacial definido na seção 3.2. Novamente considere sobre  $S$  avaliar o monitoramento estatístico de um processo estocástico  $\mathcal{X} = \{X_t(s_\ell), t = 1, 2, \dots \text{ e } \ell = 1, 2, \dots, L\}$ . A cada tempo discreto  $t \geq 1$  observa-se um vetor  $L$ -variado  $X_t = (X_t(s_1), X_t(s_2), \dots, X_t(s_L))'$ , em que  $X_t(s_\ell)$  representa algum valor de interesse, que tenha sido observado na  $\ell$ -ésima área. Sejam ainda,  $\xi$  alguma zona pertencente ao mapa  $S$  em estudo.

Seja  $\mathcal{D}_n = \{X_1, \dots, X_n\}$  o conjunto de dados acumulados até o instante  $n$  para o valor de interesse, considere o conhecimento das populações nas  $L$  áreas em estudo e  $n$  instantes em investigação, dados por  $p_{t,\ell}; t \in \{1, \dots, n\}$  e  $\ell \in \{1, \dots, L\}$  e em particular seja  $X_t(s_\ell)$  o número de casos ocorridos para o fenômeno de interesse na área  $\ell$  no tempo  $t$ . Como definido anteriormente, um *cluster* no espaço-tempo

é expresso em função de uma mudança no padrão da distribuição do sub-processo  $\{X_t(s_\ell) \in \mathcal{X} : s_\ell \in \xi; t \geq k\}$  tal que  $\xi^{k_a, k_b} = \{s_\ell \in \mathcal{S} : s_\ell \in \xi\} \times [k_a, k_b]$  representa o *cluster* no espaço-tempo. Por uma questão prática dos interesses dessa pesquisa, as definições serão restritas a um procedimento prospectivo, mas nada impede que sejam estendidas a um procedimento retrospectivo, modificando apenas o conjunto dos cilindros espaço-temporais sob investigação.

São executadas simulações de Monte Carlo. A distribuição Poisson é proposta com a utilização da informação dos casos observados em cada uma das áreas e em cada tempo sob investigação. Seja  $c_{t,\ell}$  o número de casos observados na área  $\ell$  no tempo  $t$ , o processo de simulação será tal que, para cada região e cada tempo, o número de casos seguirá uma distribuição  $\text{Poisson}(c_{t,\ell})$ . Para cada simulação de Monte Carlo, é utilizado o algoritmo de detecção de clusters escolhido, como por exemplo, o Scan Circular espaço-temporal prospectivo. Considere o conjunto das soluções obtidas pelo algoritmo utilizado. Seja  $\mathcal{F}_{t,\ell}$  a proporção de vezes que dentre as soluções mais verossímeis foi obtido algum cilindro  $\xi \times [k, n]$  tal que a área  $\ell$  pertence à zona  $\xi$ , e ainda com  $k \leq t \leq n$ , isso dentre todas as simulações de Monte Carlo executadas. Para cada área  $\ell$ ,  $\mathcal{F}_{t,\ell}$  é dita função  $\mathcal{F}$  da área espacial  $\ell$  no tempo  $t$ . É possível descrever a função  $\mathcal{F}$  espaço-temporal por:

$$\mathcal{F}_{t,\ell} = \frac{\sum_{q=1}^Q \mathbb{1}_{\{\ell; \ell \in \hat{\xi}_q^{k,n}\} \cap \{t; t \geq k\}}}{Q}$$

em que  $q \in \{1, \dots, Q\}$  representa a simulação corrente para um total de  $Q$  simulações de Monte Carlo,  $\hat{\xi}_q^{k,n}$  é o cilindro mais verossímil na  $q$ -ésima execução com base na zona espacial  $\hat{\xi}$  e altura no intervalo temporal  $[k, n]$ ,  $\mathbb{1}_A$  é função indicadora para o conjunto  $A$ .

## 4.2 ALARME PARA MUDANÇAS NO PROCESSO

Seja  $x_1, x_2, \dots$  uma sequência de observações do processo em monitoramento com densidade (ou função de probabilidade)  $f(x_t, x_{t-1}, \dots, x_1/\theta)$ . A cada tempo  $t$  o processo pode estar em dois possíveis estados, (1) fora-de-controle ou (2) sob-controle. Através da avaliação de todos os dados passados, é possível identificar o primeiro ponto no tempo para qual existem fortes evidências de uma mudança no processo. O monitoramento é feito através de uma função alarme  $\Psi_t = \Psi_t(x_1, \dots, x_t)$  e um limite de controle  $B$  é tal que o tempo de alarme  $\tau$  de uma possível mudança é dado por  $\tau = \min\{t : \Psi_t > B\}$ .

Considere que se  $\Psi_t$  excede  $B$ , então, existem indicações de que o processo está fora de controle, e um alarme é acionado. É importante observar que existe um conflito entre espaço e tempo no processo de detecção. Ao minimizar o atraso temporal na detecção os métodos tendem a perder precisão na identificação espacial. Por outro lado, ao apurar a precisão espacial ocorre uma tendência de atraso na detecção temporal.

Defina  $\tau$ , um instante desconhecido em que ocorre uma mudança no processo estocástico  $X(t)$ . O processo é dito *sob controle* em um tempo qualquer  $t$  se  $t < \tau$ . Quando  $t > \tau$ , entende-se que o processo está *fora de controle* em  $t$ . Se não houver mudança no processo, então  $\tau = \infty$ . Se o processo está efetivamente sob-controle, todos os alarmes são falsos. O valor de  $B$  é geralmente especificado de forma a minimizar a ocorrência de alarmes falsos.

Em situações de monitoramento, são realizadas análises repetidas sobre o estado do processo e como o processo pode eventualmente mudar durante o monitoramento, o nível de significância e poder do teste não são medidas apropriadas para a avaliação em decorrência de problemas de testes múltiplos. A fim de detectar *clusters* espaço-temporais existem várias propostas que utilizam artifícios de detecção de *clusters* temporais, boa parte oriundos de conceitos abordados dentro das ferramentas

de Controle da Qualidade que realizam análises repetidas de determinadas medidas sobre o estado do processo tais como:

- $ARL^0$  (*average in-control run length*): utilizada para medir a qualidade de um processo sob controle, através do número médio de observações até a mudança de status de sob-controle para fora-de-controle, levando em conta que sob hipótese nula o processo está sob-controle. Não havendo mudança de status sob hipótese nula, qualquer alarme soado será falso e sua distribuição é dada por:

$$ARL^0 = \mathbb{E}_0[\tau]$$

em que  $\mathbb{E}_0$  representa o valor esperado sobre a hipótese nula.

- $FAR$  (taxa de falsos alarmes) dada por:

$$FAR = \frac{1}{ARL^0} = \frac{1}{\mathbb{E}_0[\tau]}$$

Para altos valores de  $ARL^0$  ocorrem poucos alarmes falsos, ou seja a taxa de falsos alarmes será baixa.

- Atraso (*delay*): medida utilizada para medir a velocidade de detecção da mudança de status do processo podendo ser medida através do atraso esperado condicional (CED) sendo dado por:

$$CED(k) = \mathbb{E}_K[\tau - \nu \mid \tau \geq \nu = k]$$

que representa o tempo médio de atraso até soar o alarme dado que realmente ocorreu alteração no status do processo,  $\mathbb{E}_K$  é o valor esperado sobre a hipótese alternativa de existência de alteração no status do processo no ponto  $\nu = k$ .

Motivado pelos conceitos acima, Lima e Duczmal (2014) adaptaram os instrumentos utilizados para verificação do *status* do processo nas condições de detecção de *clusters* temporais para a condição espaço-temporal propondo a seguinte definição:

**Definição 4.2.1** *Seja  $\tau$  um tempo de alarme e  $\xi$  um conjunto desconhecido de áreas do mapa. Um procedimento para a detecção e localização do cluster espaço-temporal é representado por um par  $(\tau, \xi)$ , em que  $\{\tau = \tilde{\tau} < \infty, \xi = \tilde{\xi}\}$  representa o evento em que uma possível mudança ocorreu no tempo  $\tilde{\tau}$  e localizada em  $\tilde{\xi}$ .*

Na definição apresentada anteriormente é possível destacar a ocorrência de alguns erros ao identificar os *clusters* espaço-temporais tais como:

- **Atraso de Detecção:** Sendo o *cluster* detectado e localizado corretamente, ou seja,  $\hat{\xi} = \tilde{\xi}$ , entretanto, o verdadeiro instante de surgimento do *cluster* é o tempo  $\nu$ , mas  $\tilde{\tau} > \nu$  então o atraso na detecção é dado por  $|\tilde{\tau} - \nu|$ ;
- **Falso Alarme:** Ocorre quando o tempo de mudança detectado foi  $\tau < \nu$ , em que  $\nu$  é o instante de surgimento do *cluster* verdadeiro;
- **Falsa Localização:** O início da ocorrência do *cluster* foi acertado porém a região localizada  $\hat{\xi} \neq \tilde{\xi}$ .

### 4.3 ALARME ATRAVÉS DA FUNÇÃO $\mathcal{F}$

Como visto na seção anterior, o procedimento que soa o alarme no instante em que foi observada a efetiva mudança no processo temporal é dependente de uma quantidade  $\tau$  dada por  $\tau = \min\{t : \Psi_t > B\}$ . Em virtude dessa definição, a proposição de um limiar efetivo, porém eficiente  $B$  e uma função  $\Psi_t$  são caminhos para a construção de um processo de detecção de *clusters* emergentes. Entretanto, ao observar o problema através da perspectiva da função  $\mathcal{F}$ , não ocorre mais um procedimento de detecção de *clusters*, mas sim um procedimento para o delineamento de localizações espaço-temporais de maior e menor relevância na formação de um possível *cluster*.

A construção do procedimento de alarme através da função  $\mathcal{F}$  seguirá princípios semelhantes aos já discutidos para a função  $\mathcal{F}$  espaço-tempo com algumas alterações. Considere o procedimento já definido anteriormente para determinar a

função  $\mathcal{F}$  espaço-temporal. A cada execução do procedimento de Monte Carlo uma solução mais verossímil é considerada para efeito cálculo da função  $\mathcal{F}$ , o tempo inicial dessa solução é uma proposta para soar o alarme para cada uma das regiões pertencentes a essa solução candidata exatamente o tempo inicial do cilindro temporal dessa solução candidata.

Com isso, a construção do procedimento de alarme através da função  $\mathcal{F}$  seguirá uma formatação distinta, porém com o mesmo propósito. A cada execução de uma simulação de Monte Carlo para o cálculo da função  $\mathcal{F}$  associada a cada região e alocação temporal, uma solução mais verossímil é obtida. Suponha que a solução obtida seja determinada no intervalo temporal  $[k, n]$ . Sendo essa a verdadeira solução, o alarme deveria ter sido soado no tempo  $k$ , produzindo assim atraso nulo no procedimento de detecção. Para a referida solução, suponha que espacialmente a mesma seja representada pela zona  $\xi$ . Dessa forma, para todas as áreas pertencentes à zona espacial  $\xi$  o alarme deveria ter sido soado no tempo  $k$ .

Considerando diversas execuções de simulações de Monte Carlo, é possível, verificar a proporção de vezes que o alarme seria soado para cada uma das regiões do mapa em estudo e para cada instante de tempo sob investigação. Para  $Q$  realizações do procedimento de Monte Carlo, seja  $\mathcal{A}_{t,\ell}$  a proporção de vezes que o alarme soou para a área  $\ell$  no tempo  $t$ . A quantidade  $\mathcal{A}_{t,\ell}$  será dita alarme através da função  $\mathcal{F}$  na área  $\ell$  no instante  $t$ . É importante ressaltar que a técnica proposta para o alarme através da função  $\mathcal{F}$  não é efetivamente uma técnica para detecção de *clusters*, mas sim para delinear a intensidade que um possível alarme possa ser soado no espaço e no tempo, através da expressão:

$$\mathcal{A}_{t,\ell} = \frac{\sum_{q=1}^Q \mathbb{1}_{\{\ell, \ell \in \hat{\xi}_q^{k,n}\} \cap \{t=k\}}}{Q}$$

em que  $q \in \{1, \dots, Q\}$  representa a simulação corrente para um total de  $Q$  simulações

de Monte Carlo,  $\hat{\xi}_q^{k,n}$  é o cilindro mais verossímil na  $q$ -ésima execução com base na zona espacial  $\hat{\xi}$  e altura no intervalo temporal  $[k, n]$ ,  $\mathbb{1}_A$  é função indicadora para o conjunto  $A$ .

## 5 RESULTADOS EXPERIMENTAIS

As propostas discutidas nessa tese foram avaliadas através de dois bancos de dados reais. Inicialmente foi verificada a adequabilidade dos procedimentos, em um segundo momento a análise das respostas obtidas para o conjunto de dados reais foi comparada com procedimentos usuais já estabelecidos. Por fim, experimentos extensivos de simulações foram executados, com interesse em mensurar a efetividade das respostas obtidas pela metodologia proposta. Serão discutidos alguns aspectos iniciais sobre os bancos de dados utilizados para posteriormente a real utilização da metodologia ser abordada.

### 5.1 DADOS DE CÂNCER DE TIREOIDE - ESTADOS UNIDOS DA AMÉRICA

A tireoide é uma glândula localizada na parte frontal do pescoço, abaixo da caixa de voz (laringe) contendo duas partes, ou lóbulos. Ela tem por função produzir hormônios que regulam o metabolismo do corpo humano, ou seja controlam o uso e armazenamento de energia necessária ao corpo. O câncer de tireoide ocorre quando tumores, também conhecidos como nódulos, crescem na tireoide. Sabe-se que em sua grande maioria (cerca de 90%) esses nódulos que crescem na tireoide, são benignos (não cancerosos), mas aqueles que são cancerosos podem se espalhar por todo o corpo colocando a vida em risco.

