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

DETECÇÃO E INFERÊNCIA DE CLUSTERS POR MEIO DO FLUXO DE PESSOAS

Mestrando: Francisco da Silva Oliveira Júnior

Orientador: Prof. Sabino José Ferreira Neto

Co-orientador: Prof. Ricardo Tavares

Belo Horizonte
Maio de 2012

Francisco da Silva Oliveira Júnior

DETECÇÃO E INFERÊNCIA DE CLUSTERS POR MEIO DO FLUXO DE PESSOAS

Dissertação de Mestrado apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial na obtenção do título de Mestre em Estatística.

Orientador: Sabino José Ferreira Neto
Co-orientador: Ricardo Tavares

Universidade Federal de Minas Gerais
Belo Horizonte, Maio de 2012.

Dedicatória

Dedico este trabalho as duas pessoas mais importantes da minha vida, minhas companheiras e incentivadoras. Minha mãe Sra. Consolação, responsável pela minha existência e pela minha formação, que me deu vida e que tanto me apóia e incentiva com amor e carinho e a minha amada esposa Cíntya Oliveira, razão maior do meu viver, que me auxilia e me da força para as batalhas e vitórias diárias, que em todos os momentos me enche de alegria e amor.

Agradecimentos

Primeiramente agradeço a Deus, por estar sempre presente em minha vida, em todos os momentos, por me dar força, saúde e sabedoria, para estudar, trabalhar e realizar todos os meus sonhos.

Minha querida mãe, obrigado pelo apoio incondicional, por me motivar em todos os momentos, pela paciência, carinho e compreensão. Obrigado família, por acreditarem que este desejo se realizasse, pelo suporte, pela amizade, por estarem comigo, nesta grande batalha.

Agradeço a meu grande amor, Cíntya, por sempre estar ao meu lado, pela compreensão, pelo incentivo, pelo companheirismo, pelo amor incondicional, por fazer parte da minha história, da minha vida.

Prof. Sabino, meu grande orientador, que me guiou durante todo este processo com paciência, atenção, dedicação e ensinamentos fundamentais, de forma ética e segura. Obrigado Prof. Ricardo, meu Co-orientador, que com tanta presteza me ajudou a construir o conhecimento e a aperfeiçoar este trabalho. Obrigado por acreditarem em mim!

Ao Prof. Flávio Moura eu agradeço pela grande idéia, uma das etapas mais importantes desta pesquisa. Também sou muito grato ao professor Luiz Duczmal que nos orientou em momentos complexos do trabalho.

Agradeço aos amigos da MRV Engenharia, em especial ao Thiago Caixeta, Ricardo Paixão e Leandro Rodrigues e ao time do Desenvolvimento Organizacional, pela compreensão, pelo apoio e pelo companheirismo.

Também sou muito grato aos meus queridos amigos que sempre me incentivaram nesta luta, que contribuíram com conhecimento, com tempo, com palavras, com companheirismo.

Importante destacar os professores Gregório, exemplo de profissional, Profa. Lourdes, que confiou em mim seu nome, Profa. Glaura e às secretárias Rogéria e Rose, que me deram grande suporte.

Obrigado, esta conquista é de todos nós!

Resumo:

Este trabalho propõe um método de detecção de clusters adaptando o método tradicional, Scan Circular, de forma a utilizar o fluxo de pessoas como medida de proximidade, interação entre regiões do mapa, para identificar um conjunto de regiões com risco elevado de ocorrência de um determinado evento de interesse. O fluxo de pessoas entre duas regiões é estimado através do método gravitacional, sendo diretamente proporcional aos produtos dos produtos internos brutos e inversamente proporcional ao quadrado das distâncias entre regiões. Usamos também o método linear generalizado gravitacional para estimar o fluxo através de um modelo logístico que usa como variáveis preditoras índices de desenvolvimento econômico, social e a distância. O desempenho dos métodos propostos foi comparado com o método tradicional Scan Circular a partir de simulações de clusters usando uma base de casos de homicídios e também analisando a situação real. Em todos os casos simulados, as técnicas propostas tiveram melhores resultados de poder, sensibilidade e valor preditivo positivo que o método tradicional, com exceção para clusters simulados com formato regular. Dentre as técnicas propostas, a técnica do modelo linear generalizado apresentou resultados ligeiramente superiores aos da técnica do modelo gravitacional. Na aplicação das técnicas à situação real de casos de homicídios o modelo linear generalizado gravitacional apresentou resultados mais coerentes com a realidade. Em conclusão consideramos que os métodos propostos são boas alternativas para detecção de clusters irregulares e ou desconexos.

Palavras Chave: cluster espacial; estatística espacial Scan Circular; interação entre regiões; modelos gravitacionais; detecção de clusters irregulares.

Abstract:

This work proposes a cluster detection method that adapts the traditional circular scan method, in the way how the proposed method uses the flow of people as a measure of proximity, interaction between regions of a map to identify a set of regions with a high risk of occurrence of some specific event. The flow of people between two regions is estimated by the gravitational method as proportional to the product of their gross domestic product and inversely proportional to the square of the distance between them. We also use a gravitational generalized linear model method to estimate the flow of people by a logistic model with social and economic development indices and the distance as predictor variables. The performance of the proposed methods was compared with the traditional circular scan simulating clusters from a database of real cases of homicides and also analyzing the real picture. In all simulated cases the proposed techniques overcame the circular scan with better results of detection power, sensibility and positive predictive value, except for regular shaped simulated clusters. Considering the proposed techniques the gravitational generalized linear model presented slightly better results than the gravitational model concerning the simulated clusters. When applied to the real situation of homicides cases the gravitational generalized linear model presented results more consistent with reality. In conclusion we consider that the proposed methods are good alternatives for detection of irregular and or disconnected clusters.

Key words: spatial clusters; circular spatial scan statistics; regions interaction; gravitational models; irregular clusters detection.

Lista de Figuras:

Lista de Tabelas:

Sumário:

Capítulo 1
Apresentação

1.1 Introdução:

A identificação de regiões onde existe maior incidência de algum evento, seja ele uma doença, um tipo de crime, ou até algum tipo de degradação ambiental, apresenta grande importância tendo em vista que isto poderá facilitar as ações voltadas para reduzir ou aumentar a ocorrência destes fenômenos. Sendo assim, existe grande atenção voltada para identificação de regiões de risco discrepante em relação à ocorrência de algum evento.

Dado um mapa dividido em regiões, definimos um cluster como um conjunto destas regiões onde a incidência de casos de determinado fenômeno de interesse é expressivamente maior (ou menor) do que nas demais regiões do mapa. Em outras palavras, um indivíduo estar localizado em uma das regiões do cluster apresenta um risco significativamente maior (ou menor) do que fora do cluster.

Este tipo de análise pode ser aplicado em diversas áreas do conhecimento como na epidemiologia, onde o evento de interesse pode ser a manifestação de doenças específicas, na área da segurança, na avaliação da incidência de determinados crimes, na área ambiental, no caso do monitoramento de áreas de degradação e ainda na área da biologia no caso da concentração de determinadas espécies da fauna e flora, dentre outras.

Esta avaliação pode ser reativa como no caso da investigação de alarme de alta incidência de determinado evento, proativa, quando do monitoramento contínuo de regiões com alta incidência do evento ou até etiológica na busca das características de um evento previamente desconhecido.

Esta investigação pode ser realizada de forma a localizar regiões no espaço com maior ou menor incidência de casos (cluster espacial), e pode também ser localizado considerando o tempo (cluster espaço temporal). Nesta pesquisa serão tratados casos de detecção de clusters espaciais, ou seja, localização de regiões com risco de ocorrência significativamente distinto em relação às demais.

Lawson *et al.* (1999), Moore e Carpenter (1999), Lawson (2001), Glaz *et al.* (2001), Balakrishnan *et al.* (2002) e Buckeridge *et al.* (2005) destacam que os métodos de detecção e inferência de clusters espaciais são muito importantes para as diversas áreas do conhecimento.

Dentro deste contexto torna-se necessário a utilização de técnicas estatísticas adequadas a identificação de regiões que possuem um risco discrepante em relação à ocorrência destes eventos, sendo os métodos de Estatística Espacial adequados para esta avaliação. Porém, não existe na literatura um consenso quanto à técnica mais adequada para estes problemas num contexto geral.

Os métodos de detecção de clusters espaciais, em geral, partem do princípio de que existe um mapa dividido em regiões e que, para cada uma dessas regiões, é conhecida a população sob risco e o número de casos observados para um determinado evento. Estas regiões são representadas por pontos escolhidos arbitrariamente, denominados centroides.

Estes métodos utilizam janelas móveis que varrem a área de estudo, fazendo a contagem de casos das regiões cujos centroides caem dentro da janela e comparando o número de casos observados com o número de casos esperados para esta região. O número de casos esperados é determinado considerando mesmo risco de incidência para todas as regiões.

Os testes alteram sistematicamente o tamanho das janelas e avaliam a significância estatística do risco dentro do conjunto de regiões cujos centroides pertencem a janela. Em outras palavras, se a discrepância entre o número de casos observados e o número esperado for grande o bastante estas regiões são consideradas um cluster.

O método Scan Circular (Kulldorff, 1997) apresenta-se como uma boa alternativa quando se tratam de clusters regulares, com centroides agrupados em forma próxima a um círculo. Porém, na existência de clusters irregulares, com centroides agrupados em forma diferente do círculo, esta estatística apresenta baixo poder (Duczmal *et al*. 2009), podendo subestimar ou superestimar a verdadeira região de risco de ocorrência do evento de interesse.

No trabalho de Duczmal et. al. (2009) são descritas diversas situações onde podem existir clusters irregulares, destacando-se os casos de problema de tráfego, poluição, vigilância sindrômica e outros, devido às características geográficas como presença de rios,

litoral, montanhas, etc. Os autores apresentam diversas técnicas utilizadas para melhorar a capacidade de detecção de clusters irregulares.

O critério de proximidade entre regiões utilizado no método Scan Circular é a distância euclidiana entre os centroides das regiões. Porém, sabe-se que a distância euclidiana é só uma das formas de se medir a proximidade entre regiões e nem sempre é a mais adequada.

Existem situações em que o fluxo de pessoas, por exemplo, pode ser utilizado como medida de proximidade. Dentre estas, destaca-se a investigação de ocorrência de doenças contagiosas como gripe, tuberculose, sarampo, etc. Neste sentido, Duczmal e Buckeridge (2006) apresentam uma técnica de detecção de clusters que leva em consideração o fluxo de pessoas entre a residência e o trabalho como critério para determinação do cluster.

A Figura 1, apresentada a seguir, mostra um esquema contendo a distância euclidiana entre Belo Horizonte, Ouro Preto e Ouro Branco, além do fluxo estimado de pessoas entre estas mesmas regiões.

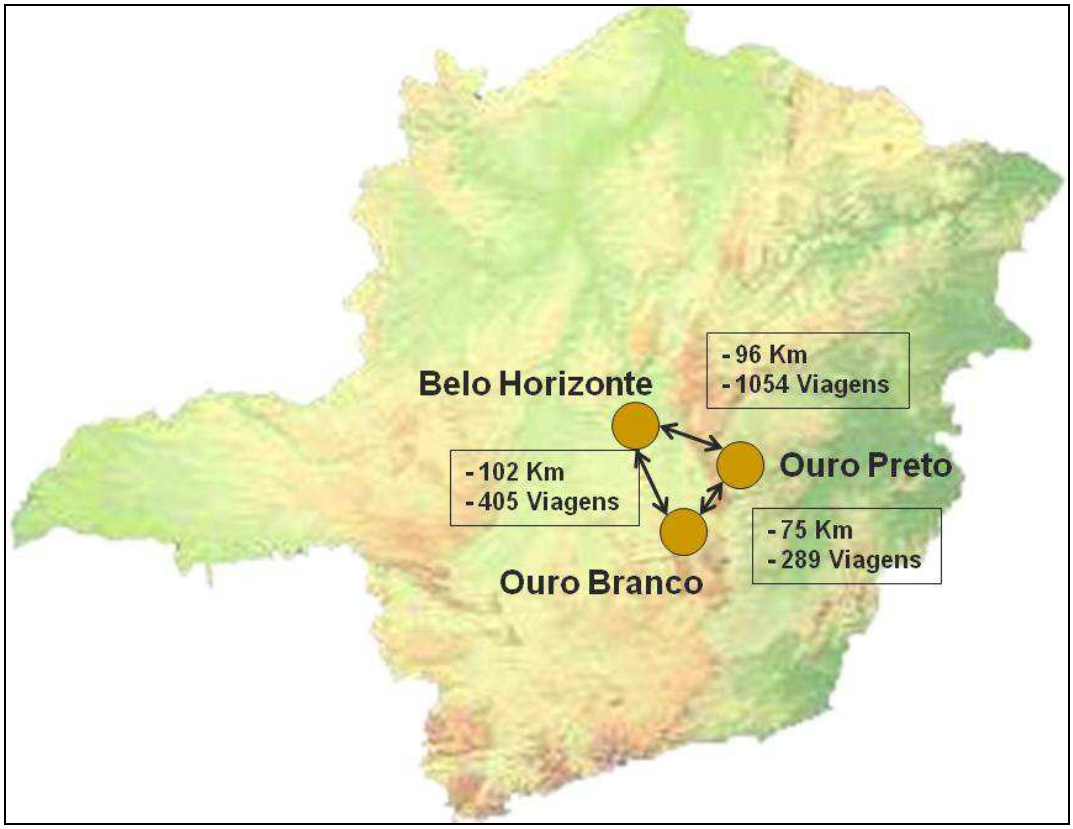

Figura 1: Distância e total de viagens entre três cidades de Minas Gerais.

A partir da análise da Figura 1, podem ser estabelecidas as seguintes questões:

- Qual a região mais próxima de Ouro Preto? Belo Horizonte ou Ouro Branco?
- Qual a influência do fluxo de pessoas no risco de um determinado evento?
- Como identificar clusters através de um método que leve em consideração o fluxo de pessoas?
- Qual a influência dos fatores socioeconômicos na interação entre regiões?
- Qual a influência dos fatores socioeconômicos no risco de um determinado evento?
- Como identificar clusters através de um método que leve em consideração informações socioeconômicas das regiões do mapa?

Este trabalho apresenta uma técnica de identificação de clusters mais apropriada para detectar regiões cujo risco seja discrepante. Procura-se também entender a influência da distância entre regiões para ocorrência de determinado evento e como isto pode impactar no poder de detecção do cluster.

Pretende-se determinar como o fluxo de pessoas impacta no risco de ocorrência de determinado evento quando este é altamente influenciado pelo contato entre pessoas e como a informação da interação entre as pessoas das regiões de interesse pode contribuir para identificação do cluster. Para tanto, serão realizados estudos de simulação e aplicação a um conjunto de dados reais, para os casos de homicídios e casos de ocorrência de gripe (influenza A). A seguir serão apresentados os objetivos do estudo.

1.2 Objetivos:

1.2.1 Objetivo Geral:

Propor uma metodologia de detecção de clusters que utilize o fluxo de pessoas como uma medida de proximidade entre regiões do mapa, ou seja, utilizando uma nova estrutura de vizinhança, de forma a identificar um conjunto de regiões com risco elevado de ocorrência do evento de interesse com maior eficiência.

1.2.2 Objetivos Específicos:

- Adaptar o método Scan Espacial de forma a considerar o fluxo de pessoas como critério de proximidade entre regiões;
- Comparar o desempenho da técnica proposta através de simulações com o método Scan Circular tradicional que usa a distância entre os centroides das regiões como medida de proximidade;
- Comparar o método tradicional com a proposta no caso de situações reais.

Este trabalho está organizado da seguinte forma: O Capítulo 2 repassa o histórico dos principais métodos de detecção de clusters até o presente momento. Já o Capítulo 3 mostra a metodologia tradicional e a proposta para utilização da interação entre pessoas como critério para estimar a proximidade entre regiões. O Capítulo 4 apresenta os resultados do método tradicional e o método proposto neste trabalho. Estes resultados estão divididos em avaliações numéricas via simulações Monte Carlo e aplicações a dados reais de homicídios e gripe influenza. No Capítulo 5 é realizada a discussão e conclusão do trabalho e o Capítulo 6 apresenta as propostas para trabalhos futuros.

Capítulo 2
Revisão de Literatura

Dentro do contexto de localização de regiões de maiores incidências de casos, o primeiro método para detectar clusters foi proposto por Choynowsky (1959), baseado em "*quadrats*" para estudar a distribuição espacial de casos de tumores na Polônia. Este método consistia em avaliar cada área individualmente, ocorrendo o problema de múltiplos testes e incapacidade de detecção de clusters que não seguissem as delimitações geográficas.

Anos depois, surgiu o Geographical Analysis Machine (GAM), desenvolvido por Openshaw *et al*. (1987). Esta técnica se baseou na idéia de Choynowsky, utilizando aqui múltiplos círculos de raio *R*, sobrepostos, de forma a permitir a detecção de clusters com formatos aproximadamente circulares.