De acordo com Cramer et al. (2010) e DeSantis et al. (2014) houve um grande crescimento no número de casos do câncer de tireoide nos Estados Unidos a partir dos anos 90. Essa moléstia se tornou o câncer em avaliação com maior crescimento de incidência de novos casos dentre todos os cânceres malignos, mesmo envolvendo os dois sexos. A causa exata do câncer de tireoide não é conhecida, mas indivíduos com certos fatores de risco são mais vulneráveis que outros à doença. Exemplos podem ser encontrados em indivíduos que fizeram tratamentos com radiação na cabeça, pescoço ou tórax, especialmente na infância ou adolescência, além de pessoas com histórico familiar dessa doença.



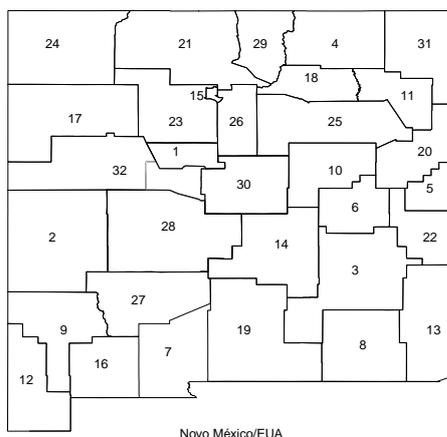


Figura 3: Mapa do Novo México/EUA.

As regiões codificadas numericamente na Figura 3 são nominadas conforme a Tabela 1.

Tabela 1: Sub-regiões do mapa do Novo México/EUA

código	região	código	região
1	Bernalillo	17	McKinley
2	Catron	18	Mora
3	Chaves	19	Otero
4	Colfax	20	Quay
5	Curry	21	Rio Arriba
6	De Baca	22	Roosevelt
7	Dona Ana	23	Sandval
8	Eddy	24	San Juan
9	Grant	25	San Miguel
10	Guadalupe	26	Santa Fé
11	Harding	27	Sierra
12	Hidalgo	28	Socorro
13	Lea	29	Taos
14	Lincoln	30	Torrance
15	Los Alamos	31	Union
16	Luna	32	Valencia

Existem alguns propósitos bem específicos nesse estudo, um deles diz respeito a mensurar a intensidade associada a cada localização espaço-temporal no mapa em estudo, outro está associado a produzir uma estratégia de visualização eficiente para a interpretação das informações.

A Figura 4 apresenta o efeito de intensidade medido através da função  $\mathcal{F}$  em cada área do mapa, associado à cada intervalo temporal. Uma explanação inicial é bastante importante para a adequada interpretação da Figura 4. O grau de intensidade associado à cada região é representado por sua respectiva coluna através de uma barra de cores que é mais escura a medida que a região é mais intensa, ou seja, com maior chance de pertencer a um possível *cluster* existente no mapa em estudo. As colunas associadas a cada uma das regiões são subdivididas ao longo dos 20 tempos nos quais está dividido o intervalo temporal em estudo. Os níveis de intensidade são representados como em um mapa de calor, mostrando a variação do menos intenso ao mais intenso. Dessa forma, um tom mais escuro na região  $X$  no tempo  $Y$  significa muita intensidade no cilindro com base na região  $X$  delimitado pelo intervalo temporal  $[Y, 20]$ . Todos os intervalos temporais são limitados superiormente pelo valor 20, isto se deve a natureza prospectiva desse estudo.

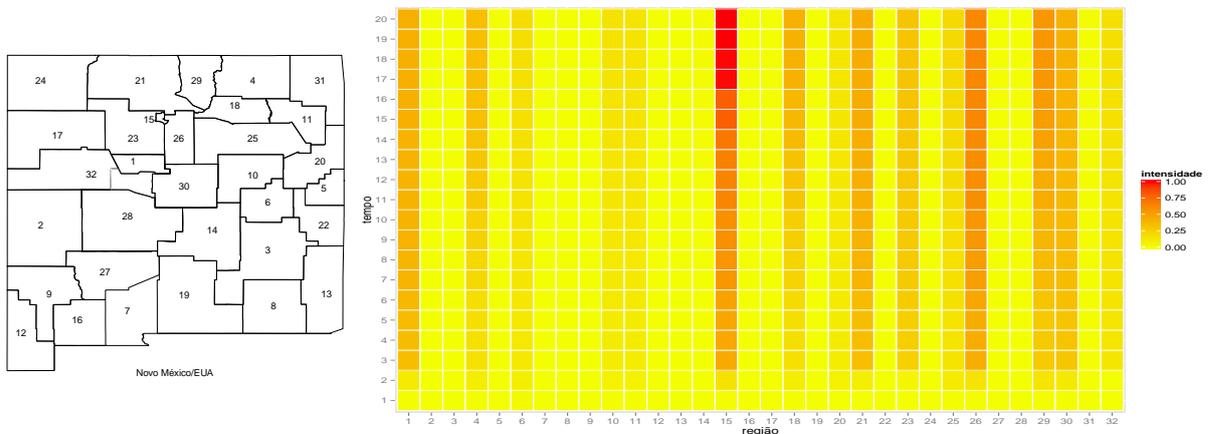


Figura 4: Função  $\mathcal{F}$  para câncer de tireoide no Novo México/EUA.

Os procedimentos computacionais foram realizados seguindo a definição dada no Capítulo 4. Foram executadas 1000 sequências distintas, cada uma dessas sequências composta por 10.000 simulações de Monte Carlo. Para cada tempo e espaço sob investigação, foi computada a média da função  $\mathcal{F}$  dentre as 1000 sequências.

É fácil ver uma elevada intensidade para a região 15 em todos os intervalos temporais, mas uma intensidade destacada para o cilindro definido pelos intervalos

temporais [17, 20] e [15, 20]. Esse resultado sugere uma atenção especial para o cilindro definido pelo intervalo temporal [15, 20], mas desde o tempo 3 a região 15 já se mostra bastante intensa. A sugestão de mais foco no intervalo [15, 20] se deve ao interesse em minimizar o atraso na constatação pelo surgimento de um possível *cluster*. Outras regiões como 1, 4, 18, 21, 23, 26, 29 e 30 apresentam também uma intensidade que se destaca em relação ao restante do mapa, mas claramente abaixo da intensidade na região 15.

A Figura 5 apresenta o efeito do alarme através da função  $\mathcal{F}$ , ainda associado a cada área do mapa em estudo. As colunas associadas a cada uma das regiões são subdivididas ao longo dos 20 tempos nos quais está dividido o intervalo temporal em estudo. Um tom mais escuro na região  $X$  no tempo  $Y$  agora significa muita intensidade para soar o alarme para a região  $X$  no instante de tempo  $Y$ .

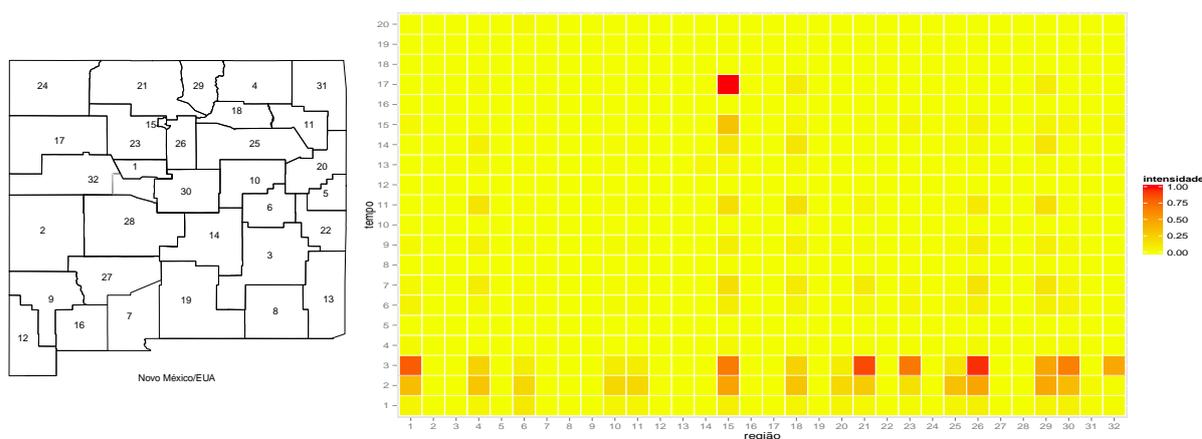


Figura 5: Alarme através da função  $\mathcal{F}$  para câncer de tireoide no Novo México/EUA.

A interpretação da Figura 5 somente tem sentido se avaliada em associação à Figura 4. Dado que a Figura 4 sugere uma altíssima intensidade na região 15, a Figura 5 sugere que a tendência maior é que o alarme soe no tempo 17, entretanto uma relevante intensidade no tempo 15 pode ser notada. O fato de muitas regiões apresentarem um sinal de intensidade relevante, porém menor que o da região 15, fica evidenciado pela intensidade presente para soar o alarme nos tempos 2 e 3 em diversas regiões, anteriormente observado na Figura 4.

A solução sugerida pela Estatística Scan espaço-temporal prospectiva coincide com o cilindro com base na região 15 no intervalo de tempo [17, 20]. Entretanto, a estratégia usual da Estatística Scan não permite delinear se vizinhanças desse cilindro são focos potenciais para se associarem a um possível *cluster*. Apenas um cilindro é apresentado como solução, isso se deve ao fato de apenas uma solução ter sido considerada significativa segundo o procedimento da Estatística Scan espaço-temporal prospectiva.

A Figura 6 sugere o que seria o efeito da Estatística Scan espaço-tempo prospectiva sob a ótica da relevância por regiões e instantes de tempo (solução significativa considerando um nível de significância de 5%). Nota-se a perda de muita informação ao redor de um possível *cluster* verdadeiro.

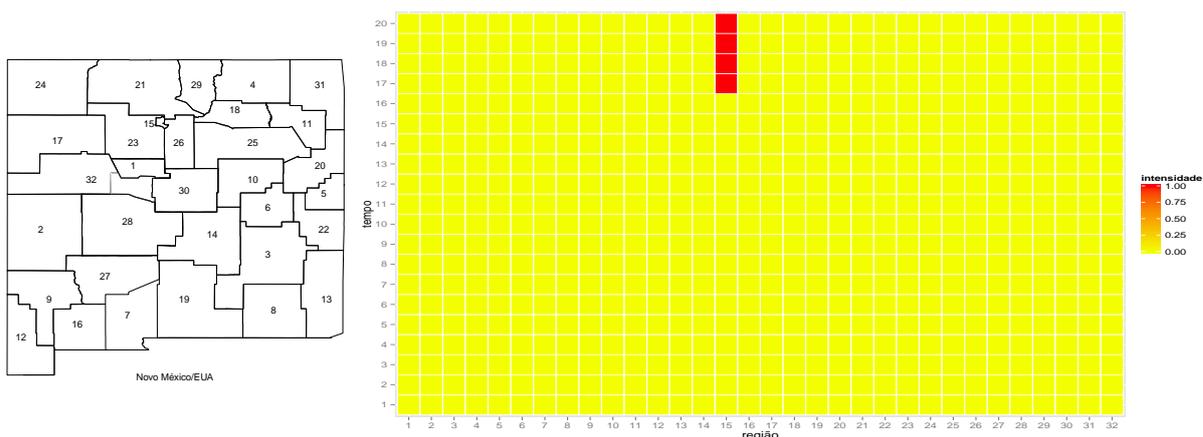


Figura 6: Solução para câncer de tireoide através do Scan Prospectivo.

A Figura 7 reproduz o efeito do alarme através da Estatística Scan espaço-temporal prospectiva. Por razões óbvias, como a metodologia apresentou somente uma solução significativa, existe um único tempo como proposta de tempo específico para o alarme. Dessa forma, não são mostrados potenciais tempos de alarme em outros instantes sob investigação, isso poderia ser visto somente com alguma técnica capaz de captar o efeito de intensidade nas possíveis ocorrências.

A confirmação do grande poder informativo da metodologia proposta pode ser observada através da análise da Figura 4 que sugere que a intensidade medida atra-

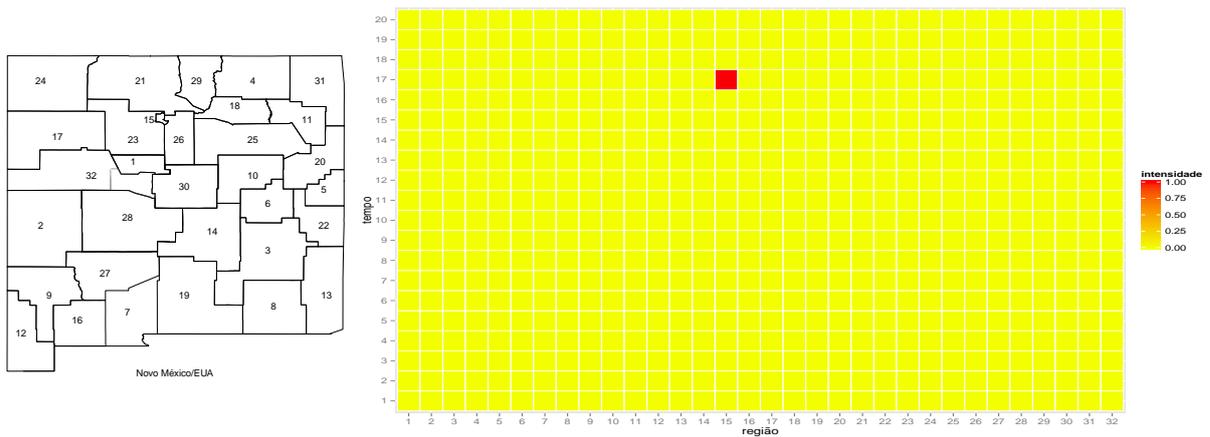


Figura 7: Alarme para câncer de tireoide através do Scan Prospectivo.

vés da função  $\mathcal{F}$  é efetiva nas regiões 1, 4, 18, 21, 23, 26, 29 e 30 mesmo que menores que a intensidade na região 15. É possível observar que esse conjunto de regiões é conexo no mapa em estudo e ilustra que o efeito tende a se dissipar nesse percurso como pode ser visto na Figura 8. É importante ressaltar que a intensidade verificada na Figura 8 foi obtida apenas através do último instante de tempo sob investigação, portanto é meramente ilustrativa na análise temporal, mas bastante relevante na análise espacial.