Turbull *et al*. (2009) apresentaram o Cluster Evaluation Permutation Procedure (CEPP), que também utiliza zonas circulares, sendo os círculos construídos de forma a apresentar a mesma população *P*. Bessag e Newell (1991), desenvolveram o teste TBN, onde o número de casos *K* é definido como o tamanho do cluster a ser procurado.

Já em 1995, Tango desenvolveu o teste C_λ onde o tamanho do cluster é determinado por λ que é um parâmetro de escala de uma função que mede a proximidade entre áreas pertencentes ao conglomerado.

Porém, de acordo com Moura (2006), estes métodos apresentam um grande problema na definição à priori do parâmetro que caracteriza o tamanho do cluster: No GAM, o raio *R* do conglomerado; no CEPP, o raio *P* populacional; no TBN o raio *K* de casos e no C_λ o parâmetro λ. Como os clusters são desconhecidos, é necessário realizar o teste repetidas vezes usando diferentes valores dos parâmetros, numa espécie de tentativa e erro, podendo existir vício de pré-seleção. Além destes problemas, existe ainda a questão dos múltiplos testes, podendo ser o nível de significância maior que o pré-estabelecido.

Kulldorff (1997) propôs a Estatística Scan Espacial para detectar áreas com elevada taxa de incidência, baseando-se nas ideias do GAM e CEPP, permitindo uma avaliação global dos resultados.

A estatística de varredura espacial que utiliza a janela circular (Método Scan Circular) apresenta resultados muito bons quando os clusters são formados por um conjunto de regiões com formato aproximadamente circular (Kulldorff, 1997).

No entanto, quando se trabalha com clusters com formato irregular (diferentes do circular), este método apresenta algumas deficiências, podendo subestimar ou superestimar o cluster, com o teste apresentando baixo poder.

Além da irregularidade do cluster pode ocorrer ainda o fato de existirem soluções com alta razão de verossimilhança, entretanto, dentre as regiões desta solução, algumas podem ter sido obtidas através de uma junção, sem restrições, de regiões com elevado risco no mapa.

Estas zonas, assim formadas, podem se espalhar por todo mapa fazendo com que o significado geográfico do possível cluster fique prejudicado. Diante deste fato, alguns autores sugeriram métodos de penalização por irregularidade dos clusters.

Tendo em vista estas dificuldades, Ronald e Murray (2001) propuseram a utilização da razão de verossimilhança ponderada, sendo os pesos um fator de correção para o teste, por irregularidade.

Em 2004, Duczmal e Assunção apresentaram uma nova proposta, utilizando uma estrutura de grafo no mapa e o método Simulated Annealing (SA), sendo os clusters determinados como zonas conectadas com formato irregular. Desta forma, o método permitiu a detecção de áreas com padrão geométrico arbitrário, com poder não muito menor que o Scan para áreas circulares.

Duczmal *et al*. (2006) apresenta a penalização por compacidade geométrica, onde as zonas que possuem área muito irregular são penalizadas. Moura (2006) apresenta a penalização por ocupação circular, comparando a área do cluster identificado com o menor círculo centrado em um dos centroides das regiões contidas no cluster e que contém os centroides das demais regiões do cluster. Yiannakoulias *et. al.* (2007) mostrou a penalização por não conectividade.

Estas medidas de penalização podem ser utilizadas como multiplicador da razão de verossimilhança, reduzindo o valor da mesma proporcionalmente a não regularidade dos clusters, para evitar a detecção de clusters sem significado geográfico. Uma revisão bibliográfica completa deste assunto pode ser encontrada em Duczmal *et al*. (2009).

Considerando as restrições do método Scan Circular, pode-se utilizar o fluxo de pessoas como uma medida de interação entre regiões, tendo em vista que determinados eventos podem se espalhar pelo mapa através do contato ou interação entre as pessoas das regiões que fazem parte deste mapa.

Neste sentido, Duczmal e Buckeridge (2006) apresentam em seu artigo uma técnica de detecção de clusters que leva em consideração o fluxo de pessoas como critério para determinação do cluster. Eles descrevem uma extensão do método tradicional, a estatística Scan Espacial, que usa informações do fluxo de pessoas entre a residência e o trabalho para pesquisar o cluster de uma determinada doença, resultante da exposição no local de trabalho.

O objetivo destes autores é detectar clusters em situações onde a exposição a uma determinada doença ocorre no local de trabalho, onde temos somente a informação do endereço residencial destas pessoas para avaliação e identificação das regiões de risco.

Todos os trabalhos citados apresentam novas propostas de correção da estatística Scan Espacial para tentar um aumento da qualidade de detecção de cluster quando este não é formado por regiões conexas do mapa de interesse. Porém, em nenhum destes casos, houve a proposta de se alterar o conceito de interação entre regiões, que pode estar fortemente associado à ocorrência dos eventos de interesse como doenças, crimes, devastação, etc.

Este trabalho propõe uma alteração no método Scan Circular tradicional, no sentido de se trabalhar com uma medida de interação entre regiões que seja mais informativa, que leve em consideração o fluxo de pessoas entre as regiões do mapa e não somente a proximidade física entre as mesmas.

A este método daremos o nome de Scan Espacial Adaptado e a medida de interação entre as regiões será o fluxo de pessoas, estimado através do modelo Gravitacional. O modelo Gravitacional (Signorino, 2011) estima a força do fluxo de pessoas entre regiões de um mapa através de medidas econômicas e sócio-demográficas das regiões de interesse como por exemplo, o Produto Interno Bruto (PIB), o Índice de Desenvolvimento Humano (IDH) e a *renda per capita* (RPC).

Capítulo 3
Metodologia

3.1 Introdução:

Este capítulo se propõe a descrever a metodologia de trabalho utilizada nesta pesquisa de forma a responder aos seus objetivos. Para tanto, será utilizado o método Scan Circular de Kulldorff, técnica tradicionalmente utilizada e que apresenta melhores resultados quando existem clusters regulares.

Será apresentado também o método Scan Espacial Adaptado, que utiliza o fluxo de pessoas como medida de proximidade, sendo este estimado através do método de Modelos Gravitacionais.

Posteriormente, serão apresentados os métodos de comparação das técnicas tradicional e modificada em termos de qualidade de detecção de clusters, através de medidas como o poder de detecção, sensibilidade e valor de predição positivo (PPV).

3.2 Estatística Scan Espacial de Kulldorff:

3.2.1 Método Scan Circular

Vários métodos têm sido desenvolvidos para detecção de clusters, mas, em geral, o Scan Circular têm se mostrado mais eficiente que os demais, principalmente nos casos em que o cluster apresenta formato regular. Kulldorff (1997) propôs a Estatística Scan Espacial para detectar áreas com elevada taxa de incidência. A estatística de varredura espacial que utiliza a janela circular (Scan) apresenta resultados muito bons quando se tratam de clusters formado por regiões regulares.

Considere um mapa dividido em m regiões, onde cada região R_i tem uma população N_i e um número C_i de casos. Seja N a população total e C o total de casos do mapa. Uma zona z é definida como um subconjunto conexo de regiões do mapa, sendo o conjunto de

todas as zonas denotado por *Z*. Neste contexto, um cluster é definido como a zona mais verossímil dentre todos os possíveis subconjuntos conexos do mapa em estudo.

Considerando o modelo de Poisson, o número de casos em cada região C_i é uma variável aleatória com distribuição de Poisson com parâmetro λ_i que representa o número esperado de casos.

Assim, $\lambda_i = p_i \cdot N_i$, onde p_i é a probabilidade de um indivíduo da região R_i vir a ser um caso. Com isto, $C_i \sim Poisson(\lambda_i = p_i \cdot N_i)$. Quando o risco de ocorrência do evento é o mesmo para todas as regiões do mapa, a probabilidade de ocorrência *p* é estimada por *C/N*.

Por outro lado, a hipótese alternativa supõe a existência de pelo menos uma zona $z \in Z$, onde a probabilidade *p* de um indivíduo ser um caso seja maior (ou menor) do que a probabilidade *q* de um indivíduo em outra região do mapa fora de *z* ser um caso. Assim, temos:

$$\begin{cases} H_0 : p = q \\ H_a : p > q \; ou \; p < q, z \in Z \end{cases} \tag{1}$$

Seja *L(z)* a verossimilhança sob a hipótese alternativa e L_0 a verossimilhança sob H_0. Considerando o número esperado de casos dentro da zona *z*, sob H_0 como:

$$\lambda_z = C \diagup N \cdot N_z \tag{2}$$

Assim, a razão das verossimilhanças é dada por (Kulldorff, 1997):

$$LR(z) = \begin{cases} \dfrac{L(z)}{L_0} = \left(\dfrac{C_z}{\lambda_z}\right)^{C_z} \left(\dfrac{C - C_z}{C - \lambda_z}\right)^{C - C_z}, \; se \; \dfrac{C_z}{\lambda_z} > 1 \\ 1 \qquad\qquad\qquad\qquad\qquad\qquad , \; caso \; contrário \end{cases} \tag{3}$$

Como o logaritmo é uma função estritamente crescente, se $\hat{z}$ maximiza *LR(z)*, então $\hat{z}$ maximiza *LLR(z)* que é definido como o logaritmo da razão de verossimilhança. Assim, para cada uma das zonas *z*, teremos uma *LLR(z)*. A Estatística Scan Espacial é definida como:

$$T = \max_{z \in Z} \{LLR(z)\} \quad (4)$$

A zona $\hat{z}$ que maximiza *T* será a zona mais verossímil.

Fixado um centroide no mapa, inicia-se com um raio (distância entre o centroide e o centroide mais próximo) e assim, sucessivamente, deve-se aumentar seu raio até atingir um determinado percentual da população total do mapa.

A estatística Scan Espacial *T* é definida como a maior verossimilhança observada no conjunto de todas as janelas circulares com raios variáveis centrados no centroide de cada uma das regiões.

3.2.2 Significância Estatística

A significância estatística é avaliada a partir da comparação da estatística de teste encontrada com a distribuição empírica obtida por simulações de Monte Carlo, realizadas sob a hipótese nula. Essa simulação consiste em construir milhares de réplicas do mapa original em que o número total de casos *C* é distribuído aleatoriamente por todas as regiões sob H_o. Para cada réplica teremos um valor da estatística T e o conjunto delas obtido pela simulação gerará uma distribuição empírica da estatística de teste. A Figura 2 descrita a seguir mostra uma ilustração da distribuição empírica e a estatística de teste.

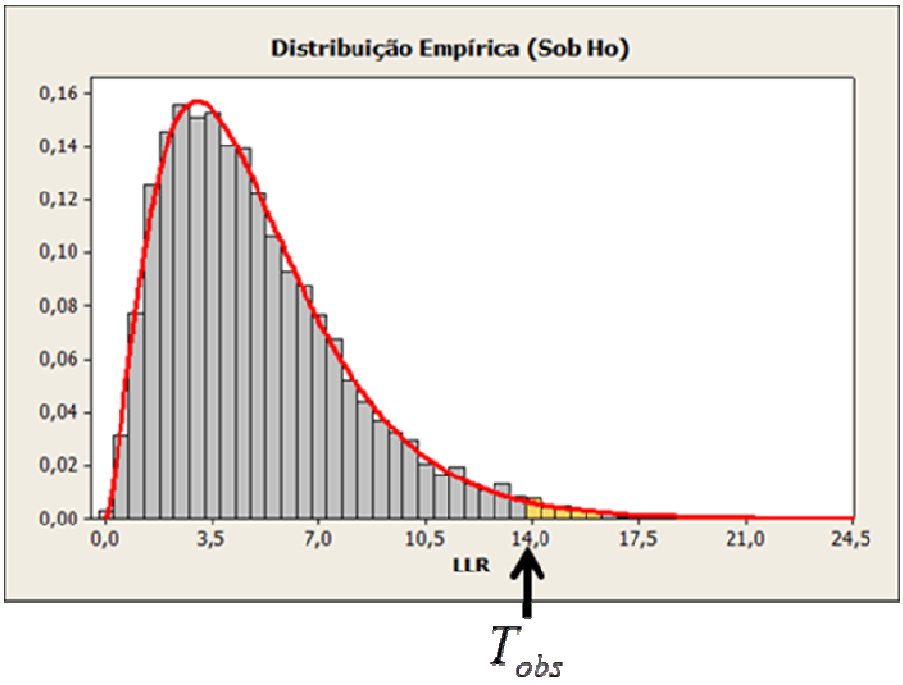

Figura 2: Distribuição empírica de T, sob H_o.

Perceba que os valores simulados representam a distribuição empírica da estatística de teste, sob a hipótese nula. A significância do cluster obtido é representada pela probabilidade de ocorrência de valores maiores ou iguais ao observado. Assim, se T_{obs} é um valor altamente provável se comparado à distribuição sob H_o, conclui-se que o cluster obtido não é significativo, ou seja, conclui-se pela não existência de clusters.

Por outro lado, se T_{obs} tem baixa probabilidade de pertencer ao conjunto de dados sob a hipótese nula, então se conclui que existe uma zona com risco de ocorrência do evento diferente das demais, ou seja, que o cluster obtido é significativo.

3.2.3 Algoritmo Scan Circular

Esta seção trata da apresentação do algoritmo utilizado para obtenção da estatística Scan Espacial em seu formato original. Inicialmente será discutida a lógica utilizada para detecção, segundo o método de Kulldorff (1997) e logo após será apresentado o algoritmo. A Figura 3 mostra o esquema da lógica da metodologia de análise.

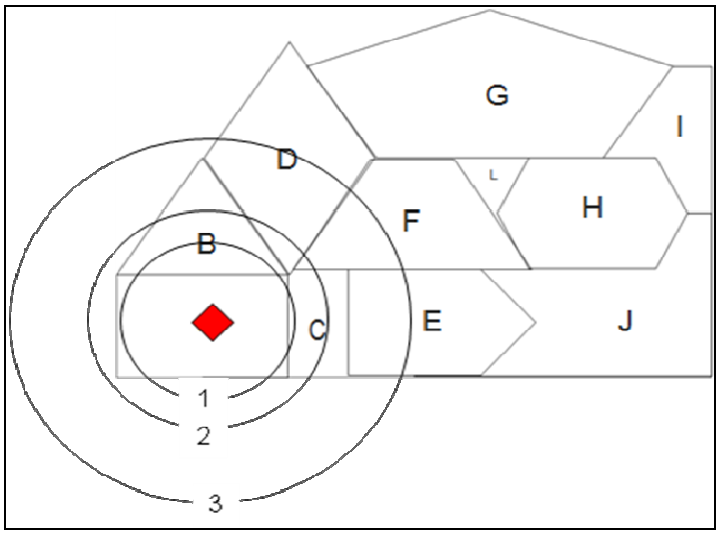

Figura 3: Descrição do procedimento de varredura do método de Kulldorff.

Considere um mapa dividido em 11 regiões, a saber A, B, C, ... , L. As letras estão alocadas no centróide das regiões. Os círculos representam a janela de varredura considerada e o índice dos círculos, a ordem de varredura.

Por exemplo, inicialmente escolhe-se uma região, a região A. Para esta região é realizado o cálculo da verossimilhança, de acordo com o número de casos e a população da região escolhida. Num primeiro passo, considerando o círculo 1, temos que a janela circular representa o subconjunto formado pelas regiões A e B, para o qual é calculada novamente a verossimilhança. Já em uma segunda etapa, a região C é incluída no subconjunto anterior, formando um novo subconjunto para um novo calculo da verossimilhança. A escolha destas regiões é definida segundo a distância euclidiana e o raio dos círculos representa a distância entre a região A e as outras regiões avaliadas.

Assim, o método consiste em representar as regiões, determinadas pelas letras, e escolher uma dentre todas. Calcula-se a distância entre o centroide da região escolhida e as demais, da mais próxima até a mais distante. A região escolhida inicialmente é determinada como uma zona. Calcula-se a estatística de teste para zona selecionada, considerando a população sob risco e o número de casos dentro desta região.

O próximo passo consiste em selecionar a primeira região mais próxima da que foi tratada no primeiro passo. A proximidade é determinada em termos da distância euclidiana

existente entre as regiões. Agora a zona a ser avaliada é determinada pelo conjunto da região selecionada de forma aleatória e a primeira região mais próxima.

A população sob risco e o número de casos são atualizados para o cálculo da verossimilhança. Este passo é repetido até que a zona represente um percentual, definido a priori, da população total do mapa. E em seguida uma segunda região é escolhida aleatoriamente, repetindo-se os mesmos passos até que a população da zona represente o percentual determinado como regra de parada. Esta lógica é apresentada a seguir, na forma de algoritmo.

Início:

1. Represente cada uma das regiões no mapa pelo seu centroide;
2. Para toda região do mapa faça;
3. Calcule as distâncias entre o centroide escolhido no passo 2 e os centroides das outras regiões, ordenando-as em de forma crescente, e guardando-as em um vetor;
4. Crie um círculo centrado no centroide da região escolhida no passo 2 e continuamente aumente o seu raio de acordo com as distâncias encontradas no passo 3. Considere a zona *z* como o conjunto das regiões cujos centroides estejam dentro do círculo. Para cada região que entrar no círculo, atualize C_z, N_z e calcule *LLR(z)*. Registre a zona de maior *LLR(z)* até o momento;
6. Registre a zona com maior *LLR(z)*;
7. Utilize simulações de Monte Carlo para construir a distribuição empírica da estatística de teste, sob Ho;
8. Avalie a significância do teste a partir do *LLR(z)* encontrado.