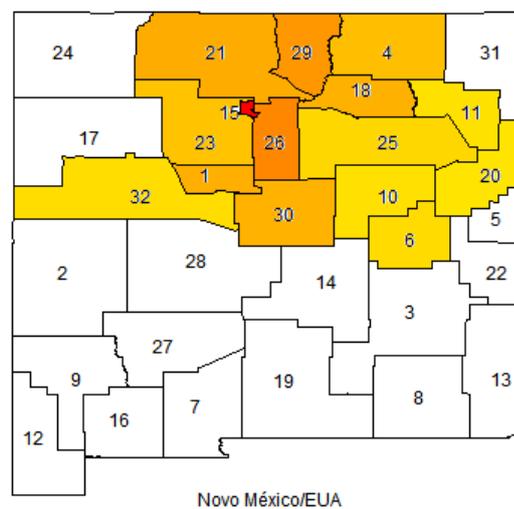


Figura 8: Dissipação espacial do efeito da função  $\mathcal{F}$  espaço-temporal.

Quando a solução obtida através da Estatística Scan espaço-temporal prospectiva é considerada, não existe dissipação do efeito. Apenas uma solução significativa composta por uma única região foi encontrada, nesse caso a região 15 denomi-

nada Los Alamos

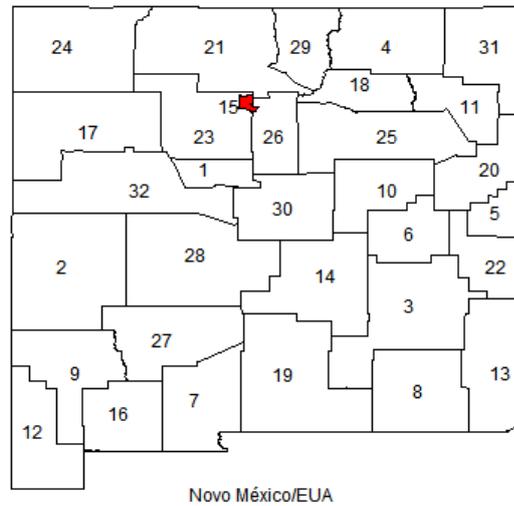


Figura 9: Dissipação espacial do efeito da Estatística Scan Prospectiva.

Já a solução sugerida através da razão de verossimilhança adaptativa (LIMA; DUCZMAL, 2014) apresenta o cilindro com base na região 15, mas no intervalo de tempo  $[15, 20]$ . Novamente a estratégia em discussão não permite delinear se vizinhanças desse cilindro são focos potenciais para se associarem a um possível *cluster*. Essa solução ilustra o elevado efeito de Intensidade já verificado a partir do tempo 15, através do uso da função  $\mathcal{F}$  espaço-temporal (veja Figura 4).

A Figura 10 sugere o que seria o efeito visual da única solução significativa fornecida através da razão de verossimilhança adaptativa (LIMA; DUCZMAL, 2014) sob a perspectiva da relevância por regiões e instantes de tempo.

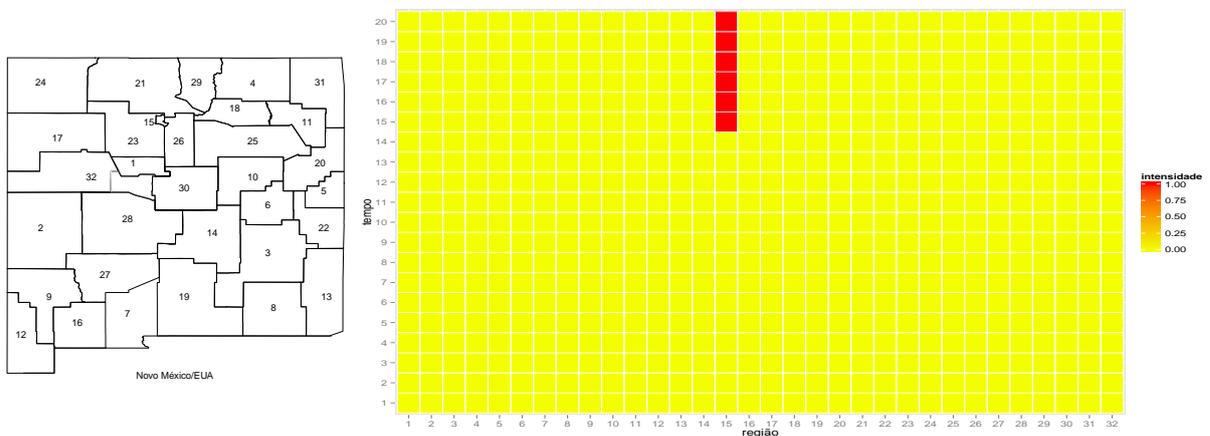


Figura 10: Solução para câncer de tireoide (Razão de verossimilhança adaptativa).

A Figura 11 sugere apenas um tempo específico para soar o alarme não mostrando potenciais alarmes em outros instantes sob investigação.

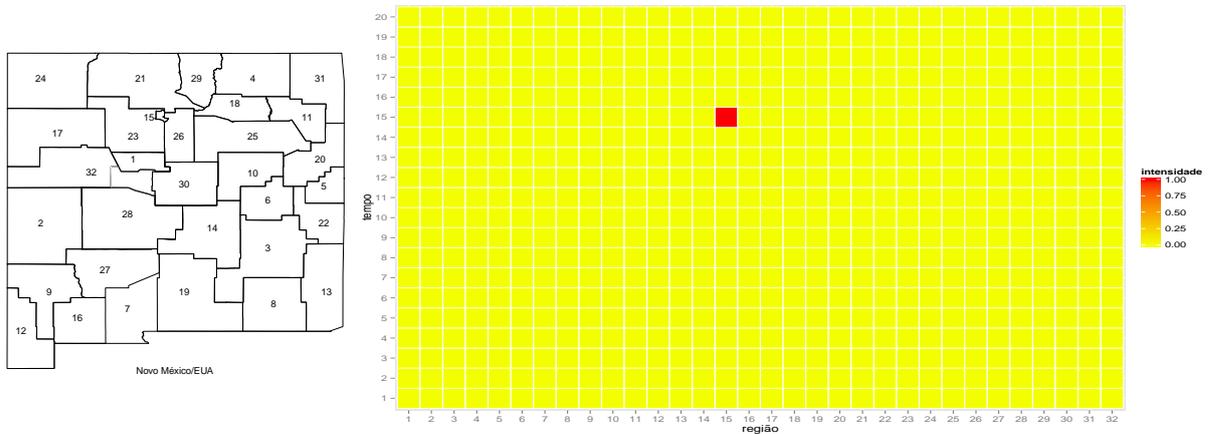


Figura 11: Alarme para câncer de tireoide (Razão de verossimilhança adaptativa).

Assim como visto anteriormente para a Estatística Scan espaço-temporal prospectiva, quando a solução obtida através da razão de verossimilhança adaptativa (LIMA; DUCZMAL, 2014) é considerada, não existe dissipação do efeito. Apenas uma solução significativa composta por uma única região foi encontrada.

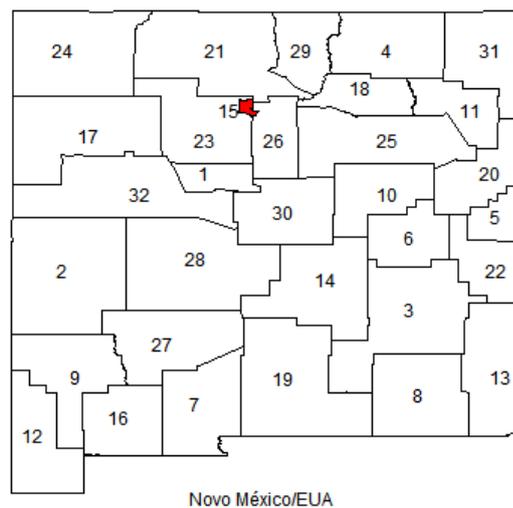


Figura 12: Dissipação espacial do efeito da Razão de verossimilhança adaptativa.

## 5.2 DADOS DE HANSENÍASE NO ESTADO DO AMAZONAS NO BRASIL

A Hanseníase ainda atinge a população de vários países. O Brasil é o segundo país no ranking de incidência dessa doença, e apresenta no estado do Amazonas um dos maiores focos do país. Segundo fontes do Ministério da Saúde brasileiro, a taxa de incidência de Hanseníase caiu mais de 25 por cento nos últimos 10 anos em todo o país. No estado do Amazonas, o índice em 2014 foi de 1,29 casos para cada 10 mil habitantes.

A Hanseníase é uma doença crônica, contagiosa transmitida de pessoa para pessoa ou pelas vias respiratórias através do contato com a bactéria *Mycobacterium leprae*. Um dos principais sintomas da doença é o aparecimento de manchas esbranquiçadas, avermelhadas ou de tom marron em qualquer parte do corpo. Normalmente, no local dessas manchas a pessoa não sente dor; calor; nem frio. Outros sinais importantes para identificar a Hanseníase é a presença de caroços e formigamento nos pés e mãos. Existe uma forte correlação dessa doença com as condições de vida da população e o acompanhamento epidemiológico é realizado através do coeficiente de detecção de novos casos.

Trata-se de uma doença curável, porém, se não tratada, o seu agravamento pode se tornar bastante preocupante. Atualmente, em praticamente todo o mundo, o tratamento é oferecido gratuitamente, e há várias campanhas para a erradicação da doença. Os países com maiores incidência são os menos desenvolvidos ou com condições precárias de higiene. Em 2011, o Ministério da Saúde brasileiro registrou no Brasil mais de 33 mil casos da doença.

Somente na zona urbana de Manaus, de acordo com a Fundação Alfredo da Mata (Fuam), 4104 casos ocorreram no período de 1998 a 2004. Essa doença está relacionada com as condições de vida da população. O coeficiente de detecção de novos casos, usado no seu monitoramento, é obtido através da divisão do total de novos casos pela população sob risco e depois multiplicando o resultado por 10.000

habitantes.

Como referencial para acompanhamento epidemiológico foram adotados parâmetros do Ministério da Saúde do Brasil, no qual estabelece como hiperendêmico os valores de referência para os coeficiente de detecção:  $\geq 4,0$  por 10.000 habitantes como muito alto; entre 4,0 e 2,0 por 10.000 habitantes como alto; entre 2,0 e 1,0 por 10.000 habitantes como médio; entre 1,0 e 0,2 por 10.000 habitantes como baixo.

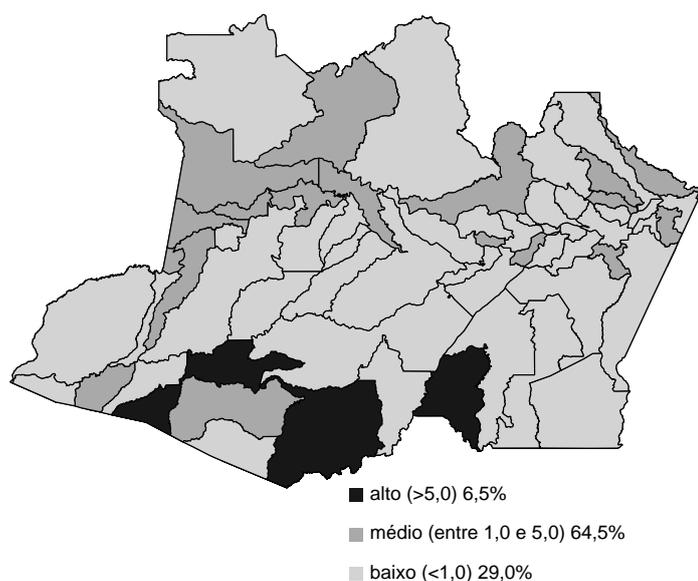


Figura 13: Prevalência do Hanseníase (por 10 mil habitantes) no Amazonas/BRA.  
 FONTE: Secretaria de Saúde do Amazonas/BRA.

O mapa apresentado na Figura 13 ilustra as taxas de prevalência da Hanseníase no estado do Amazonas no ano de 2005. Os municípios de Envira, Itamarati, Lábrea e Humaitá mostram altas taxas. O foco no estudo da incidência dos casos do Hanseníase ao longo do tempo também é de grande interesse. Isso se deve a alguns motivos, em particular: entender as rotas de propagação da doença munindo-se de informações capazes de subsidiar o desenvolvimento de ferramentas de controle epidemiológico eficazes e de baixo custo; traçando ações de combate e erradicação da doença.

Foram utilizados os dados de registros de novos casos da Hanseníase em menores de 15 anos no estado do Amazonas no período de janeiro de 2008 até de-

zembro de 2009, sendo um forte indicador do nível endêmico o estudo dessa faixa etária abordada. Os dados de população e coordenadas geográficas foram obtidos na Secretaria de Planejamento do estado do Amazonas (SEPLAN), e o conjunto de dados foi obtido através da Central de Dados do Sistema Único de Saúde do Ministério da Saúde brasileiro (DATASUS) e através do Instituto Brasileiro de Geografia e Estatística (IBGE).

A região específica em estudo no estado do Amazonas/BRA, subdividida em 62 regiões pode ser visualizada na Figura 14.

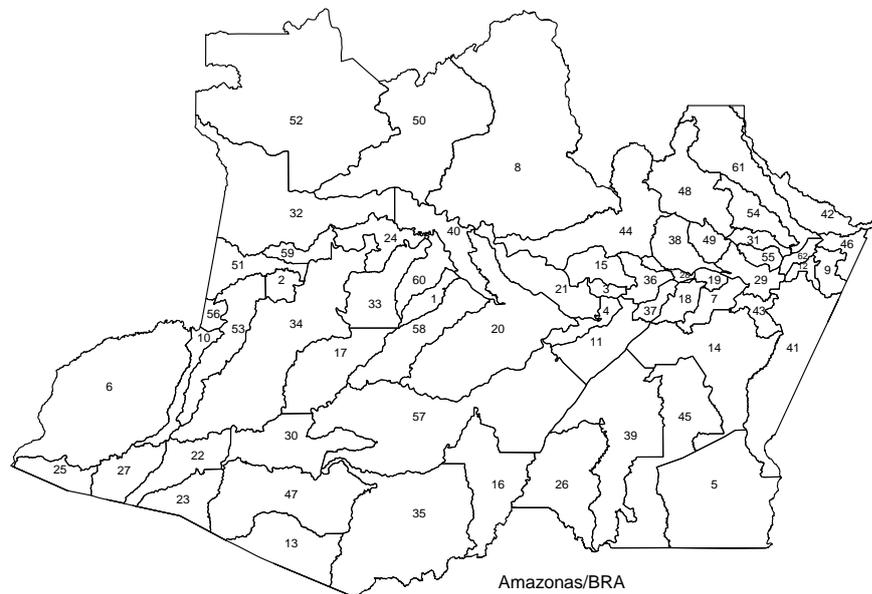


Figura 14: Mapa do Amazonas/BRA.