Fim.

Como mencionado anteriormente, quando se trabalha com clusters de formatos irregulares, este método apresenta algumas deficiências, podendo subestimar ou

superestimar o verdadeiro cluster, apresentando baixo poder, de acordo com o apresentado pela Figura 4 e 5, descritas a seguir.

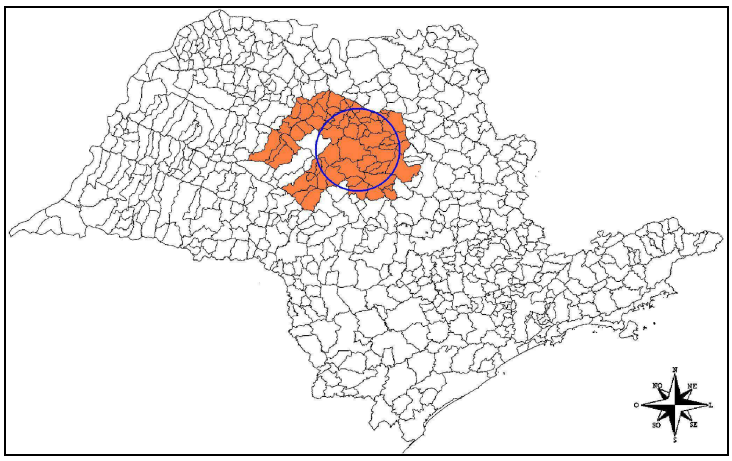

Figura 4: Subestimação do cluster.

Perceba, de acordo com a Figura 4, que as regiões que fazem parte do cluster identificado pela janela circular representam somente uma parte das regiões do verdadeiro cluster (hachurada), ou seja, o cluster real pode ser subestimado devido a sua irregularidade.

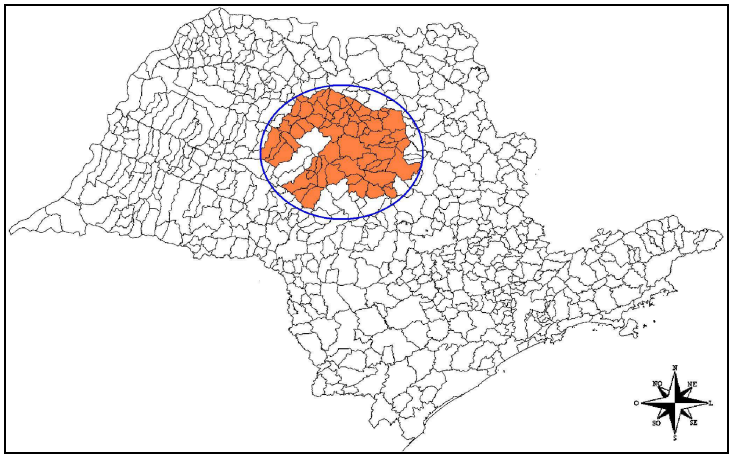

Figura 5: Superestimação do cluster.

A partir da Figura 5 é possível observar que as regiões do verdadeiro cluster representam somente uma parte das regiões do cluster identificado, ou seja, o cluster real está sendo superestimado devido a sua irregularidade. Além disto, esta técnica não

possibilita a detecção de clusters formados por regiões desconexas, ou seja, que não possuem fronteira geográfica comum.

3.3 Fluxo de Pessoas:

Existem situações onde o fluxo de pessoas representa de forma mais adequada a proximidade entre regiões, no sentido de que quanto maior o fluxo de pessoas, mais "próximas" estarão as regiões. Por exemplo, no caso de doenças contagiosas, a interação entre as pessoas de diferentes regiões pode ser um fator importante para a proliferação do número de casos.

De acordo com esta nova visão, quanto maior é o fluxo entre pessoas mais próximas estarão as regiões. Por isto, será utilizado o inverso do fluxo como uma medida de proximidade entre as regiões. Assim, a distância entre duas regiões será representada pelo inverso do fluxo.

Neste sentido, torna-se importante determinar uma forma de quantificar o fluxo de pessoas entre as regiões de interesse. Este pode ser representado pelo número de pessoas que vão de uma região para outra, ou até mesmo pelo número de ônibus, carros, aviões, etc, que saem de uma determinada área com destino à outra região.

O número de viagens de ônibus pode ser uma medida importante para estimar o fluxo entre pessoas. A Figura 6, descrita a seguir, mostra a matriz do fluxo de ônibus e a matriz de proximidade. Perceba que não serão considerados fluxos internos, tendo em vista que o método escolhe inicialmente uma região e avalia a verossimilhança somente para os casos dentro desta região.

Tabela 1: Fluxo de pessoas:

Cidades	A	B	C	D
A	-	50	30	20
B	50	-	40	30
C	30	40	-	50
D	20	30	50	-

$$Fluxo(i, j) = N^{o}\ viagens\ entre\ i\ e\ j \quad (5)$$

Tabela 2: Inverso do fluxo de pessoas:

Cidades	A	B	C	D
A	-	0,020	0,033	0,050
B	0,020	-	0,025	0,033
C	0,033	0,025	-	0,020
D	0,050	0,033	0,020	-

$$D_f(i,j) = \frac{1}{Fluxo(i,j)} \qquad D_f(i,i) = 1 \tag{6}$$

A proximidade é avaliada segundo o fluxo de pessoas entre regiões distintas. Assim, o inverso do número de viagens (ida e volta) será utilizado como medida do grau de proximidade entre regiões. A medida do fluxo de pessoas entre cidades pode ser estimada pelo número de viagens de ônibus intermunicipais, por exemplo.

Estas informações foram levantadas junto ao DER-MG, que forneceu o número de viagens de ônibus de ida e volta existente entre todas as cidades do Estado de Minas Gerais, tanto da Região Metropolitana de Belo Horizonte quanto os intermunicipais de dezembro de 2010.

A partir deste levantamento criou-se a matriz de fluxos, que é representada em formato de matriz sintetizada, mostrando o número de viagens entre as Cidades de Minas Gerais, tendo em vista que a matriz completa, de acordo com os 853 municípios de Minas Gerais, contém 853 linhas e 853 colunas.

Tabela 3: Distribuição do número de viagens entre os municípios de Minas Gerais.

Cidades	Abadia D.	. . .	BH	BeloOrien.	BeloVale	Berilo	Berizal	Bertópolis	Betim	. . .	Wenc.Braz
Abadia D.	-		0	0	0	0	0	0	0		0
⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
BH	0		-	0	68	0	0	62	35723		0
BeloOrien.	0		0	-	0	0	0	0	0		0
BeloVale	0		68	0	-	0	0	0	0		0
Berilo	0		0	0	0	-	0	0	0		0
Berizal	0		0	0	0	0	-	0	0		0
Bertópolis	0		62	0	0	0	0	-	0		0
Betim	0		35723	0	0	0	0	0	-		0
⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
Wenc.Braz	0		0	0	0	0	0	0	0		-

Fonte: DER/MG 2010

Perceba que para a grande maioria dos pares de municípios origem/destino, não existe um fluxo de ônibus, segundo os dados disponibilizados pelo DER/MG referente ao ano de 2010. Desta forma, torna-se necessário utilizar técnicas de imputação ou estimação para determinar o fluxo de pessoas de forma mais precisa. Dentre as possibilidades, destacam-se 4 opções:

- **Tratar o inverso do fluxo como sendo igual a 1:**

Quando o fluxo entre regiões é igual a 0 (nulo). O inverso do fluxo tem que ser o maior possível, ou seja, 1 que é o resultado máximo possível para o inverso dos fluxos.

Esta técnica tem o problema de não levar nenhuma outra informação em consideração e da grande quantidade de valores iguais a 1 que serão definidos arbitrariamente .

- **Fluxo Médio Entre Vizinhos:**

Considerando duas cidades que possuem fluxo zero, esta técnica consiste em calcular o fluxo médio entre uma Cidade e as Vizinhas da outra Cidade e vice-versa, retirando a média dos resultados obtidos ao final. Desta forma o fluxo estimado seria representado pelo fluxo médio entre as Cidades e seus vizinhos mais próximos (Ohmann, 2011).

Esta técnica é considerada mais coerente que o fluxo igual a 1, pois leva em consideração informações adicionais trazidas pelo fluxo entre os vizinhos. Porém, a questão a ser respondida aqui é qual o número de vizinhos a serem considerados para determinação do fluxo médio.

- **Critério Bayesiano:**

O critério Bayesiano pode ser utilizado no sentido de corrigir o fluxo zero de uma certa região, a partir do fluxo médio das regiões vizinhas, principalmente aquelas com

população pequena. Marshall (1991) propõe um estimador Bayesiano empírico para corrigir a taxa levando-se em consideração o efeito da variabilidade entre os tamanhos populacionais das áreas que constituem um mapa. A ideia é corrigir a taxa observada (bruta) para se obter uma nova taxa (Bayesiana) que seja mais precisa quando a população for pequena.

Ainda aqui, outras informações importantes não são levadas em consideração como, por exemplo, o desenvolvimento econômico e social das regiões, o que pode influenciar de forma decisiva na determinação do fluxo de pessoas.

- **Modelo Gravitacional:**

Os Modelos Gravitacionais são técnicas de estimação do fluxo de pessoas, ou melhor, da força de interação entre regiões, através de informações econômicas e sóciodemograficas como, por exemplo, o PIB (Produto Interno Bruto), a *renda per capita* e o IDH (Índice de Desenvolvimento Humano), dentre outras. Este método atribui maior força entre relações cujos resultados de desenvolvimento econômico e sóciodemograficos são significativos, sem deixar de levar em consideração a distância entre regiões, que também é uma medida muito informativa sobre a interação entre elas (Signorino *et al*. 2011).

Dentro do modelo gravitacional pode-se utilizar o método de regressão logística de forma a estimar a força de interação entre regiões, representada por uma probabilidade, que é estimada através de medidas sóciodemograficas e do fluxo entre regiões. A seção a seguir apresenta o detalhamento desta técnica.

3.4 Modelos Gravitacionais: Estimação do Fluxo de Pessoas

A interação espacial é um termo amplo que abrange qualquer movimento no espaço. Esta interação espacial pode se dar através de uma viagem ao trabalho, migração, informação sobre fluxos de mercadorias/serviços, etc. Os modelos gravitacionais são os

tipos mais utilizados de modelos de interação. Eles são formulações matemáticas utilizadas para analisar e prever os padrões de interação espacial (Signorino *et al*. 2011).

A mobilidade no sentido de fluxo de pessoas é um tipo de interação espacial cuja análise geralmente depende da aplicação de esquemas do tipo gravidade que considera a quantidade de interações entre territórios distantes como uma função crescente de suas forças econômicas e como uma função decrescente das distâncias que os separam. Formalmente, os modelos gravitacionais para interações espaciais foram desenvolvidos a partir das teorias de Newton e baseiam-se na expressão:

$$T_{ij} = \frac{A(i)B(j)}{C(d_{ij})} \qquad (7)$$

Em que T_{ij} representa a força de interações entre as locações *i* e *j*, ou seja, o fluxo entre as regiões. As funções *A(i)* e *B(j)* representam as ponderações não especificadas relacionadas com a origem e destino, respectivamente, que contem atributos relativos às suas massas econômicas, e $C(d_{ij})$ é definida como uma função de distância que conta o efeito de distância *d* em *T*. As especificações de *A(i)* e *B(j)* são estabelecidas de acordo com as suposições associadas aos parâmetros/variáveis do modelo.

Considere a seguinte situação, uma cidade A, situada a 50 km da cidade B e 70 km da cidade C, fisicamente está mais próxima à cidade B. Porém, o PIB das três cidades são respectivamente, R$ 100 milhões, R$ 200 milhões e R$ 300 milhões. Considerando que o conceito do modelo gravitacional, de acordo com a fórmula (7), o fluxo entre as cidades A e C é maior, por estas juntas possuírem maior desenvolvimento econômico, sendo este um fator importante para demonstrar a interação entre as cidades.

Importante ressaltar que o PIB é apenas uma das possíveis formas de representar o desenvolvimento econômico, ou a força de interação entre regiões, e que podem ser utilizadas como medida de interação, no lugar do fluxo, outras medidas como o tamanho das populações, o desenvolvimento socioeconômico, representado pelo IDH, dentre outros.

O modelo de gravidade considera pelo menos dois elementos básicos: (1) impactos de escala: por exemplo, cidades com grandes populações, ou grande desenvolvimento socioeconômico tendem a gerar e atrair mais atividades que cidades com pequenas

populações; (2) impactos de distância: por exemplo, em lugares mais distantes, as pessoas/atividades poderão estar mais separadas ou próximas.

Signorino *et al*. (2011) destacam que para este contexto, não há ainda um consenso geral sobre a especificação formal da relação distância-decaimento, sobre o uso de funções potências ou exponenciais, equações (8) e (9) a seguir:

$$C(d_{ij}) = d_{ij}^{\beta} \tag{8}$$

$$C(d_{ij}) = \exp(\beta\, d_{ij}) \tag{9}$$

Sendo β é um parâmetro positivo que mede o efeito da distância. Segundo Signorino *et al.* (2011) as funções exponenciais são mais apropriadas para modelar interações de curta distância (ex.: mobilidade urbana local), enquanto as especificações via funções de potência são mais recomendadas para interações de longa distância (ex.: fluxos de migração).

Ultimamente, estudos envolvendo análise de riscos para populações que residem/trabalham em áreas contaminadas tem se destacado. Signorino *et al*. (2011) aplicaram um modelo gravitacional para analisar o risco residencial de trabalhadores no complexo petroquímico de Gela (Sicília, Itália), principal fonte de poluição industrial na região.

Um modelo de regressão logística foi ajustado para medir a capacidade do pólo de Gela atrair fluxos de comunidade de outros locais. A metodologia adotada foi definir a probabilidade de encontrar comunidades de municípios fora de Gela com uma função relacionada à economia das áreas origem e das distâncias entre cada área destino. Através de modelo linear generalizado controlando a idade e o período, compararam-se taxas de mortalidade nesses locais e confirmaram um maior risco de câncer associado com residência em Gela.

Neste trabalho, utilizaremos dois modelos de interações gravitacionais para estimar o fluxo entre regiões: o primeiro utilizando a fórmula de proporcionalidade, descrita em (7) e o segundo utilizando o modelo de regressão logística, como na aplicação descrita por Signorino *et al*. (2011).

3.4.1 Modelo Gravitacional Tradicional:

O modelo de interação gravitacional utilizando a fórmula de proporcionalidade será estimado através de uma medida de desenvolvimento econômico entre regiões e uma medida de distanciamento, seguindo a lei da atração gravitacional, de Newton.

A proximidade será diretamente proporcional ao desenvolvimento econômico, a saber o PIB de cada uma das regiões e inversamente proporcional à distância entre as mesmas. Como sugerido por Signorino *et al*. (2011) na fórmula (8), no caso de interações de longa distância, que é o que este estudo busca representar o valor do parâmetro β será 2, de forma a seguir de forma fiel a lei de atração dos corpos de Newton. Assim, a força de interação T_{ij} entre as regiões *i* e *j*, será estimada através da fórmula (4), descrita a seguir:

$$T_{ij} = \frac{PIB(i)PIB(j)}{d_{ij}^{\,2}} \tag{10}$$

Os resultados obtidos a partir da equação (10) serão utilizados como medida do fluxo de pessoas entre as regiões *i* e *j*. Assim, o fluxo entre as regiões *i* e *j* será determinado de forma proporcional ao peso que esta interação representa dentre todas as interações.

Seja F_t o número total de viagens de ida e volta entre as Cidades de Minas Gerais, o fluxo será determinado por:

$$F_{(grav)ij} = F_t \left[\left(T_{ij} \right) \Big/ \sum_{ij} \left(T_{ij} \right) \right] \tag{11}$$

No modelo tradicional de Kulldorff, esta medida será utilizada para representar a força de interação entre as regiões ao invés da distância euclidiana. Esta proposta será denominada de Estatística Scan Adaptada Gravitacional (GRAV). A Tabela 4, descrita a seguir mostra os resultados de força do fluxo, estimados a partir do modelo GRAV.

Tabela 4: Distribuição da força de interação (fluxo) entre regiões de Minas Gerais a partir do modelo GRAV.