A proposta desse estudo associado aos dados da Hanseníase no estado do Amazonas/BRA aponta para a efetiva utilização da técnica proposta através da Função  $\mathcal{F}$  espaço-tempo e também do Alarme através da função  $\mathcal{F}$ . O interesse central não é comparativo entre metodologias, mas sim uma possível confirmação do alarmante problema da disseminação dessa doença na região amazônica. Em virtude disso, apenas a função  $\mathcal{F}$  e a já difundida estatística Scan espaço-temporal prospectiva serão utilizadas. Os procedimentos de simulação de Monte Carlo foram executados seguindo as mesmas quantidades utilizadas no estudo com os dados do estado

do Novo México/EUA.

As regiões codificadas numericamente são descritas conforme a Tabela 2.

Tabela 2: Sub-regiões do mapa do Amazonas/BRA

código	região	código	região
1	Alvarães	32	Japurá
2	Amaturá	33	Juruá
3	Anamã	34	Jutaí
4	Anori	35	Lábrea
5	Apuí	36	Manacapuru
6	Atalaia Do Norte	37	Manaquiri
7	Autazes	38	Manaus
8	Barcelos	39	Manicoré
9	Barreirinha	40	Maraã
10	Benjamin Constant	41	Maués
11	Beruri	42	Nhamundá
12	Boa Vista do Ramos	43	Nova Olinda do Norte
13	Boca do Acre	44	Novo Airão
14	Borba	45	Novo Aripuanã
15	Caapiranga	46	Parintins
16	Canutama	47	Pauini
17	Carauari	48	Presidente Figueiredo
18	Careiro	49	Rio Preto da Eva
19	Careiro da Várzea	50	Santa Isabel do Rio Negro
20	Coari	51	Santo Antônio do Içá
21	Codajás	52	São Gabriel da Cachoeira
22	Eirunepé	53	São Paulo de Olivença
23	Envira	54	São Sebastião do Uatumã
24	Fonte Boa	55	Silves
25	Guajará	56	Tabatinga
26	Humaitá	57	Tapauá
27	Ipixuna	58	Tefé
28	Iranduba	59	Tonantins
29	Itacoatiara	60	Uarini
30	Itamarati	61	Urucará
31	Itapiranga	62	Urucurituba

A Figura 15 apresenta o efeito de intensidade medido através da função  $\mathcal{F}$  em cada área do mapa associado à cada intervalo temporal. A interpretação é mesma já discutida através dos dados de câncer de tireoide no estado do Novo México/EUA. O número de subdivisões do mapa em estudo aqui é quase o dobro do estudo anterior, entretanto não ocorre prejuízo na estratégia de visualização.

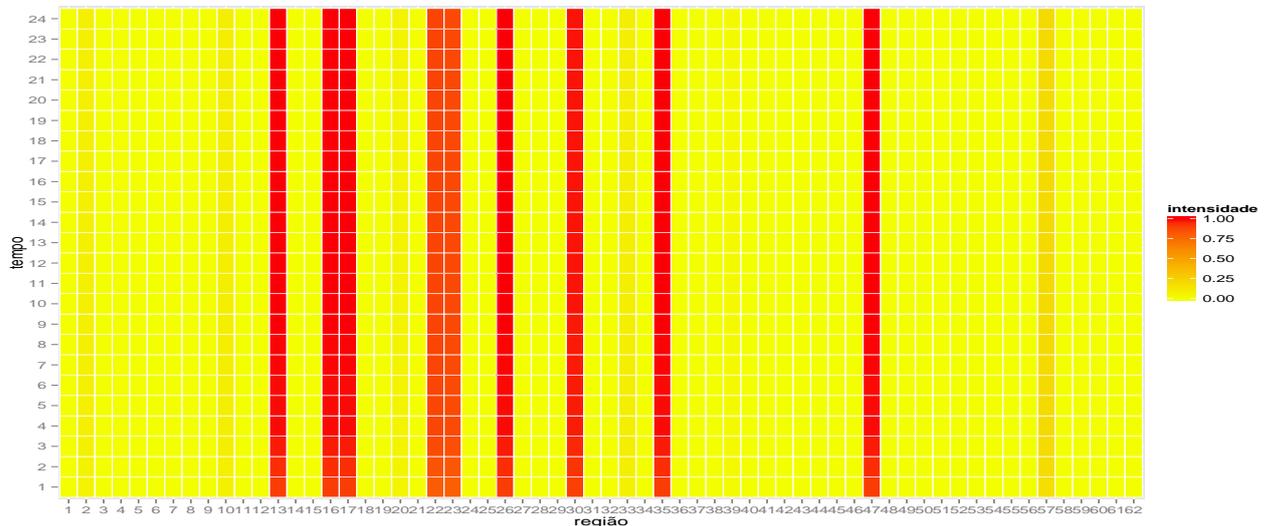
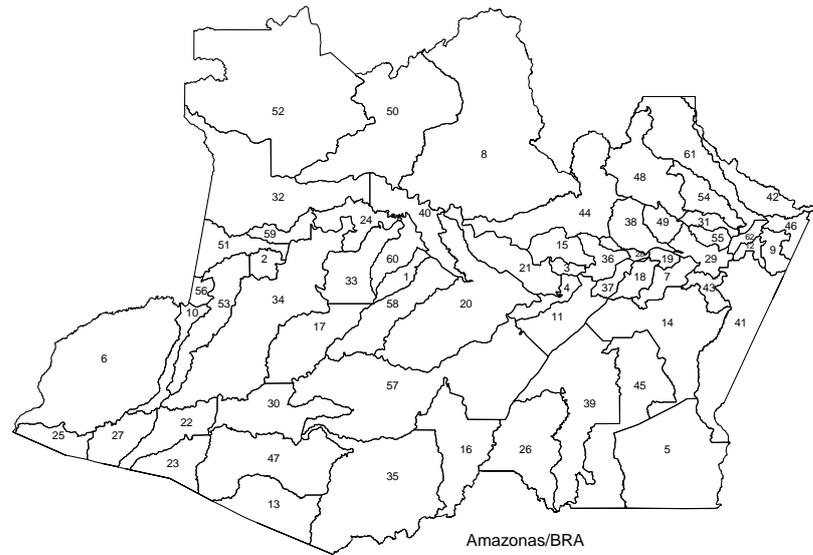


Figura 15: Função  $\mathcal{F}$  para Hanseníase no Amazonas/BRA.

É fácil ver uma elevada intensidade medida através da função  $\mathcal{F}$  para as regiões 13, 16, 17, 26, 30, 35 e 47, além de uma intensidade considerável nas regiões 22 e 23. Esse alto efeito de intensidade é detectado logo no primeiro instante temporal sobre investigação. Os referidos municípios com alta intensidade são: Boca do Acre, Canutama, Carauari, Humaitá, Itamarati, Labréa e Pauini. Os outros também com intensidade elevada são Eurinepé e Envira.

Como visto anteriormente na Figura 13, mesmo em 2005, antes do período sob investigação, municípios como Envira, Itamarati, Lábrea e Humaitá já apresentavam

taxas de prevalência bastante elevadas. Isso é uma provável explicação para o fato da função  $\mathcal{F}$  apresentar valores elevados desde o primeiro tempo sob investigação.

A Figura 16 apresenta o efeito do alarme através da função  $\mathcal{F}$ , ainda associado a cada área do mapa em estudo. As regiões com intensidade destacada para o alarme são, como previsto, as mesmas discutidas anteriormente, e todas elas com o alarme sendo soado logo no primeiro tempo sob investigação.

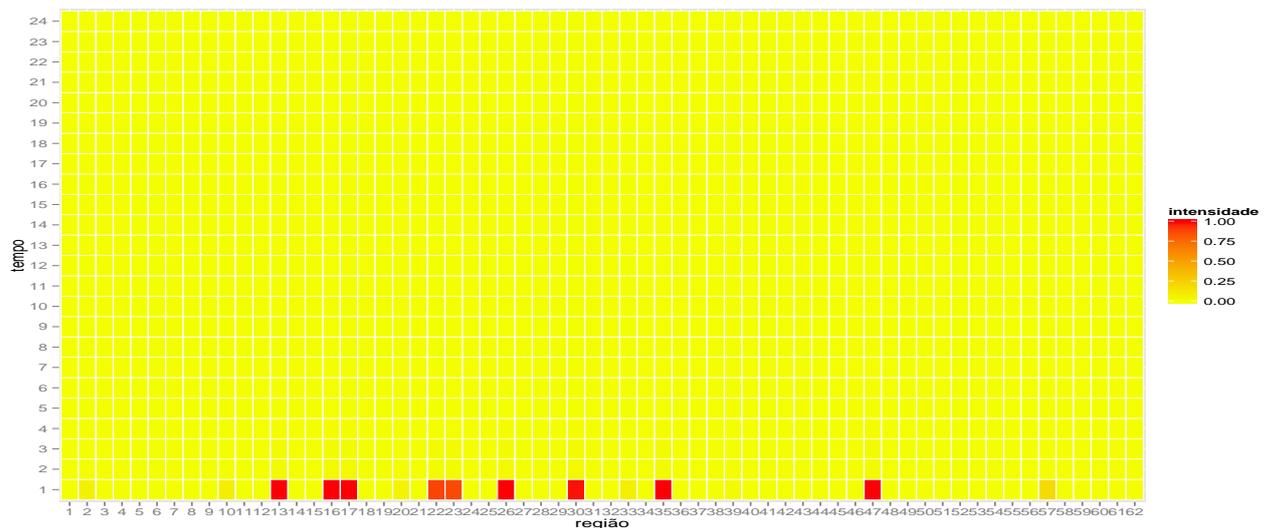
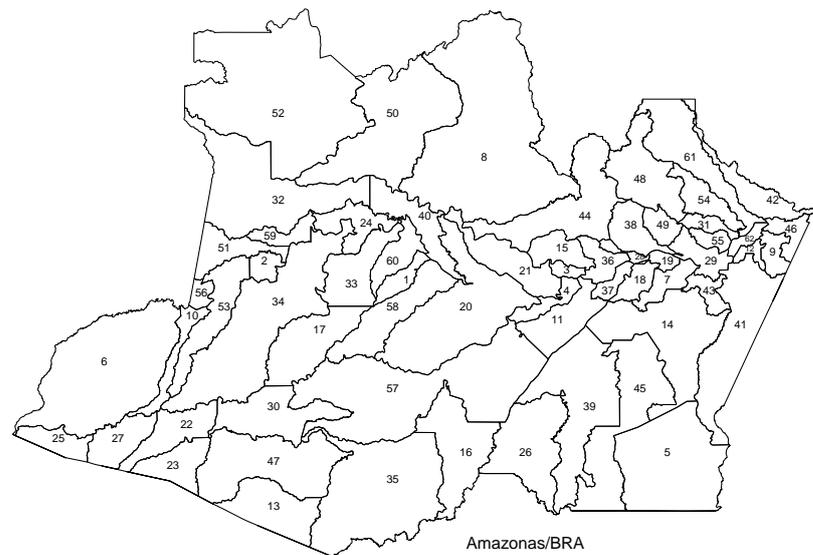


Figura 16: Alarme através da função  $\mathcal{F}$  para Hanseníase no Amazonas/BRA.

A Figura 17 deixa mais claro esse efeito visual, mesmo considerando a diferença temporal entre a Figura 13, com informações de 2005, e a Figura 17, com infor-

mações de 2008 e 2009, a comparação entre as duas imagens é um tanto elucidativa sobre o delineamento de regiões que efetivamente se tornaram alvo para investigação de problemas sobre a Hanseníase através da avaliação da função  $\mathcal{F}$  espaço-tempo e do Alarme através da função  $\mathcal{F}$ .

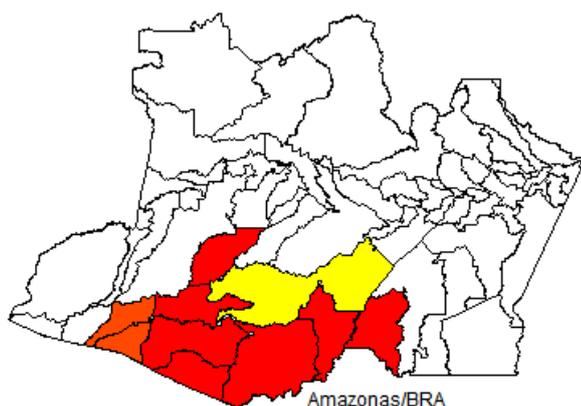


Figura 17: Dissipação espacial da função  $\mathcal{F}$  no Amazonas/BRA.

Em um segundo momento, o conjunto de dados com os casos de Hanseníase entre 2008 e 2009 no estado do Amazonas/BRA foi submetido à metodologia de detecção através da Estatística Scan espaço-temporal prospectiva. A expectativa inicial era verificar uma solução semelhante a imagem verificada na Figura 17.

Para não ocorrer nenhuma influência na metodologia Scan, não foi utilizado qualquer código que tenha sido implementado nessa pesquisa, mas sim o programa original disponibilizado em <http://www.satscan.org/>. O programa disponível, em sua versão prospectiva, investiga somente cilindros cujo comprimento do intervalo de tempo seja no máximo 90% do tempo total sob investigação. Dessa forma, para um intervalo de tempo sob investigação com comprimento  $n$ , o maior intervalo de tempo a ser pesquisado é dado por  $[0.1 \times n; n]$ , com a base sendo definida por todas as zonas espaciais em avaliação.