Cidades	Abadia D.	. . .	BH	BeloOrien.	BeloVale	Berilo	Berizal	Bertópolis	Betim	. . .	Wenc.Braz
Abadia D.	-		13,190	0,112	0,035	0,022	0,009	0,011	11,027		0,019
⋮		⋱									⋮
BH	13,190		-	71,784	799,335	17,045	9,667	8,631	6234,079		21,715
BeloOrien.	0,112		71,784	-	0,720	0,495	0,311	0,204	41,970		0,136
BeloVale	0,035		799,335	0,720	-	0,032	0,014	0,011	619,124		0,045
Berilo	0,022		17,045	0,495	0,032	-	0,096	0,179	11,885		0,008
Berizal	0,009		9,667	0,311	0,014	0,096	-	0,064	6,129		0,003
Bertópolis	0,011		8,631	0,204	0,011	0,179	0,064	-	6,665		0,003
Betim	11,027		6234,079	41,970	619,124	11,885	6,129	6,665	-		17,753
⋮										⋱	⋮
Wenc.Braz	0,019		21,715	0,136	0,045	0,008	0,003	0,003	17,753		-

A seguir será feita a descrição do método GLM gravitacional.

3.4.2 Modelo GLM gravitacional

O segundo modelo de interação gravitacional, será construído a partir de uma medida de proximidade que é uma função que leva em consideração o desenvolvimento econômico das regiões, representado pelo PIB, do desenvolvimento social das regiões, representado pelo IDH e da distância entre regiões.

Para construção do modelo gravitacional, utilizando estes parâmetros para estimar a força de interação entre as regiões, será utilizado um modelo de regressão logística, sendo o PIB, o IDH e a distancia as variáveis explicativas e a resposta representada pelo fluxo, sendo este igual a 1 quando existem viagens entre duas regiões e 0 quando não existem.

Seja P_{ij} a força de interação entre os municípios i e j, e 1-P_{ij} a força de não interação entre i e j. Para estimar a força da interação, será utilizado um modelo linear generalizado com função de ligação Binomial, a regressão logística binária. A variável resposta, dicotômica foi construída a partir de uma matriz de fluxo de ônibus de acordo com o que foi mencionado anteriormente.

O PIB e o IDH serão representados pela média aritmética das medidas obtidas para as duas regiões avaliadas, como descrito em (12) e (13).

$$PIB_{ij} = \frac{PIB(i) + PIB(j)}{2} \tag{12}$$

$$IDH_{ij} = \frac{IDH(i) + IDH(j)}{2} \tag{13}$$

A expressão do modelo de regressão logística é apresentada em (14).

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_0 + \beta_1 d_{ij} + \beta_2 PIB_{ij} + \beta_3 IDH_{ij} \tag{14}$$

Os coeficientes do modelo são representados por β_i que são os parâmetros da regressão logística. A estimação destes é feita através do método de máxima verossimilhança. Neste estudo, estes parâmetros representam o peso com que cada uma das características influenciam o fluxo de interações entre as regiões. Assim, β_0 é o intercepto, β_1 representa a influência da distância, β_2 representa a influência do PIB e β_3 a influência do IDH.

As medidas obtidas para o PIB e distância entre as regiões apresentam grande dispersão, prejudicando a estimação dos parâmetros do modelo de regressão logística. Para minimizar o efeito negativo da dispersão na estimação dos parâmetros, será utilizado o logaritmo, que é uma função estritamente crescente, que reduz consideravelmente a variabilidade destas informações sem prejudicar a ordenação destas medidas. A expressão do modelo final de descrita em (15).

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 \log\left(d_{ij}\right) + \beta_2 \log\left(PIB_{ij}\right) + \beta_3 IDH_{ij} \tag{15}$$

A partir do modelo descrito pela equação (14) e das características de cada uma das regiões estudadas, foi estimado o modelo de regressão logística, utilizando o método de máxima verossimilhança, através do software R, versão 2.13.0. A Tabela 5, descrita a seguir apresenta os resultados do modelo de regressão logística, estimado a partir dos dados de fluxo, PIB, IDH e distancia entre as Cidades do estado de Minas Gerais.

Tabela 5: Estimação dos parâmetros do modelo.

Coeficiente	Estimativa	Erro Padrão	P-valor
(β_0) – *Intercepto*	-23,283	0,806	< 0,001
(β_1) - $log(d_{ij})$	-2,416	0,041	< 0,001
(β_2) - $log(PIB_{ij})$	1,067	0,024	< 0,001
(β_3) - IDH_{ij}	6,124	1,162	< 0,001

A partir da análise da Tabela 5 é possível observar que todas as características utilizadas para estimar a força da interação entre as regiões, apresentam efeito significativo, sendo as probabilidades de significância menores que 0,001.

Assim, a força de interação entre quaisquer duas regiões estudadas pode ser obtida através da substituição das estimativas obtidas para os coeficientes na expressão (14) e inserindo a média dos valores das características de PIB, IDH das regiões *i* e *j* e distância entre as mesmas.

Obtemos portanto outra medida da força de interação entre as regiões, representada por P_{ij}. Este fluxo será determinado de forma proporcional ao peso que esta interação representa dentre todas as interações. Da mesma forma, seja F_t o número total de viagens de ida e volta entre as Cidades de Minas Gerais, o fluxo $F_{(glm)\,ij}$ será determinado por:

$$F_{(glm)ij} = F_t \left[\left(P_{ij} \right) \Big/ \sum_{ij} \left(P_{ij} \right) \right] \tag{16}$$

Esta medida também será utilizada para representar a força de interação entre as regiões, no método chamado de Scan Circular Adaptado Gravitacional GLM (GLM). A Tabela 6 apresenta um resumo da matriz de força de interação obtida a partir do método GLM.

Tabela 6: Distribuição da força de interação (fluxo) entre regiões de Minas Gerais a partir do modelo GLM.

Cidades	Abadia D.	. . .	BH	BeloOrien.	BeloVale	Berilo	Berizal	Bertópolis	Betim	. . .	Wenc.Braz
Abadia D.	-		16,220	0,061	0,021	0,010	0,003	0,004	11,276		0,011
⋮											⋮
BH	16,220		-	91,564	516,184	17,278	6,617	6,028	660,306		27,636
BeloOrien.	0,061		91,564	-	0,537	0,290	0,124	0,078	44,276		0,074
BeloVale	0,021		516,184	0,537	-	0,015	0,004	0,003	452,780		0,030
Berilo	0,010		17,278	0,290	0,015	-	0,038	0,085	9,678		0,003
Berizal	0,003		6,617	0,124	0,004	0,038	-	0,020	3,631		0,001
Bertópolis	0,004		6,028	0,078	0,003	0,085	0,020	-	3,426		0,001
Betim	11,276		660,306	44,276	452,780	9,678	3,631	3,426	-		18,881
⋮											⋮
Wenc.Braz	0,011		27,636	0,074	0,030	0,003	0,001	0,001	18,881		-

A seguir serão apresentados o método Scan Circular adaptado ao fluxo de pessoas, a ser utilizado para o GLM e o GRAV.

3.5 Método Scan Circular Adaptado ao Fluxo de Pessoas:

Como apresentado na seção 3.2, o método Scan Circular tradicional se inicia com a seleção de uma das regiões do mapa, para a qual é calculado o número de casos e a população sob risco. A partir destas informações, é obtido o valor da estatística de teste. O próximo passo é selecionar a região mais próxima geograficamente da primeira região escolhida, formando assim uma nova zona e calculando novamente a estatística de teste. Este processo se repete até que certo percentual da população total sob estudo seja atingido.

Aqui serão apresentados os métodos Scan Circular adaptados ao fluxo de pessoas. Perceba que para o método tradicional, a medida de proximidade entre regiões, critério utilizado na construção das janelas circulares é o grau de proximidade, representado pela distância euclidiana entre as regiões, ou seja, a distância euclidiana entre uma região e outra.

A estatística Scan Espacial adaptada ao fluxo de pessoas apresenta as mesmas características do método Scan Circular tradicional, porém, a medida de proximidade entre regiões passa a ser representada pelo fluxo de pessoas. De acordo com esta nova visão,

quanto maior é o fluxo de pessoas entre regiões do mapa, maior será a proximidade entre as mesmas. Em outras palavras, será criado um novo grafo baseado nos fluxos.

De acordo com o método Scan tradicional apresentado anteriormente, uma zona é definida como um conjunto de regiões conexas do mapa. Já para o método adaptado a definição de zona também se modifica, sendo agora determinada por um conjunto de regiões do mapa, não sendo necessário o atendimento ao critério de conectividade entre regiões, ou seja, qualquer região poderá ser vizinha da outra.

A matriz de distâncias cartesianas, utilizada para a realização da varredura das regiões do mapa no método tradicional será substituída pela matriz do inverso do fluxo e esta forma representará a proximidade entre regiões.

O método Scan Adaptado deverá realizar a varredura, agrupando a área escolhida inicialmente com regiões que possuem o maior fluxo, na ordem decrescente até o menor fluxo, sendo realizados os cálculos da estatística de teste, em cada passo. Este processo deverá ser repetido até se atingir um determinado percentual da população do mapa. Desta forma, o algoritmo de varredura será o mesmo utilizado no método Scan tradicional, com a substituição da matriz de distâncias pela matriz do fluxo e a ordem passa a ser decrescente. A Figura 6, descrita a seguir sintetiza a lógica do método Scan Circular Adaptado ao fluxo de pessoas.

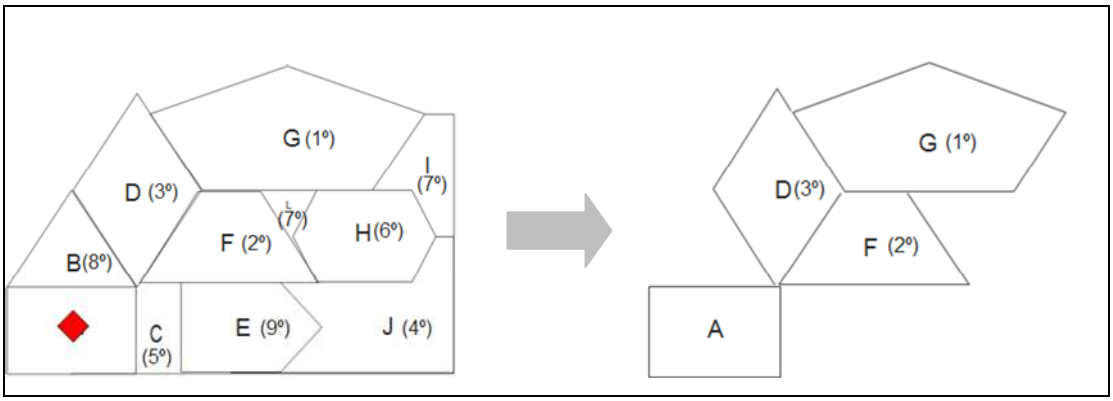

Figura 6: Descrição do procedimento de varredura do método de Kulldorff Adaptado.

Aqui também temos 11 regiões representadas pelas letras A, B, C, ..., L. As letras representam os centroides das regiões e os índices representam a ordem de interação destas

regiões com a região escolhida inicialmente, ou seja, a ordenação decrescente do fluxo da região com a região escolhida.

Por exemplo, inicialmente escolhe-se uma região de forma aleatória, a região A para a qual é calculada a estatística de teste. Num primeiro passo, considerando que a região mais próxima é a região G (1º), temos assim um subconjunto formado pelas regiões A e G, realizando assim o cálculo da estatística de teste. Já em uma segunda etapa, a região F (2º) é incluída no subconjunto anterior, formando um novo subconjunto e calculando a estatística de teste, tendo em vista que esta é a segunda região mais próxima da região A e conseqüentemente a terceira região mais próxima que passa a integrar o subconjunto.

A partir destas modificações, existirá maior chance de detecção de clusters que são formados por regiões que não estão próximas fisicamente, mas que ainda sim possuem grande interação, ou seja, grande fluxo de pessoas.

Como a matriz de proximidade não depende da distância euclidiana entre as regiões, o cluster detectado não terá que ser necessariamente composto por regiões conexas e nem terá que apresentar formato regular, que é uma das limitações do método tradicional. A seção a seguir apresenta as técnicas utilizadas para comparar o método tradicional com o método adaptado.

Para implementação dos métodos propostos, utilizou-se o software C++, seguindo o algoritmo descrito anteriormente para os dados da pesquisa, tanto na simulação, quanto na aplicação a dados reais.

3.6 Técnicas de Avaliação da Eficiência dos Métodos:

Esta seção apresenta a metodologia utilizada para realizar as comparações entre o método tradicional, Scan Circular, e o método Scan Adaptado. Para avaliar a eficiência em termos de detecção de clusters serão utilizadas medidas de poder, sensibilidade e valor de predição positiva. Para o mapa real serão produzidos clusters artificiais ou sintéticos, que denominaremos de cluster real.

Para geração dos clusters artificiais, em cada uma das hipóteses alternativas ou escolha do cluster real, o número total de casos (mantido fixo) é distribuído aleatoriamente

entre as regiões do mapa de acordo com uma distribuição multinomial. Nesta distribuição, o risco de um individuo ser um caso é maior que um para regiões escolhidas para pertencer ao cluster artificial e igual a um para as demais regiões do mapa. O método utilizado para calcular o risco relativo do cluster artificial é descrito a seguir (Kulldorff *et al*. 2003).

Seja p_z a população sob risco do cluster e P a população total do mapa. Para um número total de casos C, o número de casos observados c_z, do cluster z, tem distribuição binomial com parâmetros (C, τ_z), com $\tau_z = p_z / P$, sob a hipótese nula de que não existem clusters no mapa. A média e a variância são dadas por:

$$m_0 = C\frac{p_z}{P} \qquad e \qquad v_0 = C\frac{p_z(P-p_z)}{P^2} \tag{15}$$

Considerando a aproximação normal para a distribuição binomial, o número crítico de casos k para que um teste unilateral rejeite a hipótese nula com o nível de significância $0 < \alpha < 1$ é tal que:

$$\Phi\left(\frac{k-m_0}{\sqrt{v_0}}\right) = 1-\alpha \quad \longrightarrow \quad \frac{k-m_0}{\sqrt{v_0}} = \Phi^{-1}(1-\alpha) \tag{16}$$

Onde $\Phi(.)$ é a função de distribuição acumulada da Normal Padrão. Se α=0,05, teremos que $\Phi^{-1}(1-\alpha)$=1,645, Assim, o valor crítico k é tal que $(k-m_0)/v_0$=1,645. Sob a hipótese alternativa, considerando um risco ρ_z, para as regiões do cluster, o número total de casos Cz do cluster tem distribuição Binomial com média e variância dados por:

$$m_a = \frac{C\ p_z\ \rho_z}{(P-p_z+p_z\rho_z)} \qquad v_a = \frac{C\ p_z\ \rho_z(P-p_z)}{(P-p_z+p_z\rho_z)^2} \tag{17}$$

Observe que neste caso teremos:

$$\tau_z = \frac{p_z\ \rho_z}{(P-p_z+p_z\rho_z)} \tag{18}$$

Retomando a aproximação da normal, selecionaremos um risco relativo ρ_z de forma que:

$$\frac{k - m_a}{\sqrt{v_a}} = \Phi^{-1}(\theta) \tag{19}$$

Onde $k=m_0+1{,}645\ v_0$. (20)

Assim, o risco relativo é escolhido de forma que o poder atingido por qualquer teste para clusters espaciais tenha um limite inferior igual a θ. Para este estudo, a escolha do risco relativo foi realizada de forma que se a posição exata do cluster real for conhecida, o poder de detecção deverá ser igual a 0,999. O poder de detecção de clusters é apresentado a seguir, na seção 3.6.1.

3.6.1 Poder

O poder de detecção de clusters para um dado método de detecção representa a probabilidade de o teste detectar um cluster quando este de fato existe. O cálculo do poder é realizado segundo a fórmula a seguir:

$$Poder = P(Detectar\ Cluster \mid Cluster\ Existe) \tag{21}$$

Figueiredo (2010) apresenta o procedimento utilizado para estimar o poder de detecção dos clusters para os diversos métodos utilizados. O algoritmo é descrito a seguir:

Início:

1. Condicionado no número total de casos do evento de interesse, distribuímos estes de acordo com a hipótese nula e calculamos o valor de estatística de teste. Esse procedimento é repetido milhares de vezes.

2. Estes milhares de valores são ordenados em ordem crescente e calculamos o T_{crit} igual ao percentil 95 deste conjunto de dados, considerando um nível de significância $\alpha = 0,05$.

3. Para simular um cluster artificial distribuímos o número total de casos (fixos) pelas m regiões do mapa, onde as regiões que pertencem ao cluster tem um risco relativo elevado com valor obtido como discutido anteriormente e as demais regiões do mapa tem risco relativo igual a um.

4. Para cada distribuição dos casos totais no mapa sob a hipótese alternativa, calcula-se o valor da estatística de teste T.

5. Para cada T calculado no item anterior, compara-se este valor com T_{crit}. Se $T > T_{crit}$ consideramos que o cluster detectado é estatisticamente significativo.

6. Logo, o valor estimado do poder do teste é definido como sendo a razão entre o número de vezes que o algoritmo detecta o cluster, isto é, a quantidade de vezes que $T > T_{crit}$, e o número total de simulações.

Fim.

Assim, o poder de detecção de clusters estimado para cada um dos métodos é interpretado como a proporção de vezes que o algoritmo detecta um cluster. Pode-se perceber que o poder não avalia a relação entre o cluster real e o cluster detectado, ou seja, o poder de detecção de clusters estimado para um método qualquer não avalia qual a parcela das regiões do cluster detectado que pertencem também ao cluster real. Para esta avaliação serão utilizadas as medidas de sensibilidade e do valor de predição positiva (PPV) apresentados na próxima seção.

3.6.2 Sensibilidade e PPV

Huang *et al*. (2007) adaptaram os conceitos de sensibilidade e valor de predição positiva para estatística Scan Espacial. A sensibilidade e o valor de predição positiva são duas medidas bastante utilizadas para avaliar a eficiência de detecção de clusters, tendo em vista que estas avaliam a relação entre o cluster real e o cluster detectado.

Nos últimos anos, as medidas de PPV e sensibilidade têm sido bastante utilizadas para atestar a qualidade dos métodos de detecção de clusters espaciais. Aqui estas serão utilizadas também como critério de comparação entre os métodos estudados.

Estas medidas utilizadas para avaliar a qualidade dos procedimentos de detecção de clusters, são probabilidades condicionais, definidas a partir dos seguintes eventos:

V = Indivíduo escolhido ao acaso na população do mapa pertence a população do verdadeiro cluster.

D = Indivíduo escolhido ao acaso na população do mapa pertence a população do cluster detectado.

A sensibilidade, de forma genérica, representa a probabilidade de o teste ser positivo, sabendo-se que existe a característica de interesse, ou seja, representa a proporção de resultados positivos que um teste apresenta, quando realizado em unidades conhecidamente contendo a característica de interesse.

Adaptado ao método Scan Espacial, a sensibilidade representa a proporção do cluster real que pertencem ao cluster detectado e pode ser calculada com a razão entre a população que pertence simultaneamente ao cluster real e ao cluster detectado e a população do cluster real.

$$Sens = P(D \mid V) = \frac{População(Cluster\ Detectado \cap Cluster\ \mathrm{Re}al)}{População\ do\ Cluster\ \mathrm{Re}al} \qquad (22)$$

Para um dado método de detecção de clusters, valores elevados de sensibilidade indicam que o cluster detectado ou se aproxima bastante do cluster real ou superestima o verdadeiro cluster.

Já o valor de predição positiva (PPV), de forma genérica, representa a probabilidade de o teste dar positivo dado que a característica de interesse realmente está presente, ou seja, é a proporção de resultados satisfatórios que um teste apresenta, em relação às unidades determinadas como casos.

O PPV representa a proporção de regiões do cluster detectado pelo método que pertencem ao cluster real e pode ser calculado como a razão entre a população que pertence simultaneamente ao cluster real e a o cluster detectado e a população do cluster detectado, como descrito a seguir.

$$PPV = P(V \mid D) = \frac{População(Cluster\ Detectado \cap Cluster\ \mathrm{Re}al)}{População\ do\ Cluster\ Detectado} \tag{23}$$

Para um dado método de detecção de clusters, valores elevados de PPV indicam que o cluster detectado se aproxima muito do cluster real ou que o cluster detectado subestima o cluster real.

Capítulo 4
Inferência e Resultados

4.1 Introdução

Este capítulo apresenta os resultados obtidos para as três técnicas avaliadas na pesquisa, de forma a comparar a qualidade de detecção de clusters dos métodos estudados em várias situações de interesse.

Dentre os métodos estudados, o primeiro será o método tradicional Scan Circular de Kulldorff, que utiliza a distância como medida de proximidade entre regiões. Será avaliado também o método Scan Adaptado Gravitacional, que além da distância euclidiana entre regiões, utiliza o desenvolvimento econômico, representada pelo produto dos PIB's, de acordo com os modelos gravitacionais apresentado na seção 3.4.1 (GRAV).

Além das duas técnicas citadas anteriormente, será estudado também o método Scan Adaptado Gravitacional GLM, de acordo com o modelo de regressão logística, que além da distância e do PIB utiliza o desenvolvimento social, representado pelo IDH, apresentado na seção 3.4.2 (GLM).

Para esta avaliação serão utilizados os resultados do poder, da sensibilidade, do PPV e da consistência dos resultados com as situações reais estudadas por estes modelos. Estes resultados serão avaliados através da simulação de diversas situações além da aplicação das técnicas a dois conjuntos de dados que representam situações reais.

Os resultados de simulações consistiram na determinação arbitrária de clusters, obtidos através das técnicas apresentadas na seção 3.6. Os casos foram distribuídos pelas regiões do mapa, atribuindo-se um risco maior para as regiões que fazem parte do cluster. Desta forma foi possível comparar a qualidade de detecção dos métodos avaliados em termos de poder, sensibilidade e PPV.

O poder do teste foi calculado segundo o percentual de vezes em que a técnica de detecção de clusters detecta algum cluster em relação ao total de simulações realizadas. Já a sensibilidade e o PPV foram calculados em termos da média dos valores obtidos, dentro do total de simulações realizadas.

Para avaliar a significância real dos resultados, foram obtidas bases de dados reais representando casos de Homicídios e Gripe Influenza, dentro do estado de Minas Gerais, para realizar a comparação entre a técnica tradicional e as novas propostas, de forma a avaliar os clusters detectados.

Os resultados obtidos através da utilização das três técnicas de detecção de clusters estudadas são descritos a seguir, através da seção 4.2, que mostra as simulações e seção 4.3, que mostra a aplicação a situações reais.

4.2 Resultados Obtidos por Simulação

Esta seção apresenta os resultados obtidos através de simulação para as três técnicas estudadas. Para tanto foram criadas cinco situações, considerando clusters regulares, clusters irregulares conexos, clusters desconexos, cluster regular em forma de anel e cluster formado por dois polos.

Os resultados de simulações consistiram na determinação arbitrária de clusters, obtidos através das técnicas apresentadas na seção 3.6. Para tanto, os casos foram distribuídos, atribuindo risco igual a um para regiões fora do cluster e risco igual e maior que um para as regiões que fazem parte do cluster.

Os clusters foram simulados utilizando diferentes valores para o risco de ocorrência do evento de forma a representar diversas situações. O risco calculado para a base de dados, segundo o método apresentado na seção 3.6 foi igual a 1,14 e foram utilizados riscos iguais a 1,10; 1,25; 1,50; 2,00 e 3,00.

Para cada uma das combinações de cluster simulado e risco, foram geradas 1000 simulações. A varredura percorreu até 30% da população total do mapa, tendo em vista que valores acima deste apresentavam resultados idênticos. O p-valor foi estimado através da comparação do resultado obtido com a distribuição empírica da estatística de teste sob H_o obtida por 10.000 simulações de Monte Carlo.

As medidas de qualidade, utilizadas para comparar as técnicas de detecção de clusters foram o poder, a sensibilidade e o PPV. A capacidade de detecção de clusters é caracterizada pelo poder enquanto que a subestimação é representada por uma sensibilidade

baixa e um PPV elevado. Já a superestimação é representada por um PPV baixo e uma sensibilidade elevada.

As bases utilizadas para as simulações foram o conjunto de dados de Homicídios de 2003 a 2008 em Minas Gerais (Fonte: DATASUS). A seguir são descritos os resultados obtidos para 1000 simulações de cada uma das três situações, considerando os diversos valores de risco definidos.

4.2.1 Resultados Obtidos por Simulação: Clusters Regulares

Esta seção apresenta os resultados obtidos para simulação de casos onde o cluster apresenta regularidade de forma, ou seja, possui formato aproximadamente circular, conexo. A base original de casos representa os casos de homicídios ocorridos no Estado de Minas Gerais, de 2003 a 2008. A figura a seguir apresenta o mapa contendo o cluster regular simulado.

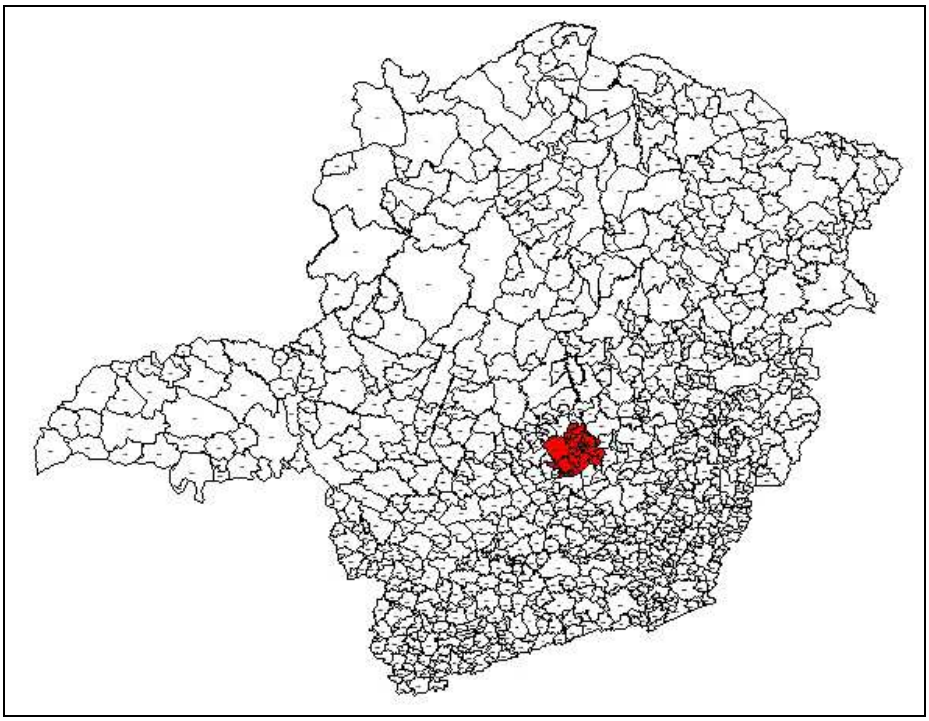

Figura 7: Cluster regular simulado.

Perceba que o cluster simulado para os dados de homicídios representa um conjunto conexo de regiões, com formato regular. Este é representado pela cidade de Belo Horizonte e mais dezenove regiões vizinhas, compondo parte da grande BH. Este cluster foi criado a partir de um cluster detectado através do método tradicional, para uma base de dados real de casos de homicídios dentro do estado de Minas Gerais.

A Tabela 7, descrita a seguir, apresenta os resultados do poder, sensibilidade e PPV, obtidos para os clusters simulados segundo várias situações de risco de ocorrência dos eventos.

Tabela 7: Resultados obtidos por simulação cluster regular.

Risco	Qualidade	TRAD	GRAV	GLM
1,10	Poder	0,987	0,984	0,984
	Sensibilidade	0,947	0,906	0,918
	PPV	0,916	0,909	0,911
1,25	Poder	1,000	1,000	1,000
	Sensibilidade	0,991	0,962	0,978
	PPV	0,987	0,951	0,967
1,50	Poder	1,000	1,000	1,000
	Sensibilidade	0,999	0,980	0,989
	PPV	0,999	0,955	0,972
2,00	Poder	1,000	1,000	1,000
	Sensibilidade	1,000	0,987	0,994
	PPV	1,000	0,956	0,971
3,00	Poder	1,000	1,000	1,000
	Sensibilidade	1,000	0,988	0,997
	PPV	1,000	0,957	0,968

Observe que, para todos os métodos de detecção de clusters utilizados, quanto maior o risco utilizado na simulação, maiores são os valores de poder, sensibilidade e PPV. Importante salientar que para riscos maiores ou iguais a 1,25, o poder de todas as técnicas é igual a 1, nível máximo desta estatística. Todas as técnicas apresentaram resultados de qualidade elevados, acima de 0,9.

A partir da análise da Tabela 3 é possível perceber que, para a simulação de clusters regulares, como apresentado pela Figura 7, o método tradicional apresenta melhores resultados de poder, sensibilidade e PPV que os métodos Scan Gravitacional (GRAV) e

Scan GLM gravitacional (GLM), independente do risco utilizado. Para o método tradicional, com riscos maiores ou iguais a 1,5, todas as medidas de qualidade já são iguais ou muito próximas do nível máximo.

As duas novas técnicas propostas apresentam resultados bem próximos ao do método tradicional, principalmente a partir de riscos iguais a 1,5. Dentre as duas novas técnicas propostas, apesar de resultados muito próximos, em geral o método GLM gravitacional (GLM), apresenta resultados mais elevados de poder, sensibilidade e PPV, se comparado com o método gravitacional (GRAV). Em relação à sensibilidade, o cluster detectado pelo método GLM se aproxima mais do cluster simulado.

4.2.2 Resultados Obtidos por Simulação: Clusters Irregulares Conexos

Os resultados obtidos para simulação de casos onde o cluster simulado é conexo e tem formato irregular, são descritos nesta seção. A base original de casos representa os casos de homicídios ocorridos dentro do Estado de Minas Gerais, em um período de 2003 a 2008. O mapa com a representação do cluster simulado é descrito a seguir pela Figura 8.

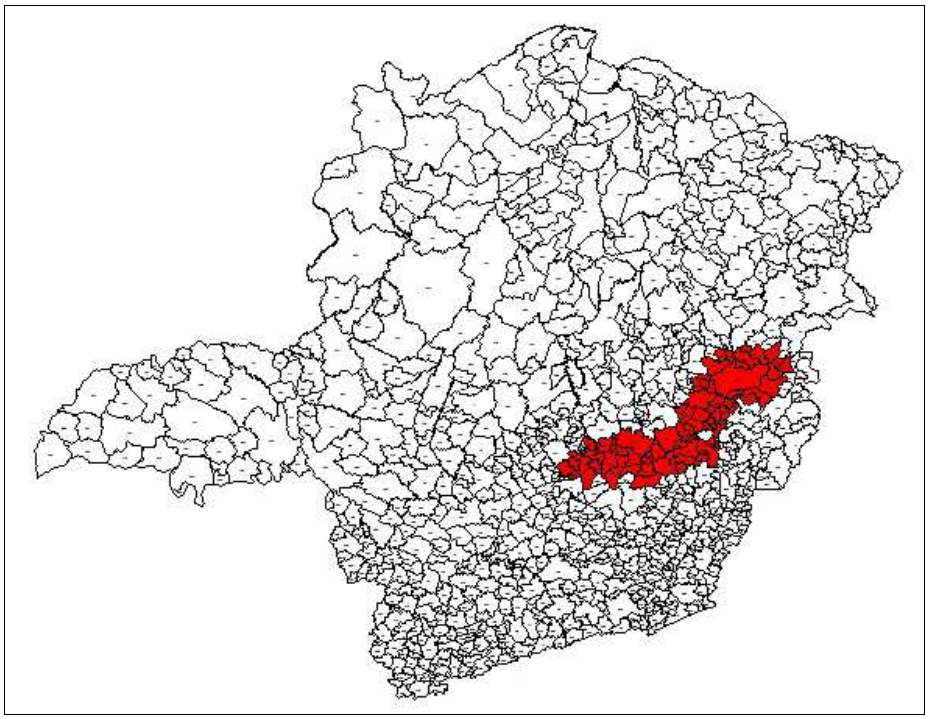

Figura 8: Cluster irregular conexo simulado.

Perceba que o cluster simulado para os dados de homicídios representa um conjunto de regiões conexo, com formato irregular. Este é representado pela estrada entre a Cidade de Belo Horizonte e a Cidade de Governador Valadares, sendo formado por 72 cidades do Estado.

A Tabela 8, descrita a seguir, mostra os resultados do poder, sensibilidade e PPV, obtidos para os clusters simulados segundo as várias situações de risco de ocorrência dos eventos.

Tabela 8: Resultados obtidos por simulação cluster conexo irregular.

Risco	Qualidade	TRAD	GRAV	GLM
1,10	Poder	0,969	0,972	0,974
	Sensibilidade	0,779	0,784	0,784
	PPV	0,898	0,916	0,922
1,25	Poder	1,000	1,000	1,000
	Sensibilidade	0,822	0,853	0,848
	PPV	0,933	0,935	0,942
1,50	Poder	1,000	1,000	1,000
	Sensibilidade	0,830	0,879	0,870
	PPV	0,939	0,929	0,940
2,00	Poder	1,000	1,000	1,000
	Sensibilidade	0,833	0,885	0,884
	PPV	0,942	0,926	0,934
3,00	Poder	1,000	1,000	1,000
	Sensibilidade	0,835	0,886	0,890
	PPV	0,942	0,925	0,930

Note novamente que, para todas as técnicas avaliadas, o poder é extremamente elevado, ficando acima de 0,96 e, a partir de riscos maiores ou iguais a 1,25, ele assume o valor máximo permitido.

Perceba que, para a simulação do cluster irregular, como apresentado pela Figura 8, os métodos propostos, em geral apresentam resultados superiores, se comparados com o método tradicional, tanto para o poder quanto para a sensibilidade e o PPV. Importante salientar que para um risco maior ou igual a 1,5, o PPV do método tradicional é mais elevado que o observado para os dois métodos propostos.