A solução sugerida pela Estatística Scan espaço-temporal prospectiva coincide com as nove áreas em vermelho e laranja escuro na Figura 17, sendo as regiões de Boca do Acre, Canutama, Carauari, Eurinepé, Envira, Humaitá, Itamarati, Labréa e

Pauini. A Figura 18 mostra essa solução através do mesmo contexto visual utilizado para analisar a Função Intensidade espaço-temporal.

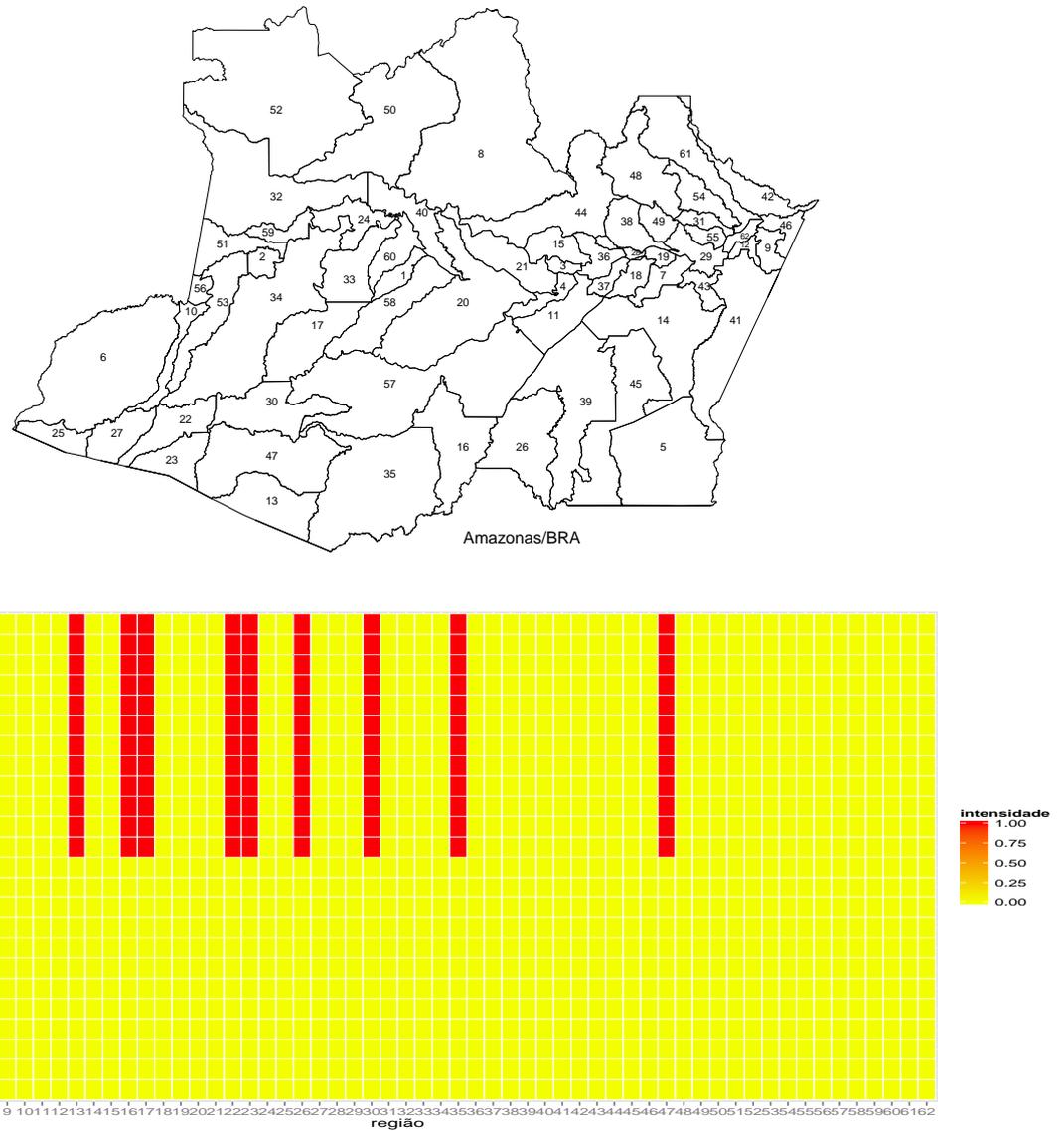


Figura 18: Solução para Hanseníase através do Scan Prospectivo.

Quanto à análise temporal, a solução apresentada é um cilindro no intervalo temporal iniciado no tempo 13 e findado no tempo 24. Isso mostra uma consonância com as conclusões advindas da Função  $\mathcal{F}$  espaço-temporal, para a análise espacial, mas uma diferença para a análise temporal. A função  $\mathcal{F}$  apresentou todo o tempo com intensidade alta nas regiões delineadas, isso não ocorre na solução através da Estatística Scan espaço-temporal prospectiva. O mais provável é que o Scan Pros-

pectivo não seja capaz de captar essa informação quando considerado um cilindro de comprimento temporal tão extenso.

Dessa forma, a solução mais verossímil obtida compreende a segunda metade do intervalo de tempo sob investigação. Nesse cenário, o alarme então seria soado, para todas as regiões anteriormente citadas, no tempo 13, ou seja, em Janeiro de 2009, como pode ser visto na Figura 19.

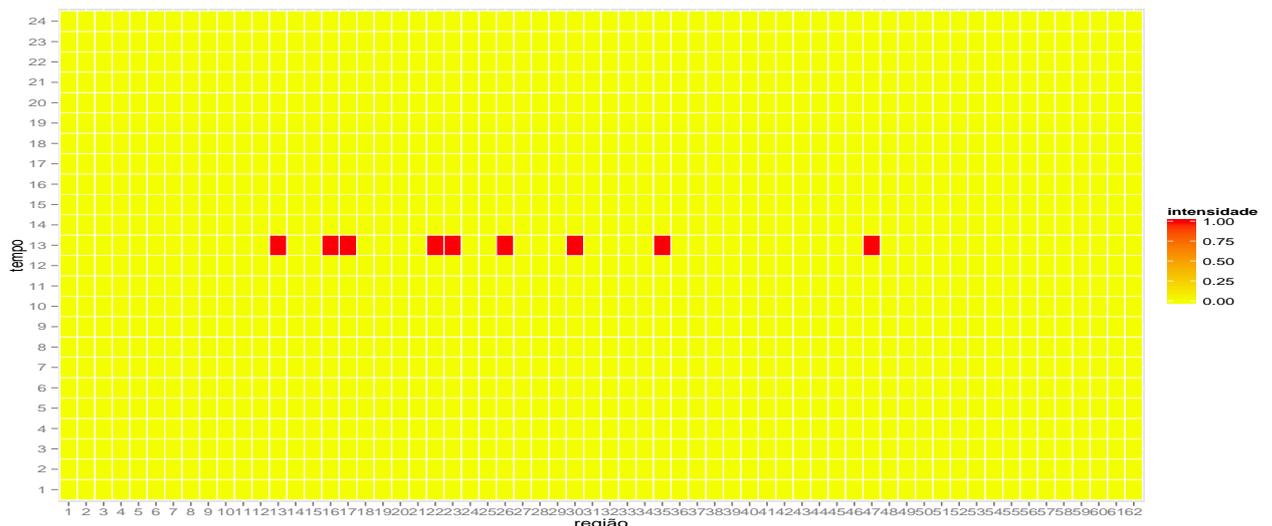
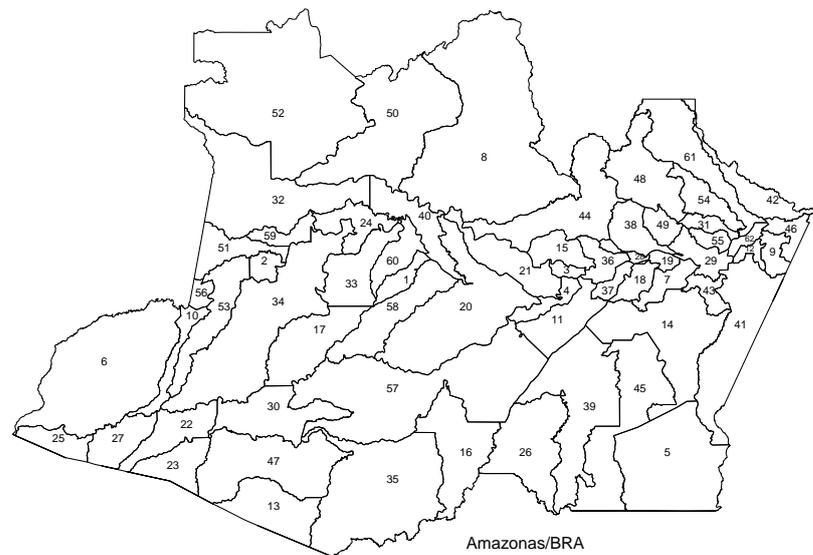


Figura 19: Alarme para Hanseníase através do Scan Prospectivo.

O poder informativo da metodologia proposta pode ser ilustrado através da comparação entre as Figuras 17 e 20. Apenas uma região não é comum às duas

imagens, mas já suficiente para notar que a Estatística Scan espaço-temporal prospectiva não é capaz de delinear o efeito de dissipação da intensidade com a qual o fenômeno ocorre nas regiões fronteiriças a um possível *cluster*, tampouco capaz de verificar dentro do próprio possível *cluster*, regiões que sejam mais ou menos intensas.

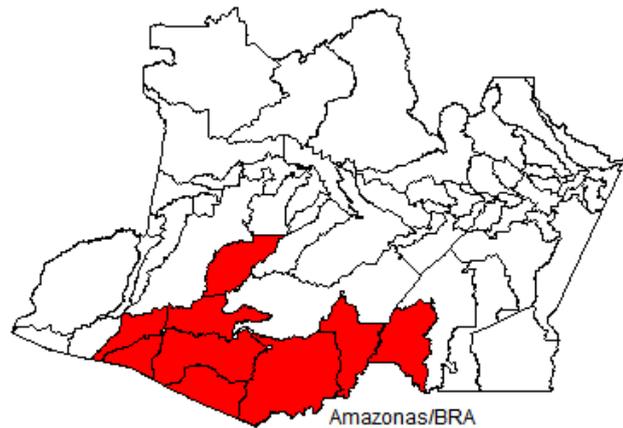


Figura 20: Dissipação para Estatística Scan Prospectiva no Amazonas/BRA.

### 5.3 AVALIAÇÃO DE DADOS SIMULADOS

Um estudo através de dados simulados foi realizado considerando inicialmente os dados originais do mapa do Novo México com dados de câncer de tireoide. Partindo da distribuição original de população ao longo dos 20 tempos utilizados anteriormente, casos simulados foram distribuídos através de simulações de Monte Carlo. Para cada área  $\ell$  do mapa, foi considerado um processo de Poisson cuja função taxa  $\mu_{t,\ell}$ , em algum tempo  $t$ , sob a validade da hipótese de não existência de *cluster* é proporcional ao tamanho populacional da área  $\ell$  no tempo  $t$ . Para fins de investigação, um *cluster* artificial será construído sobre o mapa, seguindo a premissa que a partir de um tempo  $k$  pré-fixado, um subconjunto  $\xi$  de regiões conexas previamente escolhidas terá sua função taxa ampliada seguindo o seguinte procedimento.

$$\theta_{t,\ell} = \begin{cases} \mu_{t,\ell} & \forall \ell \in \{1, \dots, L\} \text{ e } t < k; \\ \mu_{t,\ell} & \text{se } \ell \notin \xi \text{ e } t \geq k; \\ \mu_{t,\ell} \times \delta_{t,\ell} & \text{se } \ell \in \xi \text{ e } t \geq k. \end{cases}$$

em que  $\delta_{t,\ell}$  representa um aumento no risco de cada indivíduo da população contrair a doença, ou seja,  $\delta_{t,\ell} = 1 + \varepsilon$ , com  $\varepsilon > 0$ . Isso implica que os indivíduos na região  $\ell$  no tempo  $t$  possuem  $\varepsilon \times 100\%$  a mais de chance de ser um caso da doença em estudo se comparados aos demais indivíduos do mapa em estudo.

Dado algum mapa específico em estudo, quando a hipótese alternativa de que existe um *cluster* determinado pela zona espacial  $\xi$  iniciando-se no tempo  $k$  é verdadeira, a localização espacial estimada do *cluster*, dada por  $\hat{\xi}$ , fornecida pela metodologia em uso, por melhor que seja o método, não tende a coincidir com a zona  $\xi$  exatamente, mas sim em um grande percentual das áreas pertencentes a  $\xi$ .

Nesse contexto, Lima e Duczmal (2014) propõem algumas métricas para avaliação de precisão em metodologias de detecção de *clusters* espaço-temporais. Para  $Q$  simulações de Monte Carlo, indexadas por  $q \in \{1, 2, \dots, Q\}$ . A Sensitividade Condicional de um alarme  $SSC_{k,\xi}[\tau]$  é representada pela proporção de localizações espaciais pertencentes a zona  $\xi$  que foram corretamente localizadas pelo *cluster* detectado  $\hat{\xi}_q$  na  $q$ -ésima simulação de Monte Carlo, dado que  $\tau \geq k$ . Essa medida é estimada por:

$$\frac{\sum_{q=1}^Q \left( \frac{\#\{\hat{\xi}_q \cap \xi\}}{\#\xi} \right) \mathbb{1}_{\{\tau \geq k\}}}{\sum_{q=1}^Q \mathbb{1}_{\{\tau \geq k\}}}$$

Já o valor preditivo positivo condicional de um alarme,  $VPP_{k,\xi}[\tau]$  é a proporção de localizações espaciais presentes em  $\hat{\xi}_q$  que pertencem a  $\xi$  dado que  $\tau \geq k$ , estimado por:

$$\frac{\sum_{q=1}^Q \left( \frac{\#\{\hat{\xi}_q \cap \xi\}}{\#\hat{\xi}_q} \right) \mathbb{1}_{\{\tau \geq k\}}}{\sum_{q=1}^Q \mathbb{1}_{\{\tau \geq k\}}}$$

Nos dois formatos propostos,  $\mathbb{1}_A$  representa a função indicadora para o evento  $A$  e  $\#A$  a cardinalidade do conjunto  $A$ .