Dentre as duas novas técnicas propostas, apesar novamente de resultados muito próximos, em geral o método GLM gravitacional (GLM) apresenta resultados mais satisfatórios de poder, sensibilidade e PPV, se comparado com o método gravitacional (GRAV). O poder e o PPV do método GLM são mais elevados e a sensibilidade, em geral é mais elevada para o método GRAV.

4.2.3 Resultados Obtidos por Simulação: Clusters Desconexos

A situação onde o cluster é formado por regiões desconexas é apresentada nesta seção. A base original de casos novamente representa os casos de homicídios ocorridos no Estado de Minas Gerais, entre os anos de 2003 a 2008. A Figura 9, descrita a seguir, mostra o mapa contendo o cluster desconexo simulado.

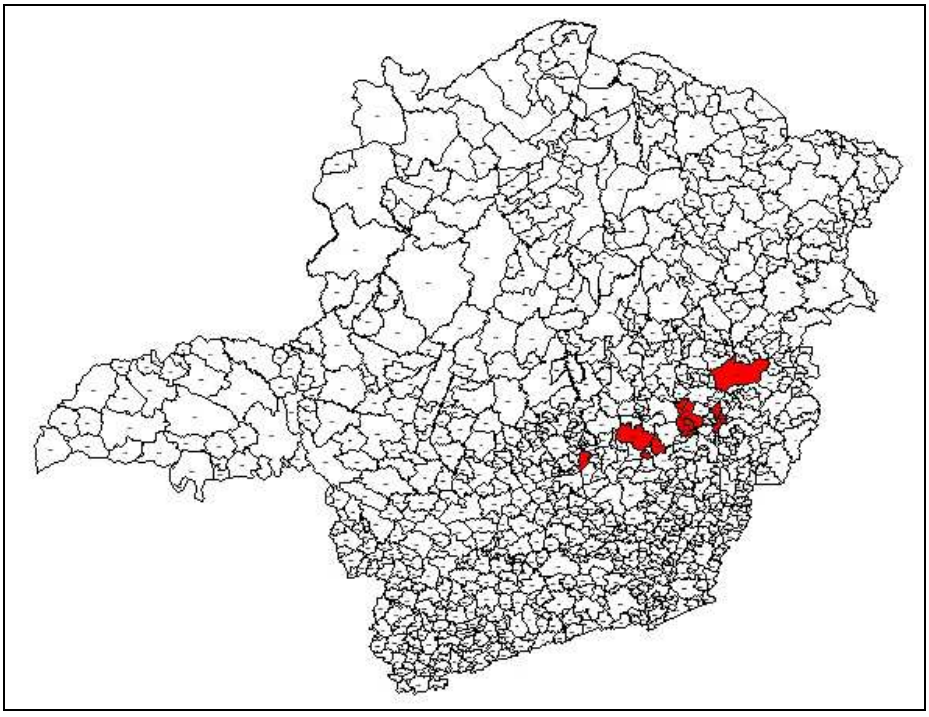

Figura 9: Cluster irregular desconexo simulado.

Para construção do cluster desconexo foram utilizadas algumas cidades que estão no caminho entre Belo Horizonte e Governador Valadares, consideradas como sendo

potenciais em termos do número de homicídios. Este cluster é formado por cidades como BH, Itabira, Coronel Fabriciano, Ipatinga e Governador Valadares, perfazendo um total de 13 cidades.

A Tabela 9, descrita a seguir, apresenta os resultados do poder, sensibilidade e PPV, obtidos para os clusters desconexos simulados segundo a Figura 9, de acordo com as várias situações de risco de ocorrência dos eventos.

Tabela 9: Resultados obtidos por simulação cluster irregular desconexo.

Risco	Qualidade	TRAD	GRAV	GLM
1,10	Poder	0,805	0,915	0,923
	Sensibilidade	0,779	0,852	0,865
	PPV	0,694	0,840	0,849
1,25	Poder	1,000	1,000	1,000
	Sensibilidade	0,827	0,898	0,915
	PPV	0,857	0,961	0,969
1,50	Poder	1,000	1,000	1,000
	Sensibilidade	0,803	0,904	0,931
	PPV	0,901	0,977	0,980
2,00	Poder	1,000	1,000	1,000
	Sensibilidade	0,828	0,903	0,940
	PPV	0,873	0,986	0,979
3,00	Poder	1,000	1,000	1,000
	Sensibilidade	0,898	0,906	0,941
	PPV	0,791	0,983	0,979

Perceba, a partir da Tabela 9, que para todos os níveis de risco de ocorrência utilizados nas simulações, as técnicas de GRAV e GLM apresentam melhores resultados de poder, sensibilidade e PPV que o método tradicional, com exceção do poder de detecção dos clusters, que para riscos maiores ou iguais a 1,25, foi igual a um para todas as técnicas avaliadas.

Isto é facilmente entendido pois se olharmos para o numerador da expressão (22) da sensibilidade e (23) do PPV, vemos que são iguais a soma das populações das regiões que pertencem simultaneamente ao cluster detectado e ao cluster real simulado. Como o Scan tradicional somente detecta subconjuntos conexos de regiões, ele somente vai detectar a parte conexa do cluster real desconexo que apresentar maior razão de verossimilhança.

Novamente, para as duas técnicas propostas pelo estudo, os resultados de poder, sensibilidade e PPV observados estão bem próximos, para todos os níveis de risco utilizados. O cluster detectado pelo método GLM se aproxima mais do cluster simulado ("real"), se comparado ao método tradicional e ao método GRAV, tendo em vista os valores de sensibilidade.

A técnica GLM gravitacional, em geral, apresenta melhores índices que a técnica gravitacional GRAV, tanto no que se refere ao poder de detecção, quanto à sensibilidade e ao PPV. Somente para os casos em que o risco de ocorrência do evento foi igual a 2 e igual a 3 é que o PPV observado para técnica GRAV apresentou melhores resultados.

4.2.4 Resultados Obtidos por Simulação: Clusters Formato de Anel

Esta seção apresenta a comparação dos resultados de simulação para situação em que o cluster real apresenta conectividade, com um formato próximo a de um anel, como se existissem regiões sem risco de ocorrência, cercada por outras várias regiões onde o risco é significativamente maior. Novamente, a base de dados original de casos representa os eventos de homicídios ocorridos no Estado de Minas Gerais, de 2003 a 2008. A Figura 10 mostra o mapa contendo o cluster simulado.

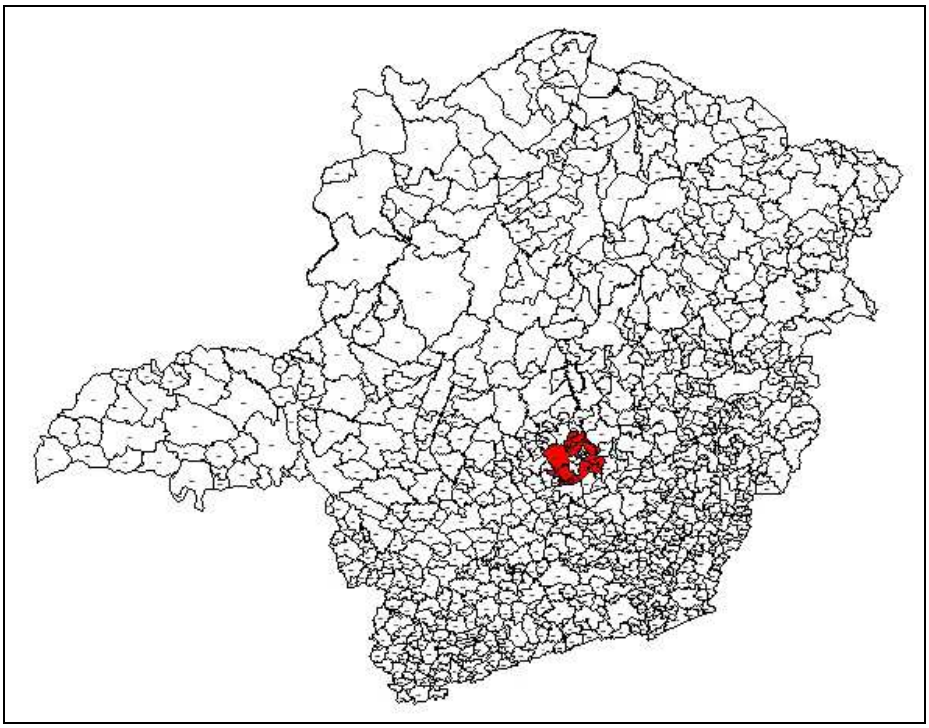

Figura 10: Cluster irregular conexo em formato de anel.

Este cluster foi criado a partir das regiões do cluster regular apresentado na seção 4.2.1, retirando-se cinco Cidades, a saber Contagem, Confins, Ribeirão das Neves, São José da Lapa e Vespasiano, perfazendo um total de 15 cidades.

Os resultados do poder, sensibilidade e PPV, obtidos para os clusters simulados, ilustrado pela Figura 10, de acordo com os vários níveis de risco de ocorrência dos eventos são descritos a seguir, pela Tabela 10.

Tabela 10: Resultados obtidos por simulação cluster formato de anel.

Risco	Qualidade	TRAD	GRAV	GLM
1,10	Poder	0,888	0,937	0,923
	Sensibilidade	0,854	0,860	0,823
	PPV	0,755	0,809	0,816
1,25	Poder	1,000	1,000	1,000
	Sensibilidade	0,891	0,930	0,866
	PPV	0,842	0,895	0,977
1,50	Poder	1,000	1,000	1,000
	Sensibilidade	0,937	0,963	0,871
	PPV	0,826	0,886	0,994
2,00	Poder	1,000	1,000	1,000
	Sensibilidade	0,990	0,994	0,874
	PPV	0,785	0,863	0,997
3,00	Poder	1,000	1,000	1,000
	Sensibilidade	0,992	0,997	0,873
	PPV	0,777	0,861	0,999

Novamente, para as duas técnicas propostas pelo estudo, os resultados de poder, sensibilidade e PPV observados estão bem próximos, para todos os níveis de risco utilizados, sendo que estas apresentam melhores resultados que o método tradicional.

Observe que para o cluster em formato de anel, o modelo GRAV e GLM apresentam melhores resultados de poder que o método tradicional, para o primeiro nível de risco, 1,1 e igual em todos os outros níveis. O modelo GRAV apresenta melhores resultados de sensibilidade, exceto quando o risco é igual a 3. Já os resultados de PPV são maiores quando é utilizado o método GLM.

4.2.5 Resultados Obtidos por Simulação: Clusters Duas Regiões Desconexas

Outra situação interessante em relação à detecção de clusters é quando existem dois polos desconexos que apresentam risco de ocorrência mais elevado em relação às outras regiões. Esta seção apresenta a comparação dos resultados de simulação para esta situação. A base de dados original de casos representa os eventos de homicídios ocorridos no Estado

de Minas Gerais, de 2003 a 2008. O cluster simulado é representado a seguir, pela Figura 11.

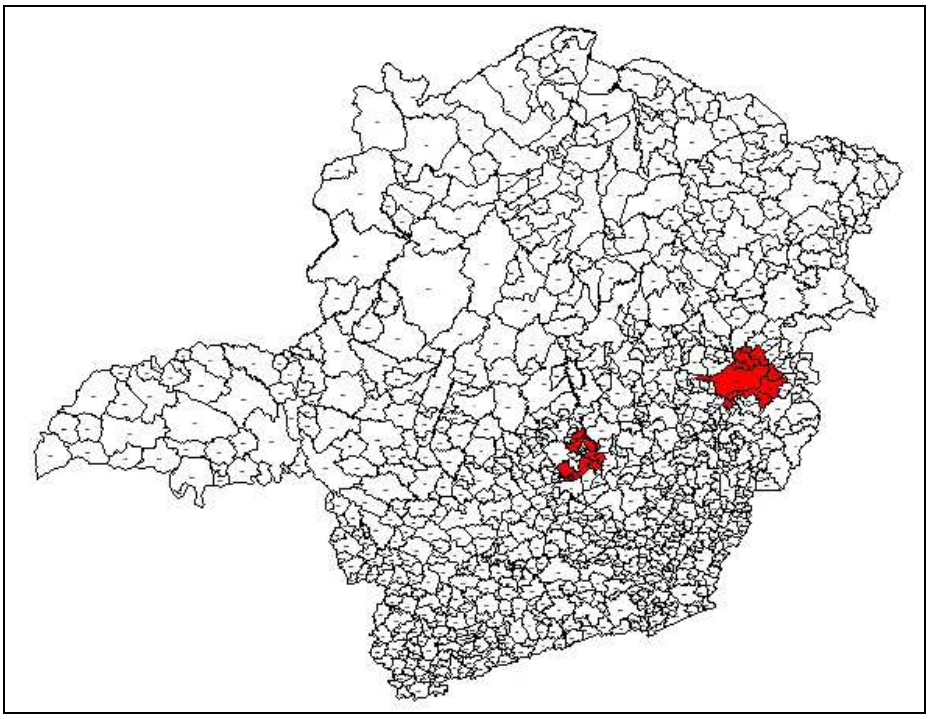

Figura 11: Cluster formado por duas regiões desconexas.

Este cluster foi obtido a partir de dois polos, o de Belo Horizonte e algumas Cidades vizinhas e o polo de Governador Valadares, formado por esta cidade e outras próximas a ela, perfazendo um total de 20 cidades.

A Tabela 11, descrita a seguir, mostra os resultados do poder, sensibilidade e PPV, de acordo com os vários níveis de risco de ocorrência dos eventos.

Tabela 11: Resultados obtidos por simulação cluster dois pólos.

Risco	Qualidade	TRAD	GRAV	GLM
1,10	Poder	0,850	0,922	0,922
	Sensibilidade	0,777	0,800	0,778
	PPV	0,751	0,819	0,833
1,25	Poder	1,000	1,000	1,000
	Sensibilidade	0,796	0,824	0,816
	PPV	0,875	0,944	0,979
1,50	Poder	1,000	1,000	1,000
	Sensibilidade	0,798	0,821	0,825
	PPV	0,891	0,969	0,993
2,00	Poder	1,000	1,000	1,000
	Sensibilidade	0,806	0,817	0,829
	PPV	0,888	0,980	0,997
3,00	Poder	1,000	1,000	1,000
	Sensibilidade	0,814	0,829	0,829
	PPV	0,820	0,966	0,999

Note que para todos os níveis de risco de ocorrência utilizados nas simulações, as técnicas de GRAV e GLM apresentam melhores resultados de poder, sensibilidade e PPV do que o método tradicional, com exceção da sensibilidade, que para um risco igual a 3 foi mais elevado para a técnica tradicional.

Para as duas técnicas propostas pelo estudo, os resultados de poder, sensibilidade e PPV observados estão muito próximos, para todos os níveis de risco utilizados. Entretanto, a técnica GLM gravitacional, em geral, apresenta resultados superiores à técnica gravitacional GRAV, tanto no que se refere ao poder de detecção, quanto à sensibilidade e ao PPV. Para um risco de 1,1 e 1,25, a sensibilidade do método GRAV é superior à do método GLM.

4.3 Aplicação a Dados Reais

Além da simulação, foi realizada a análise dos resultados obtidos para as três técnicas aplicadas a duas situações reais. Este estudo é relevante para avaliação da significância real dos resultados apresentados pelas técnicas trabalhadas.

A primeira situação utilizada para comparar a técnica tradicional com as duas técnicas propostas refere-se ao evento criminalidade, que será representado pelo número de casos de homicídios ocorridos dentro Estado de Minas Gerais em cinco anos, de 2003 a 2008. A escolha da base de homicídios se deve ao fato de estes eventos estarem diretamente ligados ao fluxo de pessoas entre regiões e ao desenvolvimento socioeconômico das mesmas.

A outra situação utilizada para comparar as técnicas propostas com o método tradicional será representada por um evento de doença, a Gripe Influenza A. Esta escolha se deve ao fato de que, além do número de casos estar diretamente ligado aos fatores socioeconômicos, esta é uma doença contagiosa e pode estar diretamente relacionada ao contato entre pessoas, ou seja, é influenciada diretamente pelo fluxo de pessoas contaminadas entre as regiões estudadas. A base de dados representa o número de casos ocorridos nas regiões de interesse entre os anos de 2008 a 2011.

Para as duas situações, utilizou-se a distribuição empírica da estatística de teste, sob H_0 para estimar o p-valor, através de 10.000 simulações. O percentual de varredura foi igual a 30% da população total do mapa, tendo em vista a estabilização do resultado para percentuais acima deste.

Os resultados obtidos para as duas situações descritas serão apresentados nas seções a seguir. Na seção 4.3.1 serão descritos os resultados obtidos para as três técnicas propostas para o conjunto de dados de homicídios e na seção 4.3.2 os resultados do conjunto de dados de gripe.

4.3.1 Resultados Base Homicídios

Esta seção apresenta os resultados obtidos para o evento de criminalidade. Estas informações compreendem o número de homicídios que ocorreram nas Cidades do estado de MG durante os anos de 2003 a 2008. Estas informações foram obtidas junto ao DATASUS (Códigos CID10: X99, Y00, Y28, Y29, X93, X94, X95, Y22, Y23, Y24, X91, X92, X96, X97, X98, Y01, Y02, Y03, Y04, Y05, Y08, Y09, Y20).