As medidas  $SSC_{k,\xi}[\tau]$  e  $VPP_{k,\xi}[\tau]$  avaliam o desempenho de um método de detecção em relação à localização espacial estimada dado que o alarme é verdadeiro. Se  $\hat{\xi} = \xi$ , tem-se que  $SSC_{k,\xi}[\tau] = VPP_{k,\xi}[\tau] = 1$ . Para  $\hat{\xi} \subset \xi$ ,  $SSC_{k,\xi}[\tau] \geq VPP_{k,\xi}[\tau]$  e se  $\xi \subset \hat{\xi}$  implica em  $SSC_{k,\xi}[\tau] \leq VPP_{k,\xi}[\tau]$ . O adequado é que ambas as medidas sejam próximas de 1 indicando muita proximidade da solução detectada em relação ao *cluster* verdadeiro.

A velocidade no processo de detecção pode ser descrita pelo atraso, ou propriamente o tempo médio de atraso  $TMA_{k,\xi}[\tau]$  até o soar do alarme afirmando que realmente existe um *cluster* emergente no mapa. Essa medida pode ser estimada por:

$$\frac{\sum_{q=1}^Q (\tau - k) \mathbb{1}_{\{\tau \geq k\}}}{\sum_{q=1}^Q \mathbb{1}_{\{\tau \geq k\}}}$$

Essas medidas propostas por Lima e Duczmal (2014) são adequadas para métodos de detecção de *clusters* espaço-temporais. Entretanto, a proposta aqui discutida não é efetivamente um método de detecção de *clusters*, mas sim, um mecanismo de delineamento de regiões e instantes temporais mais ou menos relevantes em relação a uma possível existência de *cluster*. Dessa forma, não é possível calcular tais medidas para a função  $\mathcal{F}$  e posteriormente buscar uma comparação entre ela e os métodos já difundidos.

A determinação sobre o fato do cilindro mais verossímil ser ou não efetivamente um *cluster*, depende da escolha de um nível de significância  $\alpha$  e da obtenção de um valor crítico para a estatística de teste em uso, obtido sob a validade da hipótese nula. Esse tipo de procedimento não é considerado para a utilização da função  $\mathcal{F}$ .

Com interesse em estabelecer tal comparação, será produzido o seguinte procedimento. Para cada execução de Monte Carlo contendo o *cluster* artificial, construído para o cálculo das medidas de desempenho em métodos de detecção, é executado o procedimento da função  $\mathcal{F}$ . Em outras palavras, considere o conjunto de dados simulados na  $q$ -ésima rodada de simulações, use essa distribuição de casos como sendo as taxas para execução do procedimento da função  $\mathcal{F}$ . Ao final desse procedimento, defina como *cluster* estimado  $\hat{\xi}_q$  o conjunto das regiões tais que a função  $\mathcal{F}$  para algum tempo  $t \geq k$  se encontra entre os  $\alpha \times 100\%$  (o nível de significância utilizado no método de detecção de *clusters* que será comparado) maiores valores verificados para a função  $\mathcal{F}$  na referida sequência de execuções. Dessa forma, as medidas  $SSC_{k,\xi}[\tau]$  e  $VPP_{k,\xi}[\tau]$  passam a ser calculáveis.

Quando considerada a medida  $TMA_{k,\xi}[\tau]$ , o alarme através da função  $\mathcal{F}$  será utilizado de forma análoga. O tempo de alarme estimado  $\tau$  que será utilizado no cálculo da medida  $TMA$  será o menor tempo superior ao valor  $k$  (início temporal do *cluster* artificial) dentre as regiões pertencentes ao *cluster* estimado, tal que seu alarme através da função  $\mathcal{F}$  esteja entre os  $\alpha \times 100\%$  (o nível de significância utilizado anteriormente) maiores valores verificados para a função Intensidade de alarme.

Para os testes com dados simulados foi construído um *cluster* artificial composto pelas regiões Chaves, Eddy, Lea e Otero, de codificação 3, 8, 13 e 19 a partir do tempo  $k = 11$ . O *cluster* artificial inicia-se no tempo 11 e permanece ativo até o tempo final sob investigação. O parâmetro  $\delta_{t,\ell}$  foi testado para diferentes valores  $\{1.25, 1.50, 1.75, 2.00\}$  e ainda para  $\delta_{t,\ell} = 1.3^{t-11+1}$  para  $t \geq 11$ . Nesse último caso, o risco cresce monotonicamente com o tempo. O nível de significância adotado foi



vés das medidas de  $SSC_{k,\xi}[\tau]$ ,  $VPP_{k,\xi}[\tau]$  e  $TMA_{k,\xi}[\tau]$  com as medidas similares adaptadas obtidas para a função  $\mathcal{F}$  espaço-temporal e para o alarme através da função  $\mathcal{F}$ . Quanto a Razão de verossimilhança adaptativa, existem três diferentes abordagens denominadas GMAX-ALR, GMIX-ALR e GW-ALR. Mais informações sobre a concepção dessas abordagens podem ser obtidos por Lima e Duczmal (2014). Para efeitos comparativos, as três abordagens foram utilizadas, os resultados podem ser visualizados na Tabela 3. É importante ressaltar que os métodos Scan prospectivo e ainda, GMAX-ALR, GMIX-ALR e GW-ALR são técnicas de detecção de clusters, diferente da abordagem através da função  $\mathcal{F}$  que tem o seu interesse centrado em delinear e visualizar as regiões e também instantes de tempo mais e menos relevantes para a formação de um possível *cluster* sobre o mapa em estudo.

O experimento foi conduzido para 1,000 simulações de Monte Carlo, cada uma delas contendo o *cluster* artificial, com interesse em produzir os resultados para o Scan prospectivo e as três abordagens da utilização da Razão de verossimilhança adaptativa. Cada uma dessas 1,000 simulações de Monte Carlo foi submetida ao mecanismo de cálculo da função  $\mathcal{F}$  para ser considerado como o conjunto de dados reais e então ser executado o procedimento. A partir dessa configuração (suposta como configuração real) 10,000 simulações de Monte Carlo foram produzidas para o cálculo da função  $\mathcal{F}$ . Em cada uma situação, as medidas  $SSC_{k,\xi}[\tau]$ ,  $VPP_{k,\xi}[\tau]$  e  $TMA_{k,\xi}[\tau]$  foram avaliadas como descritas anteriormente, os resultados obtidos podem ser visualizados na Tabela 3.

A Tabela 3 apresenta em **negrito** os resultados de maior eficiência. É importante sempre ressaltar que a função  $\mathcal{F}$  não representa uma metodologia de detecção de clusters, mas sim um delineamento de regiões e instantes de tempo com mais ou menos potencial de pertencerem a um possível cluster verdadeiro se este existir. É relevante lembrar esse fato, pois apenas foi feita uma adaptação para a busca de uma estratégia de verificação comparativa para caracterizar quão bem esse delineamento está sendo feito.

Tabela 3: Performance através de medidas de desempenho

	$\delta_{t,\ell}$	$SSC$	$VPP$	$TMA$
Scan prospectivo	1.25	0.684	0.578	5.945
	1.50	0.763	0.726	3.451
	1.75	0.765	0.799	1.743
	2.00	0.785	0.851	1.011
	$1.3^{t-11+1}$	0.777	0.821	1.991
GMAX-ALR	1.25	0.674	0.445	3.248
	1.50	0.799	0.514	0.955
	1.75	0.857	0.540	0.591
	2.00	0.882	0.497	0.118
	$1.3^{t-11+1}$	0.827	0.516	0.976
GMIX-ALR	1.25	<b>0.714</b>	0.432	2.611
	1.50	<b>0.803</b>	0.497	0.794
	1.75	<b>0.837</b>	0.510	<b>0.264</b>
	2.00	0.881	0.497	<b>0.092</b>
	$1.3^{t-11+1}$	0.836	0.511	0.898
GW-ALR	1.25	0.702	0.417	<b>2.403</b>
	1.50	0.780	0.498	<b>0.774</b>
	1.75	0.834	0.518	0.363
	2.00	0.870	0.493	0.187
	$1.3^{t-11+1}$	0.795	0.516	<b>0.881</b>
função $\mathcal{F}$	1.25	0.530	<b>0.721</b>	4.510
	1.50	0.653	<b>0.765</b>	2.910
	1.75	0.791	<b>0.972</b>	1.720
	2.00	<b>0.883</b>	<b>0.929</b>	0.880
	$1.3^{t-11+1}$	<b>0.879</b>	<b>0.974</b>	0.979

Considerando tudo isso, o mecanismo utilizando a Razão de verossimilhança adaptativa se mostra eficiente quando avaliado através da métrica  $SSC_{k,\xi}[\tau]$ , em outras palavras, o método tende a identificar uma grande parcela do verdadeiro *cluster*. Através da abordagem GMIX-ALR, os valores foram os mais elevados para os casos em que  $\delta_{t,\ell} = \{1.25, 1.50, 1.75\}$ . Já quando  $\delta_{t,\ell} = \{2.00, 1.3^{t-11+1}\}$  a função  $\mathcal{F}$  foi mais efetiva, entretanto a Razão de verossimilhança adaptativa ainda produziu valores bastante altos.

Por outro lado, quando analisada a medida  $VPP_{k,\xi}[\tau]$ , uma notória superioridade para a função  $\mathcal{F}$  é verificada, ou seja, o delineamento através da função  $\mathcal{F}$  informa que uma grande parcela da região efetivamente mais intensa pertence ao cluster verdadeiro. Vale lembrar que, para efeito de avaliação da função  $\mathcal{F}$ , foi con-

siderado como cluster detectado o conjunto das regiões tais que a função  $\mathcal{F}$  para algum tempo  $t \geq k$  se encontra entre os  $\alpha \times 100\%$  (o nível de significância utilizado anteriormente) maiores valores verificados para a função  $\mathcal{F}$  na referida sequência de execuções.

Para a medida  $TMA_{k,\xi}[\tau]$ , a Razão de verossimilhança adaptativa ainda se mostrou mais eficiente. Entretanto, o método sabidamente mais utilizado e difundido, a técnica através da Estatística Scan espaço-temporal prospectiva, apresentou, em média, um atraso na detecção superior ao atraso verificado para função  $\mathcal{F}$  através de seu alarme através da função  $\mathcal{F}$ . Novamente deve ser frisado que não se trata de uma técnica de detecção ou um mecanismo para soar alarmes endêmicos, mas sim uma ferramenta de delineamento e visualização de tempos potencialmente relevantes para soar o alarme de surgimento de *clusters*.

### 5.3.1 FUNÇÃO $\mathcal{F}$ PONDERADA

A proposição de uma técnica similar ao formato discutido anteriormente, porém ajustado com algum tipo de ponderação é uma tentativa de tornar a função  $\mathcal{F}$  mais sensível no processo de delineamento da formação de possíveis *clusters* existentes nos mapas em estudo. O procedimento de utilização da função  $\mathcal{F}$  espaço temporal e da medida do alarme através da função  $\mathcal{F}$  prima por delinear as áreas com maior e menor possibilidade de pertencerem a um possível *cluster* existente no mapa em estudo. Assim como verificar os instantes temporais mais e menos prováveis para determinar o instante de surgimento de um *cluster* espaço temporal.

A avaliação é dependente de uma sequência de execução de simulações de Monte Carlo para o cálculo da função  $\mathcal{F}$ , anteriormente definida por  $\mathcal{F}_{t,\ell}$ , e para o alarme através da função  $\mathcal{F}$ , anteriormente definida por  $\mathcal{A}_{t,\ell}$ . As medidas são obtidas considerando a proporção de vezes que cada região e instante temporal pertencem a solução mais verossímil obtida em cada execução de Monte Carlo. Por outro lado, ao considerar todas as simulações de Monte Carlo, existem rodadas de simulação

com soluções de função de verossimilhança muito elevada, ao passo que em outras rodadas de simulação, a função de verossimilhança para a solução mais verossímil atinge valores mais moderados. Essa diferença pode ser incorporada ao processo de análise, levando em conta uma ordenação entre todas as soluções mais verossímeis obtidas através das simulações de Monte Carlo.

A proposição de estudos através da função  $\mathcal{F}$  foi iniciada em um problema puramente espacial. Esse texto apresenta estudos para avançar para uma proposição espaço temporal. Dessa forma, incluir uma nova formatação para a função  $\mathcal{F}$ , baseada em algum tipo de ponderação, já em seu formato espaço temporal, sem um estudo prévio, puramente espacial, seria no mínimo uma atitude apressada. Ao considerar essa visão, será proposto aqui um estudo inicial de simulação considerando a proposição da formatação que leva em conta a ordenação de todas as soluções na visão puramente espacial.

Para um mapa dividido em  $L$  regiões contíguas com população  $p_\ell$  e  $c_\ell$  casos do fenômeno de interesse para  $\ell \in \{1, 2, \dots, L\}$ , a função  $\mathcal{F}$  puramente espacial considera uma sequência de  $Q$  simulações de Monte Carlo em que o número de casos simulados para cada uma das regiões segue distribuição Poisson com taxa  $c_\ell$ . A função  $\mathcal{F}$  associada a cada uma das regiões é dada pela proporção de vezes que a região pertence à solução mais verossímil dentre todas as  $Q$  execução de Monte Carlo.

Considere agora modificar o procedimento; para  $Q$  simulações de Monte Carlo executadas para a obtenção do valor da função  $\mathcal{F}$ , defina  $LLR_q$  como o valor da função de verossimilhança para a solução mais verossímil obtida na  $q$ -ésima rodada de simulação. Considere ainda o conjunto dos valores da função de verossimilhança para a solução mais verossímil obtidos nas  $Q$  simulações,  $\mathcal{L} = \{LLR_1, LLR_2, \dots, LLR_Q\}$  e o conjunto  $\mathcal{L}^* = \{LLR_{(1)}, LLR_{(2)}, \dots, LLR_{(Q)}\}$  o conjunto ordenado desses valores.