Durante o período estudado, foram detectados 20.912 casos, somando-se as ocorrências de cada uma das Cidades do Estado, sendo 4.182 casos por ano, dentro do Estado de Minas Gerais. A média anual da população foi de 19 milhões de habitantes (média dos 6 anos) e a taxa de incidência de casos considerando o período foi de 22 casos por 100 mil habitantes.

Para avaliar a existência de clusters, regiões onde o risco de homicídios é significativamente maior que nas outras áreas, foram utilizadas o método Scan Circular e as duas outras propostas, o Scan Gravitacional, adaptado para base de desenvolvimento econômico (GRAV) e a base de desenvolvimento socioeconômico (GLM), utilizando um percentual de varredura de 30%.

Inicialmente, foi realizada a avaliação da técnica Scan Circular. A partir desta análise detectou-se um cluster, formado por 20 cidades que estão localizadas na Grande BH. A Figura 12, descrita a seguir, apresenta o mapa contendo a descrição do cluster detectado pelo método tradicional.

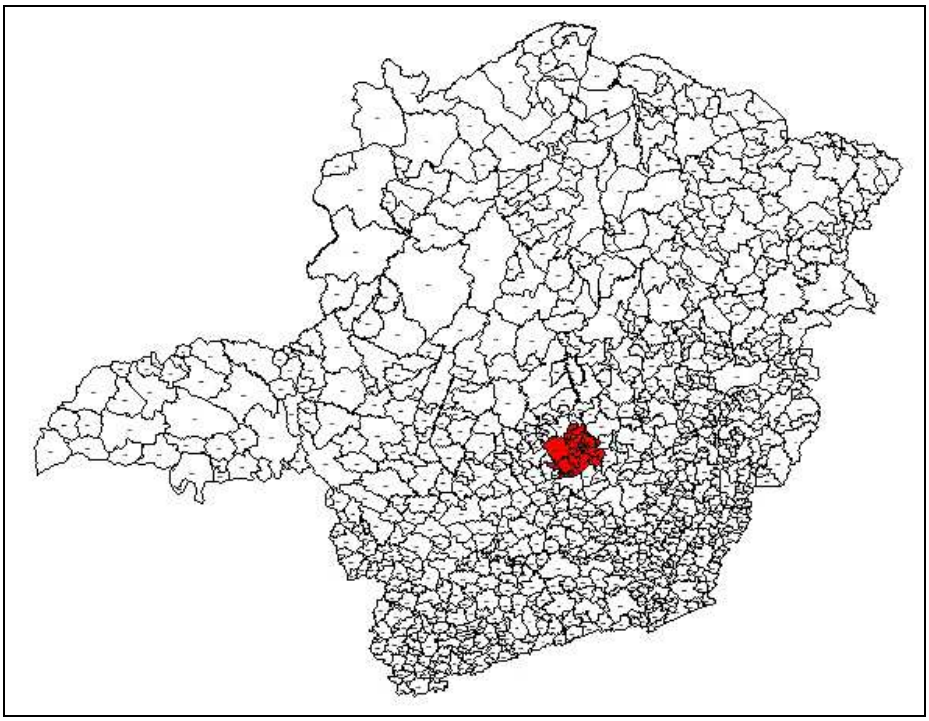

Figura 12: Cluster detectado base homicídios - Método Scan Circular.

Observe, a partir da Figura 12, que o cluster detectado pelo método tradicional apresenta conectividade e forma aproximadamente circular, configurando-se como um cluster regular, característico da técnica tradicional. A Tabela 12, descrita a seguir, apresenta os resultados obtidos de acordo com o método tradicional.

Tabela 12: Cluster detectado Base Homicídios - Método Scan Circular.

Características	**TRAD**
Número de Regiões	20
População	22.824.939
Número de Casos	11.631
Taxa / 100 mil	51
Risco	2,33
LLR	4.822,67
P-valor	< 0,001

Perceba que o cluster formado pelas 20 cidades localizadas na região metropolitana de Belo Horizonte apresentam 11.631 casos para uma população de quase 23 milhões de pessoas, sendo a taxa de incidência de 51 casos para 100 mil habitantes. O risco de

ocorrência de um homicídio é 2,3 vezes maior dentro do cluster em relação às demais regiões. O cluster detectado apresenta significância estatística, sendo a verossimilhança igual a 4.822,67, a partir da qual, a probabilidade de significância (p-valor), que ficou abaixo de 0,001.

A seguir são descritos os resultados obtidos com a utilização do modelo GRAV. Com esta técnica, detectou-se um cluster significativo, formado por 18 cidades também localizadas na Grande BH. O cluster detectado pelo modelo GRAV é descrito a seguir pela Figura 13.

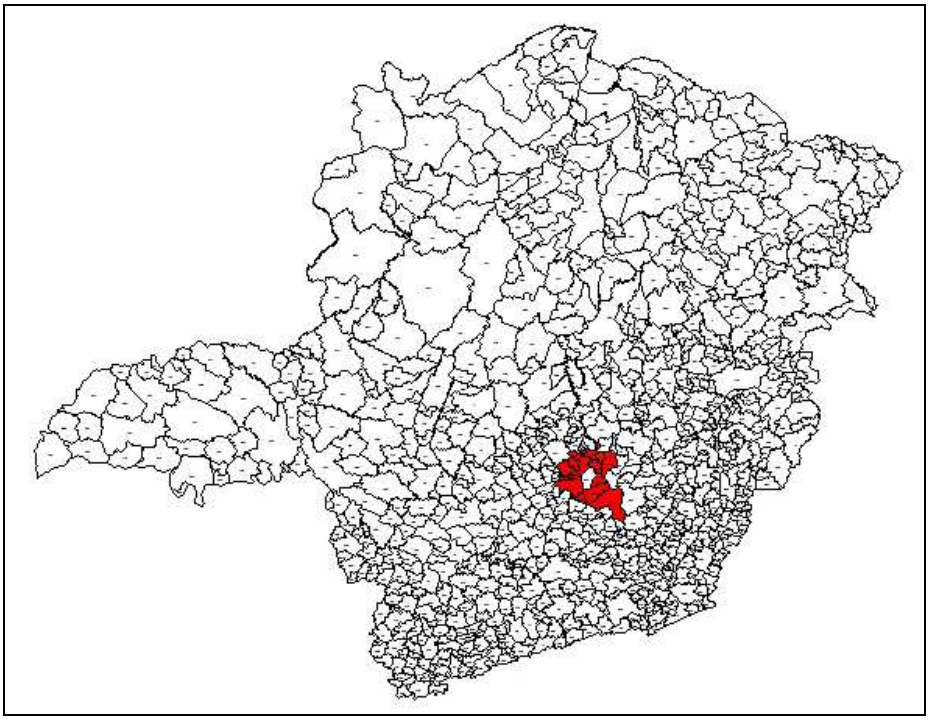

Figura 13: Cluster detectado base homicídios - Método Scan Gravitacional (GRAV).

O cluster detectado pelo método GRAV apresenta conectividade, e forma aproximadamente circular, porém existem duas Cidades, localizadas na parte central que não fazem parte do cluster, configurando-se como um cluster em formato de anel, fugindo do tradicional cluster circular. Além disto, se comparado ao tradicional, este substitui as cidades situadas à oeste de BH por regiões situadas à leste. A Tabela 13, descrita a seguir, apresenta a descrição dos resultados obtidos de acordo com o método gravitacional.

Tabela 13: Cluster detectado base homicídios - Método Scan Gravitacional (GRAV).

Características	**GRAV**
Número de Regiões	18
População	22.556.655
Número de Casos	11.433
Taxa / 100 mil	50,7
Risco	2,32
LLR	4.652,63
P-valor	< 0,001

Como mencionado, o cluster detectado é formado por 18 cidades, localizados na região metropolitana de Belo Horizonte. São 11.433 casos para uma população de aproximadamente 22,5 milhões de pessoas, sendo a taxa de incidência de 51 casos para 100 mil habitantes. Este cluster apresenta um risco de ocorrência de homicídio 2,3 vezes maior dentro do cluster em relação às demais regiões. A verossimilhança foi igual a 4.652,63, sendo o cluster detectado considerado significativo, de acordo com a probabilidade de significância (p-valor), que ficou abaixo de 0,001.

Os resultados obtidos com a utilização do modelo GLM são caracterizados pela Figura 14. Com esta técnica detectou-se um cluster significativo formado por 17 cidades, também localizadas na região metropolitana de Belo Horizonte. Os resultados são descritos a seguir.

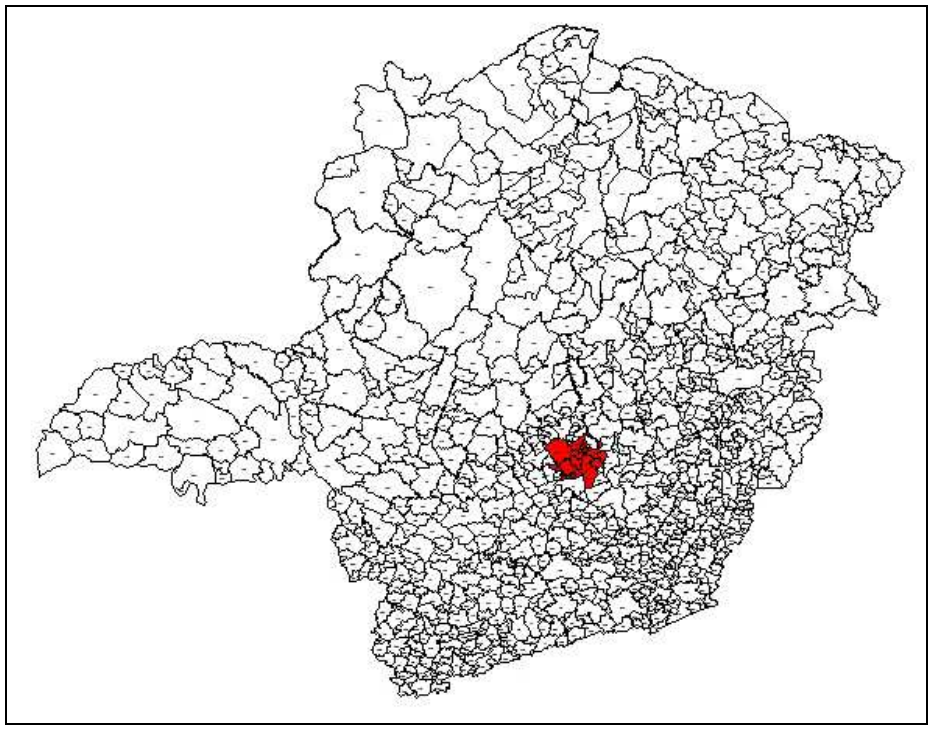

Figura 14: Cluster detectado base homicídios - Método Scan GLM.

Utilizando o método Scan GLM, percebe-se que o cluster detectado também está localizado na região central do Estado. Note que a forma do cluster detectado é mais irregular que os apresentados anteriormente, fugindo ainda mais das características de detecção do método tradicional. A Tabela 10 mostra os resultados obtidos para o cluster estudado.

Tabela 14: Cluster detectado base homicídios - Método Scan GLM.

Características	**GLM**
Número de Regiões	17
População	22.701.984
Número de Casos	11.585
Taxa / 100 mil	51,0
Risco	2,33
LLR	4806,01
P-valor	< 0,001

A partir da análise da Tabela 14 é possível observar que o cluster detectado apresenta um total de 11.433 casos, dentro do período estudado e uma população de aproximadamente 22,7 milhões de pessoas, sendo a taxa de incidência de 51 casos para 100 mil habitantes. O risco de um indivíduo que esteja dentro das regiões que compreendem o cluster ser assassinado é 2,3 vezes maior que nas demais regiões do Estado de MG. A verossimilhança foi igual a 4.806,01, sendo a probabilidade de significância do teste (p-valor) menor que 0,001. Desta forma, conclui-se que o cluster detectado pela técnica GLM é significativo.

A Tabela 15, descrita a seguir, apresenta as Cidades que fazem parte dos clusters detectados, para cada um dos três métodos utilizados na pesquisa.

Tabela 15: Descrição das cidades pertencentes aos clusters detectados base de homicídios.

TRAD	GRAV	GLM
BELO HORIZONTE	BELO HORIZONTE	BELO HORIZONTE
BETIM	BETIM	BETIM
CAPIM BRANCO	BRUMADINHO	CONFINS
CONFINS	CAETÉ	CONTAGEM
CONTAGEM	CONTAGEM	ESMERALDAS
ESMERALDAS	IBIRITÉ	IBIRITÉ
IBIRITÉ	ITABIRITO	JUATUBA
JUATUBA	MÁRIO CAMPOS	MÁRIO CAMPOS
LAGOA SANTA	MOEDA	NOVA LIMA
MARIO CAMPOS	OURO PRETO	PEDRO LEOPOLDO
MATOZINHOS	RAPOSOS	RIBEIRÃO DAS NEVES
PEDRO LEOPOLDO	RIBEIRÃO DAS NEVES	SABARÁ
PRUDENTE DE MORAIS	RIO ACIMA	SANTA LUZIA
RIBEIRÃO DAS NEVES	SABARÁ	SÃO JOAQUIM DE BICAS
SABARÁ	SANTA LUZIA	SÃO JOSÉ DA LAPA
SANTA LUZIA	SÃO JOAQUIM DE BICAS	SARZEDO
SÃO JOAQUIM DE BICAS	SARZEDO	VESPASIANO
SÃO JOSÉ DA LAPA	VESPASIANO	-
SARZEDO	-	-
VESPASIANO	-	-

A partir da análise da Tabela 15 é possível perceber que todos os três métodos detectam clusters formados por cidades da região metropolitana ou próximos à região metropolitana de Belo Horizonte.

Fazem parte do cluster detectado pelo método tradicional cidades com maiores índices de criminalidade como Belo Horizonte, Ribeirão das Neves, Contagem, Santa Luzia, Sabará e Betim. Por outro lado, fazem parte do cluster outras Cidades como Capim Branco, Lagoa Santa, Matozinhos e Prudente de Morais, Confins, Esmeraldas, Juatuba e Pedro Leopoldo, que não possuem índices de criminalidade tão expressivos quanto o conjunto citado anteriormente.

As cidades com maior criminalidade como BH, Betim, Contagem, Ibirité, Ribeirão das Neves, Sabará e Santa Luzia também são detectadas pelo método GRAV. Porém, este também capta cidades como Caeté, Itabirito, Moeda, Ouro Preto, Raposos e Rio Acima, que não são tão citadas como regiões perigosas.

Observe que o cluster detectado pelo método GLM além das regiões consideradas perigosas da Grande BH, detecta Nova Lima, Confins, Esmeraldas, Pedro Leopoldo e São José da Lapa, que não aparecem na imprensa como cidades perigosas.

Dentre os clusters detectados pelos três métodos, perceba que o método tradicional capta 20 cidades, o GRAV 18 e o GLM 17. A interseção entre os três métodos foi de 11 cidades. O método GLM capta 5 Cidades captadas pelo método tradicional e não captadas pelo método GRAV, e o contrário não ocorre. O cluster tradicional só detecta 4 regiões não captadas por nenhum outro método, o cluster GRAV 7 e o cluster GLM 1.

Com relação às verossimilhanças, perceba que os métodos apresentam valores muito próximos, sendo que o tradicional apresenta o maior resultado, seguido pelo GLM e o GRAV. Tanto a taxa de ocorrência do evento, quanto o risco de ocorrência nos clusters detectados pelos métodos são muito próximos, sendo a taxa e o risco do modelo tradicional e do GLM ligeiramente superiores aos do GRAV.

O método GRAV apresenta uma grande distorção em relação ao método tradicional, dando peso elevado ao desenvolvimento econômico, substituindo regiões situadas a oeste de BH por cidades à leste, mais desenvolvidas economicamente.

Importante ressaltar que apesar do método GLM apresentar verossimilhança menor que o tradicional, esta diferença é muito pequena, sendo o número de cidades pertencentes

ao cluster também inferior. Veja que este método retira do cluster detectado pela técnica tradicional regiões que não apresentam tanta criminalidade como Capim Branco, Lagoa Santa, Matozinhos e Prudente de Morais.

4.3.2 Resultados Base Gripe

Os resultados obtidos para análise dos eventos de Gripe serão apresentados nesta seção. A base de dados representa o número de casos de internação por esta doença, que ocorreram nas cidades de MG durante os anos de 2008 a 2011. Estas informações foram obtidas através do DATASUS (CID 10: Influenza (Gripe)).

Durante o período estudado, foram detectados 7244 casos, sendo 1449 casos por ano, dentro das Cidades que compõe o Estado de Minas Gerais. A população sob risco foi de 19 milhões de habitantes (média dos 5 anos) e a taxa de incidência de casos foi de 7,6 casos por 100.000 habitantes.