Seja  $g$  uma função com domínio no conjunto  $\mathcal{L}$  e imagem no conjunto  $\{1, 2, \dots, Q\}$  tal que tal que  $g(LLR_q)$  representa a posição de  $LLR_q$  no conjunto orde-

nado, ou seja,  $g(LLR_q) = 1$  implica que  $LLR_q$  é o menor valor dentre as  $Q$  simulações, ou seja, o menor valor no conjunto  $\mathcal{L}$ , já  $g(LLR_q) = 2$  implica que  $LLR_q$  é o segundo menor valor no conjunto  $\mathcal{L}$ . Ao considerar  $g(LLR_q) = Q$ , isso implica que  $LLR_q$  é maior valor dentre as  $Q$  simulações, ou seja, o maior valor no conjunto  $\mathcal{L}$ . É fácil ver que o conhecimento dos conjuntos  $\mathcal{L}$  e  $\mathcal{L}^*$  é suficiente para que a função  $g$  esteja bem definida. Dada a função  $g$  e os conjuntos  $\mathcal{L}$  e  $\mathcal{L}^*$  fica possível definir a função  $\mathcal{F}$  ponderada através do formato:

$$\mathcal{F}w_\ell = \frac{\sum_{q=1}^Q \left( \frac{g(LLR_q)}{Q} \right) \mathbb{1}_{\{\ell; \ell \in \hat{\xi}_q\}}}{Q}$$

em que  $q \in \{1, \dots, Q\}$  representa a simulação corrente para um total de  $Q$  simulações de Monte Carlo,  $\hat{\xi}_q$  é o zona espacial mais verossímil na  $q$ -ésima execução,  $\mathbb{1}_A$  é função indicadora para o conjunto  $A$  e função  $g$  é a função anteriormente definida.

Um estudo experimental através de dados simulados foi proposto. O mapa com dados de câncer de tireoide no estado do Novo México/EUA foi utilizado, como os dados eram compostos por 20 instantes temporais, um único desses instantes foi considerado para fornecer uma distribuição populacional sobre o mapa do Novo México/EUA. Partindo dessa distribuição de população, 1,000 casos simulados foram distribuídos através de simulações de Monte Carlo. Para cada área  $\ell$  do mapa, foi considerado uma variável aleatória Poisson com taxa  $\mu_\ell$ . O interesse era a escolha de taxa que levassem a produção de um cenário com a existência de um *cluster* artificial em algum ponto específico porém conhecido do mapa. Posteriormente o interesse seria verificar a habilidade da nova forma de definição da função  $\mathcal{F}$  para delinear a existência desse *cluster* artificial no mapa em estudo. O procedimento do aumento do risco relativo nas regiões pertencentes ao *cluster* artificial seguiu o procedimento já bem estabelecido definido por Kulldorff, Tango e Park (2003). Para cada execução de simulação, o risco relativo associado aos indivíduos não pertencentes ao *cluster* artificial será igual a um. Já para os indivíduos pertencentes à população do *cluster*



As medidas usuais de comparação entre métodos de detecção de *clusters* espaciais são as medidas de Sensibilidade e Valor Preditivo Positivo. Entretanto, como já discutido anteriormente, o procedimento da função  $\mathcal{F}$  não é um procedimento de detecção de *clusters*, mas sim um procedimento para o delineamento da formação de possíveis *clusters* no mapa em estudo. Esse formato apresenta uma escala que condiciona as regiões entre regiões com maior e menor chance de pertencer a um possível *cluster* que exista no mapa em estudo. Dessa forma, não existe pontualmente um *cluster* detectado pelo método para que se faça possível o cálculo de Sensibilidade e Valor Preditivo Positivo.

Para contornar esse problema a mesma estratégia utilizada no estudo espaço temporal com *clusters* artificiais será abordada aqui. Ao final de um procedimento simulado da função  $\mathcal{F}$ , defina como *cluster* estimado  $\hat{\xi}_q$  o conjunto das regiões tais que a função  $\mathcal{F}$  se encontra entre os  $\alpha \times 100\%$  maiores valores verificados para a função  $\mathcal{F}$  na referida sequência de execuções ( $\alpha$  é um nível de significância previamente escolhido, associado à execução do teste de hipóteses que determina pela tomada de decisão a favor ou não da existência de um *cluster* no mapa em estudo). Defina ainda, o *cluster* real (artificial) por  $\xi$ . Dessa forma, as medidas Sensibilidade e Valor Preditivo Positivo (VPP) passam a ser calculáveis através do seguinte formato:

$$\text{Sensibilidade} = \frac{\text{população}(\hat{\xi}_q \cap \xi)}{\text{população}(\xi)}$$

$$\text{VPP} = \frac{\text{população}(\hat{\xi}_q \cap \xi)}{\text{população}(\hat{\xi}_q)}$$

Considerando esse conjunto de medidas para *cluster* artificialmente construído os dois métodos apresentaram os resultados da tabela 4.

A formulação ponderada apresentou um ganho tanto em Sensibilidade quanto em Valor Preditivo Positivo. Não se trata de uma evolução altamente significativa, mas é um resultado suficiente para induzir uma maior investigação sobre a

Tabela 4: Avaliação espacial através de medidas de desempenho

	Sensibilidade	VPP
função $\mathcal{F}$	0.683	0.937
função $\mathcal{F}$ ponderada	0.753	0.963

utilização dessa nova proposição. O ganho é mais efetivo em termos de Sensibilidade, isso mostra que a nova técnica tende a descrever melhor uma grande porção do *cluster* que possivelmente existe no mapa em estudo.

Um segundo parâmetro de comparação foi a produção de mapas com escala de cores comparando o delineamento ao longo de todas as execuções fornecido através de cada uma das duas propostas a função  $\mathcal{F}$  original e a função  $\mathcal{F}$  ponderada. Esses mapas podem ser visualizados na figura 23.

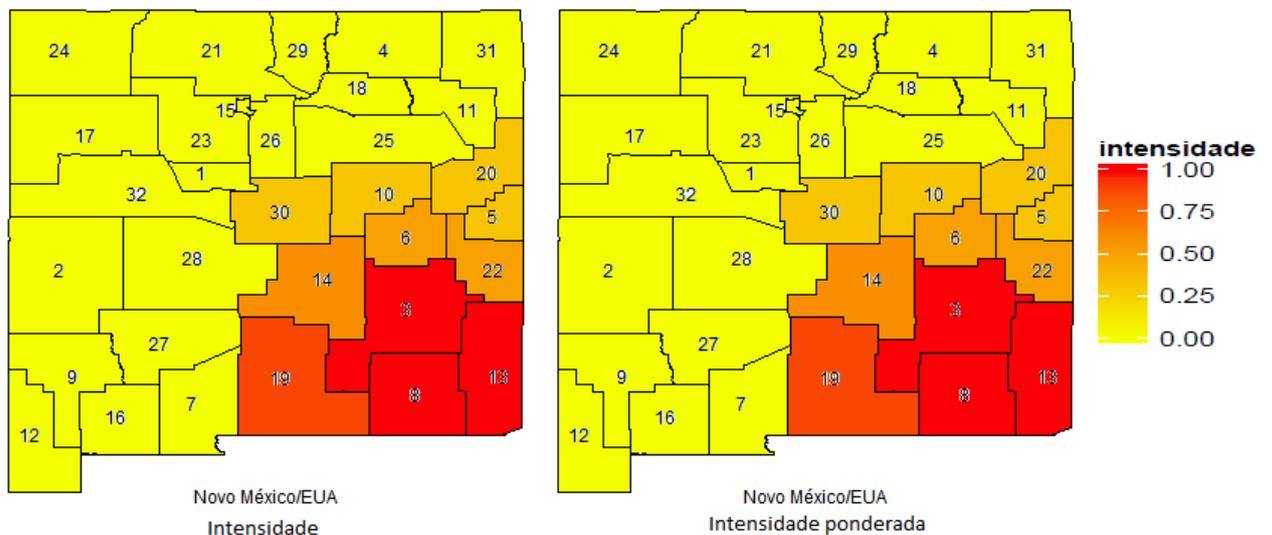


Figura 23: Comparação entre os delineamentos propostos.

A visualização através da escala de cores mostra o delineamento que a metodologia faz para a formação do possível *cluster* existente no mapa em estudo. A primeira constatação é óbvia, os dois métodos se mostraram extremamente hábeis para delinear o *cluster*. Visualmente a escala de cores foi incapaz de diferenciar os dois métodos. Ao comparar a média da função  $\mathcal{F}$  e da função  $\mathcal{F}$  ponderada ao longo dos 1000 procedimentos executados, dentre todas as regiões nas quais o mapa em estudo estava dividido, a máxima diferença entre a Intensidade e a Intensidade pon-

derada foi inferior à 1.5%, justificando a igualdade entre as duas visualizações.

Esses resultados são promissores para um estudo espacial, fica claro que a função  $\mathcal{F}$  já está bem estabelecida, seria apenas uma estratégia de melhoria, mas o delineamento visual para a formação do possível *cluster* no mapa em estudo já é suficientemente adequado no procedimento original. Obviamente esse resultado motiva estudos posteriores que ultrapassam o escopo dessa tese. Vale lembrar que esse texto faz uma análise espaço temporal e não se dedica a aprofundar estudos puramente espaciais. Por outro lado, serve como um motivador para propor um procedimento experimental no caso espaço temporal para a função  $\mathcal{F}$  ponderada. Esse experimento será discutido posteriormente.

A mesma definição dos conjuntos  $\mathcal{L}$  e  $\mathcal{L}^*$  assim como a função  $g$  podem ser abordados para o casos espaço temporais sem nenhuma perda de generalidade. Dada a função  $g$  e os conjuntos  $\mathcal{L}$  e  $\mathcal{L}^*$  fica possível definir a função  $\mathcal{F}$  espaço temporal ponderada através do formato:

$$\mathcal{F}w_{t,\ell} = \frac{\sum_{q=1}^Q \left( \frac{g(LLR_q)}{Q} \right) \mathbb{1}_{\{\ell; \ell \in \hat{\xi}_q^{k,n}\} \cap \{t; t \geq k\}}}{Q}$$

em que  $q \in 1, \dots, Q$  representa a simulação corrente para um total de  $Q$  simulações de Monte Carlo,  $\hat{\xi}_q^{k,n}$  é o cilindro mais verossímil na  $q$ -ésima execução com base na zona espacial  $\hat{\xi}$  e altura no intervalo temporal  $[k, n]$ ,  $\mathbb{1}_A$  é função indicadora para o conjunto  $A$  e função  $g$  é a função anteriormente definida.

Para esse novo formato, ao considerar as  $Q$  simulações de Monte Carlo, quanto maior for o valor da Estatística de teste obtida maior será a contribuição para a obtenção da função  $\mathcal{F}$  espaço temporal ponderada.

De forma análoga será considerada a nova proposição para o alarme através da função  $\mathcal{F}$  dada por:

$$Aw_{t,\ell} = \frac{\sum_{q=1}^Q \left( \frac{g(LLR_q)}{Q} \right) \mathbb{1}_{\{\ell; \ell \in \hat{\xi}_q^{k,n}\} \cap \{t; t=k\}}}{Q}$$

em que  $q \in 1, \dots, Q$  representa a simulação corrente para um total de  $Q$  simulações de Monte Carlo,  $\hat{\xi}_q^{k,n}$  é o cilindro mais verossímil na  $q$ -ésima execução com base na zona espacial  $\hat{\xi}$  e altura no intervalo temporal  $[k, n]$ ,  $\mathbb{1}_A$  é função indicadora para o conjunto  $A$  e função  $g$  é a função anteriormente definida.

Os testes para verificação da eficiência dessa proposta alternativa serão realizados através do mesmo cluster artificial utilizado anteriormente para testar a eficácia da função  $\mathcal{F}$  espaço temporal e também com o mesmo volume total de simulações. Novamente serão utilizadas as medidas adaptadas de  $SSC_{k,\xi}[\tau]$ ,  $VPP_{k,\xi}[\tau]$  e  $TMA_{k,\xi}[\tau]$  obtidas para a função  $\mathcal{F}$  espaço temporal e para o alarme através da função  $\mathcal{F}$ . Os resultados podem ser observados na tabela 5.

Tabela 5: Comparação de performance

	$\delta_{t,\ell}$	$SSC$	$VPP$	$TMA$
função $\mathcal{F}$ ponderada	1.25	<b>0.537</b>	0.687	<b>4.434</b>
	1.50	0.652	0.761	<b>2.887</b>
	1.75	0.780	<b>0.973</b>	<b>1.631</b>
	2.00	<b>0.907</b>	0.893	<b>0.708</b>
	$1.3^{t-11+1}$	<b>0.887</b>	0.973	<b>0.960</b>
função $\mathcal{F}$	1.25	0.530	<b>0.721</b>	4.510
	1.50	<b>0.653</b>	<b>0.765</b>	2.910
	1.75	<b>0.791</b>	0.972	1.720
	2.00	0.883	<b>0.929</b>	0.880
	$1.3^{t-11+1}$	0.879	<b>0.974</b>	0.979

Ao considerar a tabela 5 é possível verificar uma ligeira superioridade considerando a análise através da Sensitividade Condicional do alarme  $SSC_{k,\xi}[\tau]$ , assim como uma suave inferioridade quando analisado o Valor Preditivo Positivo Condicional do alarme,  $VPP_{k,\xi}[\tau]$ . Nas duas métricas as variações verificadas são flutuações meramente marginais, não parece haver indício efetivo de que um ou outro formato seja capaz de fornecer um avanço de qualidade efetiva no processo de detecção. Por outro lado, quando avaliado o Tempo Médio de Atraso  $TMA_{k,\xi}[\tau]$  até o soar do alarme,

sistematicamente a proposta através da medida da função  $\mathcal{F}$  ponderada apresenta resultados melhores, ou seja, com um menor atraso médio para a que o alarme seja soado. Entretanto, em nenhum dos cinco casos analisados, o ganho alcançado foi suficiente para superar os resultados verificados através de outras abordagens que não a medida da função  $\mathcal{F}$  (veja Tabela 3).

Essa análise deixa claro que, apesar de bem fundamentada, essa abordagem diferente é desnecessária por não se mostrar capaz de fornecer um ganho efetivo na análise do processo de delineamento de clusters espaço-tempo através da medida da função  $\mathcal{F}$  espaço temporal. Vale ainda lembrar que o caso puramente espacial apresentou alguma evolução que deve ser investigada com maior profundidade em trabalhos que sejam propostos para estudos puramente espaciais.

## 6 CONSIDERAÇÕES FINAIS

A investigação e detecção de forma rápida e precisa do surgimento de novas doenças, a identificação de fatores de risco ambientais, ou de localizações nas quais se faz necessária a intervenção de políticas públicas, são informações de grande importância. Esse conhecimento auxilia a tomada de decisões e minimiza os gastos envolvidos, além de contribuir com o planejamento nas atuações em situações emergenciais. Dessa forma, estudos e estratégias de identificação de *clusters* espaço-temporais se fazem necessários e são de grande importância na conduta dessas e em diversas outras situações.

O ponto de partida deste trabalho foi no estudo da aplicação da Função  $\mathcal{F}$ , proposta para delineamento e análise de fronteiras geográficas, através de uma medida de plausibilidade de cada região ser parte de uma anomalia no mapa em estudo. Trata-se de uma importante ferramenta de visualização, e que de forma bastante eficaz leva em conta a variabilidade intrínseca do número de casos observados, o que geralmente não é considerado diretamente nos algoritmos existentes utilizados para detectar *clusters*.

Com o foco na detecção de *clusters* espaço-temporais, foi realizada abrangente abordagem em referencial teórico envolvendo o conteúdo em questão. Isso propiciou a construção de uma revisão bibliográfica bastante detalhada e atualizada, servindo de suporte para elaboração desse trabalho e possivelmente será de grande utilidade para pesquisadores que futuramente busquem trabalhar com assuntos de Estatística Espacial. Dentro dessa revisão bibliográfica foi feita uma abordagem mais detalhada no método de detecção de *clusters* Scan espaço-tempo, que é uma extensão da Estatística Scan Espacial, através da substituição da abordagem de varredura circular pela abordagem de varredura cilíndrica.

O método Scan espaço-tempo foi utilizado para detecção de *clusters* em sua forma prospectiva, com a utilização de dois *benchmarks* de dados reais, bem como

em um conjunto de dados simulados que foi construído a fim de aferir a qualidade apresentada pelo novo procedimento proposto. Dessa forma, foi possível comparar a nova técnica com a solução apresentada por metodologias anteriores.

A nova proposta é uma versão adaptada da função  $\mathcal{F}$  espacial para a situação espaço-temporal, ou seja, no contexto espacial o foco era o de medir a plausibilidade de que uma determinada região pertença ou não a um cluster mais verossímil, já na versão espaço-temporal, será levada em conta além da utilização da informação dos casos observados em cada uma das áreas, o tempo de ocorrência dos mesmos.

A adaptação foi executada de uma forma simples e objetiva sendo muito funcional como um instrumento de auxílio na detecção de mudanças efetivas no processo temporal e de identificação de clusters. Para isso, foi acoplado à Função  $\mathcal{F}$  uma técnica de monitoramento, via utilização de alarme, para constatação da intensidade de possíveis instantes de mudança de status no processo temporal, a fim de detectar a anomalia o mais rápido possível. Como dito anteriormente, foi proposto um novo método de monitoramento de alarme para identificação de mudança no processo do status de sob controle para fora de controle para a abordagem espaço-temporal, denominado alarme através da função  $\mathcal{F}$ .

Com interesse em aplicar a nova proposta, foram utilizados dois conjuntos de dados reais, um envolvendo casos de câncer de tireoide no estado do Novo México nos Estados Unidos durante o período de 1973 a 1992, e outro envolvendo casos de Hanseníase em menores de 15 anos no estado do Amazonas no período de 2008 a 2009.

O *benchmark* de dados de câncer de tireoide é muito utilizado em pesquisas envolvendo detecção de *clusters* espaço-temporais. Encontra-se disponível na plataforma SatScan (<http://www.satscan.org/>), e é portanto um conjunto de dados de fácil acesso, o que justifica sua enorme utilização. Outro fator relevante para sua utilização é a de que os casos de câncer de tireoide se tornaram os casos com maior

incidência dentre todos os tumores malignos nos Estados Unidos, o que incentiva buscas por métodos de detecção rápida e eficiente dessa anomalia.

Com relação ao *benchmark* de casos de Hanseníase no Amazonas, foi levado em conta a faixa etária de pessoas menores de 15 anos por ser uma faixa com fortes indicadores de risco endêmico. O mapa em questão apresenta mais regiões em estudo do que o mapa do Novo México, o que não comprometeu a qualidade da solução obtida pela Função  $\mathcal{F}$  espaço-temporal. Estudos anteriores já foram realizados nessa região, como por exemplo o estudo apresentado na proposição da técnica através da Razão de verossimilhança adaptativa, entretanto o intervalo temporal dos dados em análise era diferente do aqui utilizado.

Visando mensurar a qualidade dos resultados obtidos pela nova proposta, foram sugeridas novas métricas, levando em conta as situações de detecção das regiões pertencentes ao verdadeiro *cluster* e a exatidão da ocorrência do mesmo. Essas novas métricas foram utilizadas em um conjunto de dados simulados sobre o mapa do Novo México, sendo construído um *cluster* artificial teste para o cálculo das medidas de *SSC*, *VPP* e *TMA*.

Os resultados obtidos na execução de simulações de Monte Carlo foram bastante significativos em comparação aos métodos já existentes, a Estatística Scan espaço-temporal prospectiva e os métodos baseados na Razão de verossimilhança adaptativa. O desempenho da nova proposta se mostrou bastante eficaz. Uma melhor performance na medida de *VPP* em relação a Estatística Scan espaço-temporal prospectiva e também a técnica da Razão de verossimilhança adaptativa, o que corrobora a precisão ao detectar o *cluster* verdadeiro.

Levando em conta as outras métricas, *SSC* e *TMA*, os resultados obtidos foram bastante significativos e mesmo em situações nas quais a função  $\mathcal{F}$  mostrou uma performance inferior às outras estatísticas de varredura, seu desempenho foi notório. É importante sempre lembrar que não se trata de uma metodologia de detecção de

clusters, mas sim uma alternativa para o delineamento e visualização de regiões espaciais e instantes temporais mais ou menos relevantes para a formação de um possível *cluster* existente no mapa em estudo.

A consideração da proposta de função  $\mathcal{F}$  ponderada deixou indícios sobre a possibilidade de uma sintonia fina no processo de delineamento de possíveis *clusters*. Apesar de uma contribuição marginal quanto às métricas de avaliação, no caso espaço-tempo, o estudo inicial para *clusters* puramente espaciais destacou resultados mais convincentes. Nota-se uma maior Sensibilidade para a metodologia ponderada e ainda preservando altos índices para o Valor Preditivo Positivo. Ficou caracterizada uma necessidade de eventuais estudos futuros nessa direção, quando em análises puramente espaciais.

## 6.1 PROPOSTAS DE CONTINUIDADE

Foi desenvolvida uma adaptação da Função  $\mathcal{F}$  espacial para a situação espaço-temporal, ou seja, foi feito um delineamento da intensidade de regiões e o momento em que ocorreu a mudança de intensidade. Como propostas de aprofundamento e continuidade deste trabalho é possível listar alguns objetivos futuros:

- a busca por um aperfeiçoamento na apresentação dos resultados, através de gráficos de visualização mais eficientes;
- automatizar o processo de delineamento espaço-temporal através da criação de um pacote disponível para utilização no software estatístico *R*;
- estudos de mapeamento dos valores da função  $\mathcal{F}$  espacial e espaço-temporal como uma variável aleatória a fim de ajustar uma distribuição de probabilidades, o que possibilitaria aperfeiçoar a técnica de delineamento de regiões por intensidade medida através da função  $\mathcal{F}$ , o que poderia transformar a técnica também em uma metodologia para detecção de *clusters*.

## REFERÊNCIAS

- ABRAMS, A. M.; KULLDORFF, M.; KLEINMAN, K. Gumbel based p-value approximations for spatial scan statistics. *International Journal of Health Geographics*, v. 9, p. 61 (online version), 2010.
- ALMEIDA, A. C. L. et al. Data-driven inference for the spatial scan statistic. *International Journal of Health Geographics*, v. 10, p. 47 (online version), 2010.
- ASSUNÇÃO, R.; CORRÊA, T. Surveillance to detect emerging space-time clusters. *Computational Statistics & Data Analysis*, v. 53, n. 8, p. 2817–2830, 2009.
- ATHAS, W. F. *Investigation of Excess Thyroid Cancer Incidence in Los Alamos County*. [S.l.]: Division of Epidemiology, Evaluation, and Planning New Mexico - Department of Health, 1996.
- BOSCOE, F. P. et al. Visualization of the spatial scan statistic using nested circles. *Health Place*, v. 9, n. 3, p. 273–277, 2003.
- CANÇADO, A. L. F. et al. Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. *International Journal of Health Geographics*, v. 9, p. 55, 2010. (online version).
- COLEMAN, M. P. et al. Trends in cancer incidence and mortality. *IARC scientific publications*, n. 121, 1993.
- CRAMER, J. D. et al. Analysis of the rising incidence of thyroid cancer using surveillance, epidemiology and end results national cancer data registry. *Surgery*, v. 148(6), p. 1147–1153, 2010.
- DESANTIS, C. E. et al. Cancer treatment and survivorship statistics, 2014. *CA: A Cancer Journal for Clinicians*, v. 64, n. 4, p. 252–271, 2014.
- DUARTE, A. R. et al. Internal cohesion and geometric shape of spatial clusters. *Environmental and Ecological Statistics*, v. 17, p. 203–229, 2010.
- DUCZMAL, L.; CANÇADO, A. L. F.; TAKAHASHI, R. H. C. Geographic delineation of disease clusters through multi-objective optimization. *Journal of Computational & Graphical Statistics*, v. 17, p. 243–262, 2008.
- DUCZMAL, L. et al. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis*, v. 52, p. 43–52, 2007.
- DWASS, M. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, v. 28, p. 181–187, 1957.
- KENETT, R. S.; POLLAK, M. Data-analytic aspects of the shiryayev-roberts control chart: surveillance of a non-homogeneous poisson process. *Journal of Applied Statistics*, v. 23, p. 125–137, 1996.
- KNOX, E. G. Detection of low intensity epidemics: application in cleft lip and palate. *British Journal of Preventive and Social Medicine*, v. 17(3), p. 121–127, 1963.

- KNOX, E. G.; BARTLETT, M. S. The detection of space-time interactions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, v. 13(1), p. 25–30, 1964.
- KULLDORFF, M. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, v. 26(6), p. 1481–1496, 1997.
- KULLDORFF, M. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of Royal Statistical Society*, A, n. 164, p. 61–72, 2001.
- KULLDORFF, M. et al. Evaluating cluster alarms: A space-time scan statistic and brain cancer in los alamos. *American Journal of Public Health*, v. 88, p. 1377–1380, 1998.
- KULLDORFF, M.; NAGARWALLA, N. Spatial disease clusters: detection and inference. *Statistics in Medicine*, v. 14, p. 799–810, 1995.
- KULLDORFF, M.; TANGO, T.; PARK, P. J. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, v. 42, p. 665–684, 2003.
- LIMA, M.; DUCZMAL, L. Adaptive likelihood ratio approaches for the detection of space-time disease clusters. *Computational Statistics & Data Analysis*, v. 77, p. 352–370, 2014.
- MANTEL, N. The detection of disease clustering and a generalized regression approach. *Cancer Research*, v. 27(2 Part 1), p. 209–220, 1967.
- NAUS, J. I. Clustering of random points in two dimensions. *Biometrika*, v. 52, p. 263–267, 1965.
- NEILL, D. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting*, v. 25, n. 3, p. 498–517, 2009.
- NEILL, D. B. et al. Detection of emerging space-time clusters. In: *KDD '05 Proceedings of the eleventh ACM SIGKDD*. New York, NY: ACM, 2005. (International Conference on Knowledge Discovery in Data Mining), p. 218–227.
- OLIVEIRA, F. L. P. et al. Assessing the outline uncertainty of spatial disease clusters. In: *Public Health*. [S.l.]: InTech, 2012, (Methodology, Environmental and Systems Issues). p. 51–66.
- OLIVEIRA, F. L. P. et al. Nonparametric intensity bounds for the delineation of spatial clusters. *International Journal of Health Geographics*, v. 10, n. 1, p. 1, 2011.
- OLIVEIRA, F. L. P. et al. Border analysis of spatial clusters (submitted). 2015.
- PORTER, M. D.; BROWN, D. E. Detecting local regions of change in high-dimensional criminal or terrorist point processes. *Computational Statistics & Data Analysis*, v. 51, p. 2753–2768, 2007.
- PRATES, M. O.; KULLDORFF, M.; ASSUNÇÃO, R. Relative risk estimates from spatial and space-time scan statistics: are they biased? *Statistics in Medicine*, n. 33(15), p. 2634–2644, 2014.

- ROBERTS, S. W. A comparison of some control chart procedures. *Technometrics*, v. 8, p. 411–430, 1966.
- ROGERSON, P. A. Monitoring point patterns for the development of space-time clusters. *Jornal Royal Statistical Society*, n. 164, p. 87–96, 2001.
- SHIRYAEV, A. N. On the detection of disorder in a manufacturing process. *Theory of Probability and its Applications*, v. 8, p. 247–265, 1963.
- SONESSON, C. A cusum framework for detection of space time disease clusters using scan statistic. *Statistics in Medicine*, n. 26, p. 4770–4789, 2007.
- TANGO, T. Space-time scan statistics. In: *Statistical Methods for Disease Clustering*. [S.I.]: Springer New York, 2010, (Statistics for Biology and Health). p. 211–233.
- WHITTEMORE, A. S. et al. A test to detect clusters of disease. *Biometrika*, v. 74, n. 3, p. 631–635, 1987.
- YASUI, Y. et al. An empirical evaluation of various priors in the empirical bayes estimation of small area disease risks. *Statistics in Medicine*, v. 19, n. 17-18, p. 2409–2420, 2000.