Novamente foram utilizadas as três técnicas estudadas para avaliar a existência de clusters, regiões em que o risco de ocorrência da doença é significativamente maior que nas outras áreas, utilizando um percentual de varredura igual a 30%.

A seguir são apresentados os resultados obtidos a partir da técnica tradicional, o método Scan Circular. Através desta análise detectou-se um cluster, formado por somente 1 cidade, Itabirito. A Figura 15, a seguir apresenta o mapa contendo a descrição do cluster detectado pelo método tradicional.

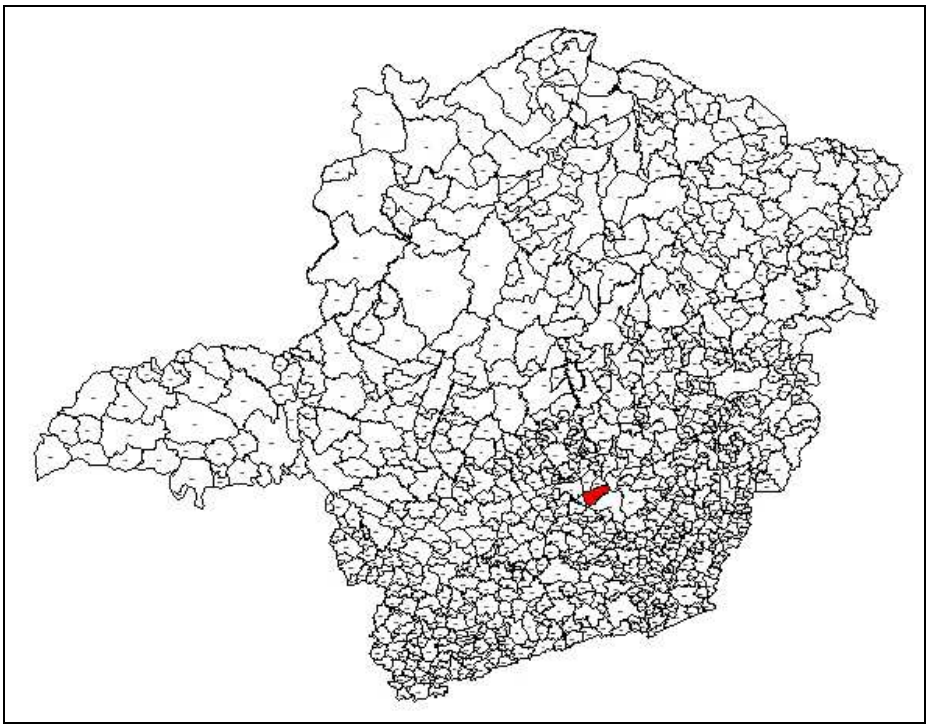

Figura 15: Cluster detectado base gripe - Método Scan Circular.

Observe que a cidade que representa o cluster está localizada na região centro-sul de Minas Gerais. Este cluster apresenta regularidade, por ser formado por somente uma cidade, Itabirito. Verificou-se que este apresenta um risco mais elevado de ocorrência de Gripe, se comparada às demais regiões do Estado. A Tabela 16, descrita a seguir apresenta os resultados obtidos de acordo com o método tradicional.

Tabela 16: Cluster detectado gripe - Método Scan Circular.

Características	**TRAD**
Número de Regiões	1
População	205.831
Número de Casos	822
Taxa / 100 mil	399,4
Risco	52,67
LLR	2498,79
P-valor	< 0,001

Para o cluster identificado, observa-se 822 casos, dentro de uma população de aproximadamente 206 mil pessoas, sendo a taxa de incidência de aproximadamente 400 casos para 100 mil habitantes. As pessoas desta cidade apresentam um risco de ser contaminado pela Gripe 53 vezes maior que em relação às demais regiões. A verossimilhança observada foi iguala a 2.489,79, sendo o cluster detectado considerado significativo, de acordo com a probabilidade de significância (p-valor), que ficou abaixo de 0,001.

A configuração do cluster detectado pelo método Gravitacional (GRAV) bem como pelo método GLM também são idênticas a do cluster detectado pelo método Scan Tradicional. Estes são formados pela Cidade de Itabirito, no centro-sul de Minas Gerais ilustrada pela Figura 15, sendo suas características iguais às descritas pela Tabela 16.

Capítulo 5
Discussão e Conclusões

A ideia desta pesquisa foi de avaliar como as medidas de interação entre regiões podem influenciar na identificação de clusters, tendo como base os resultados do método tradicional Scan Circular de Kulldorff.

Este trabalho propôs uma metodologia de detecção de clusters que utiliza o fluxo de pessoas como medida de proximidade entre regiões do mapa, de forma a identificar um conjunto de regiões com risco elevado de ocorrência de um determinado evento de interesse.

O fluxo de pessoas, utilizado como medida de interação entre regiões foi estimado através do método gravitacional, sendo o fluxo diretamente proporcional aos produtos dos PIB's e inversamente proporcional ao quadrado das distâncias, denominado GRAV. Utilizou-se também um modelo GLM para estimar o fluxo a partir de variáveis explicativas que representam o desenvolvimento econômico, social e a distância entre as regiões.

A partir destas novas medidas de interação entre regiões, foi realizada uma adaptação no método Scan Tradicional, de forma a utilizar este novo critério de proximidade entre regiões. Importante ressaltar que esta nova proposta não alterou a lógica do algoritmo tradicional.

O desempenho dos métodos propostos foram comparados com o método Scan Circular tradicional, que usa a distância entre os centroides das regiões como medida de proximidade, a partir de simulações de uma base de casos de homicídios e aplicação a dados reais, de homicídios e gripe influenza A.

Foram simuladas cinco situações, a partir da base de dados de homicídios, onde o cluster simulado apresentava regularidade, irregularidade conexa, desconexão, formato anelar e dois pólos.

Importante ressaltar que em geral, quando consideramos riscos acima de 1,25, o poder é igual a 1,000. Este é um resultado esperado, tendo em vista que o poder mede a capacidade de detecção de clusters, independente se este é o correto ou não. Além disto, o risco é demasiadamente elevado (para riscos iguais a 1,14 a chance de detecção de clusters

é de 99,9%). Assim, torna-se importante levar em consideração a sensibilidade que é a proporção do cluster real, detectado pelo método e o PPV, que é a proporção do cluster detectado, representada pelo cluster real.

O método tradicional apresentou melhores resultados de poder, sensibilidade e PPV que os outros métodos, para o caso do cluster regular, sendo que as outras duas técnicas propostas apresentaram resultados bem próximos aos obtidos a partir do método tradicional.

Porém, em todas as outras situações, as técnicas propostas tiveram melhores resultados de poder, sensibilidade e PPV que o método tradicional. Dentre as técnicas propostas, a técnica GLM apresentou resultados ligeiramente superiores aos da técnica GRAV. Importante ressaltar que os clusters detectados pela técnica GLM são muito próximos aos simulados (“real”) tendo em vista os melhores resultados observados para sensibilidade.

A aplicação das técnicas à situação real de casos de homicídios mostrou que os três métodos detectam clusters similares, com verossimilhanças muito próximas, sendo que o método tradicional sugere um cluster com um número maior de regiões que os demais.

O método GRAV tende a pesar na detecção o desenvolvimento econômico apresentando grande distorção em relação ao método tradicional, substituindo regiões situadas a oeste de BH por cidades à leste.

Já o método GLM apresenta resultados mais coerentes com a realidade, incluindo as regiões detectadas pelo método tradicional e retirando do cluster, regiões que não apresentam um índice de homicídios tão acentuado como Capim Branco, Lagoa Santa, Matozinhos e Prudente de Morais.

No caso da aplicação das metodologias a uma base de dados de casos de internação por gripe influenza, todos os três métodos detectaram clusters formados por somente uma região. Neste caso, as medidas de proximidade utilizadas nas três metodologias não influenciam no resultado e por isto os clusters são idênticos.

Este trabalho trouxe uma nova visão de interação entre regiões, mais flexível em relação à distância euclidiana, por poder ser adaptada às diversas situações, dependendo do problema estudado. A partir desta visão foi possível adaptar o método tradicional, Scan

Circular, de forma a melhorar a qualidade das detecções, quando o cluster for irregular e desconexo.

Para situações reais, também foram observados resultados mais plausíveis por conta da maior capacidade de adaptação dos métodos propostos à situação estudada. Isto em virtude desta técnica levar em conta a interação entre regiões e por conta da interação ser diretamente influenciada por fatores relacionados ao problema. Assim, foram retiradas regiões que não apresentam relevância no problema, mas que pertenciam ao cluster detectado pelo método tradicional, de forma a manter a conectividade e regularidade.

Como a proximidade considerada por este modelo apresenta grande influência do desenvolvimento econômico, percebe-se que os resultados são mais satisfatórios quando o cluster simulado contém a cidade de Belo Horizonte, por esta apresentar PIB bem superior às demais cidades, além de sua população ser consideravelmente maior.

Como agora o modelo é flexível, pode-se modificar os parâmetros de estimação do fluxo, de forma a obter melhores resultados, utilizando por exemplo critérios relacionados ao problema estudado. Além deste fato, existe uma dificuldade importante relacionada à obtenção das informações de fluxo, critério essencial para utilização dos métodos propostos.

Assim, conclui-se que os métodos que utilizam a interação entre regiões a partir do modelo gravitacional (GRAV e GLM) são boas alternativas para detecção de clusters quando temos irregularidade e não conectividade. Nestes casos, tanto o poder, quando a sensibilidade e o PPV são melhorados com a utilização das novas propostas.

Dentre os métodos propostos, o modelo GLM gravitacional se apresenta como melhor opção para os casos estudados, tendo em vista a qualidade dos resultados obtidos em termos de poder, sensibilidade e PPV e a maior proximidade com a realidade, apresentada no caso da avaliação dos casos de criminalidade.

Capítulo 6
Sugestões Para Trabalhos Futuros

A partir dos resultados e das conclusões obtidas com este estudo torna-se importante deixar sugestões para outros trabalhos relacionados à metodologia de detecção de clusters espaciais, abordando aspectos não contemplados nesta pesquisa.

As sugestões são de utilização de outras técnicas para estimação dos fluxos nulos, a utilização de outros indicadores para construção dos modelos gravitacionais para estimação do fluxo e utilização da técnica gravitacional em outros contextos de detecção de clusters.

Pode-se utilizar outras metodologias para estimação do fluxo de pessoas entre as regiões, quando este é nulo de acordo com os dados de viagens, estimando os fluxos através do fluxo médio entre os vizinhos ou utilizando o critério Bayesiano.

Dentro do contexto do fluxo de pessoas, outras variáveis podem também ser utilizadas e testadas como parâmetros para construção do modelo gravitacional, como por exemplo, a população das regiões, a renda per capita, o número de hospitais, o número de hotéis, etc. Para o caso do modelo GLM, pode-se testar outras funções de ligação para o modelo logístico, no lugar da ligação logit e outras medidas resumo para o PIB e IDH entre regiões, no lugar da média aritmética.

Como as metodologias propostas dependem do problema abordado, ou seja, do tipo de caso estudado, podem ser ainda construídos modelos diversos para estimação da interação entre regiões, levando em consideração fatores associados ao fluxo de pessoas e à ocorrência dos eventos de interesse como, por exemplo, em casos de devastação ambiental, mortalidade de animal, outros tipos de doença e crimes.

Referências

AGARWAL, D.; McGREGOR, A.; VENKATASUBRAMANIAN, S. ; ZHU, Z. **Spatial Scan Statistics Approximations and Performance Study**. Conference on Knowledge Discovery in Data Mining, 2006.

BALAKRISHNAN, N. ; KOUTRAS, M.V. **Runs and Scans with Applications, John Wiley & Sons**, New York, 2002.

BESSAG, J. ; NEWELL, J. **The detection of clusters in rare diseases**. Journal of the Royal Statistical Society, 1990. 154(1):143-155.

CANÇADO, A. L. F.; DUARTE, A. R.; DUCZMAL, L. H.; FERREIRA, S. J.; FONSECA, C. M.; GONTIJO, E. C. D. M. **Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters**. International Journal of Health Geographics, 2010. 9:55.

CANÇADO, A. L. F. **Detecção de Clusters Espaciais Através de Otimização Multiobjetivo**. Teste Phd – Departamento de Engenharia Elétrica, Universidade Federal de Minas Gerais, Belo Horizonte, 2009.

CARVALHO, A. F. **Detecção de clusters irregulares para dados pontuais através da Não-conectividade Ponderada de Grafos**. Dissertação (Mestrado em Estatística) - Departamento de Estatística, UFMG, Belo Horizonte, 2011.

CONOVER, W. J. **Practical Nonparametric Statistics**. John Wiley & Sons, New York, 1971.

CRESSIE, N. A. C. **Statistics for Spatial Data**. New York: Wiley,1993.

DWASS, M. **Modified randomization tests for nonparametric hypotheses**. Annals of Mathematical Statistics, 1957. 28:181–187.

DUARTE A. R. **Geometria e Topologia de Conglomerados Espaciais Baseados em Grafos**. Tese Phd – Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, 2009.

DUCZMAL, L. H.; ASSUNÇÃO, R. **A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters**. Computational Statistics and Data Analysis, 2004. 45, 269-286.

DUCZMAL, L. H. ; BUCKERIDGE, D. L. **A workflow spatial scan statistic, Statist. Med.**, 2006. 25, 743–754.

DUCZMAL, L. H.; CANÇADO, A. L. F.; TAKAHASHI, R. H. C.; BESSEGATO, L. F. **A Genetic Algorithm for Irregularly Shaped Spatial Scan Statistics**. Computational Statistics and Data Analysis, 2007. 52, 1, 43-52.

DUCZMAL, L. H.; CANÇADO, A. L. F.; TAKAHASHI, R. H. C. **Delineation of irregularly shaped disease clusters through multi-objective optimization**. Journal of Computational & Graphical Statistics, 2008.17:243–262.

DUCZMAL, L. H.; CANÇADO, A. L. F.; TAKAHASHI, R. H. C. **Geographic Delineation of Disease Clusters through Multi-Objective Optimization**. Journal of Computational & Graphical Statistics, 2008. 17, 243-262.

DUCZMAL, L. H.; DUARTE, A. R.; TAVARES, R. **Scan Statistics: Extensions of the scan statistic for the detection and inference of spatial clusters**. Edited by Glaz, J. and Pozdnyakov, V. and Wallenstein. Birkhauser, 2009. 153-177.

DUCZMAL, L. H., KULLDORFF, M.;HUANG, L. **Evaluation of spatial scan statistics for irregularly shaped disease clusters**. Journal of Computational & Graphical Statistics, 2006.15:428–442.

ESKELSON, Bianca, N. I.; TEMESGEN, Hailemariam; LEMAY, Valerie; BARRETT, Tara M.; CROOKSTON, Nicholas L.; HUDAK, Andrew T. **The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases**. Scandinavian Journal of Forest Research, 2009. 24: 235-246.

FIGUEIREDO, R. L. **Detecção de clusters usando a Estatística Scan Espacial Circular em conjuntos seletivos e um fator de penalização**: a ocupação circular. Dissertação (Mestrado em Estatística) – Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.

HUANG, L.; KULLDORFF, M.; GREGORIO, D. **A Spatial Scan Statistic for Survival Data**. Biometrics, 2007. 63, 109-118.

HUANG, L.; PICKLE, L.W.; STINCHCOMB, D.; FEUER, E.J. **Detection of Spatial Clusters**: Application to Cancer Survival as a Continuous Outcome. Epidemiology, 2007. 18, 73-87.

KULLDORFF, M. **A Spatial Scan Statistic.** Communications In Statistics: Theory and Methods, 1997. 26(6), 1481-1496.

KULLDORFF, M. **Spatial scan statistics**: Models, calculations, and applications. In: Recent Advances on Scan Statistics and Applications Edited by Balakrishnan and Glaz. Boston: Birkhauser, 1999. p. 303-322

KULLDORFF, M.; TANGO, T.; Park, P. J. **Power comparisons for disease clustering tests**. Computational Statistics & Data Analysis, 2003.42: 665–684.

KULLDORFF, M. (2006). Information Management Services Inc: SaTScan v7.0: software for the spatial and space-time scan statistic. Disponível em < http://www.satscan.org.>. Acesso em: 20/10/2011.

KULLDORFF, M.; MOSTASHARI, F.; DUCZMAL, L.; YIH, K.; KLEINMAN, K.; PLATT, R. **Multivariate Scan Statistics for Disease Surveillance**. Statistics in Medicine, 2007. 26, 1824-1833.

LAWSON, A.; BIGGERI, A.; VOHNING, D. B.; LESARE, E.; VIEL, J. F.; BERTOLLINI, R. **Disease Mapping and Risk Assessment for Public Health**. Wiley, London, 1999.

LAWSON, A. **Statistical methods in spatial epidemiology**. Chichester: Wiley, 2001.

LIMA, M. S. **Avaliação do Poder do Teste da Estatística Scan para Múltiplos Clusters.** 2004, 100f . Dissertação (Mestrado em Estatística) – Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, 2004.

LINDERS, G. M., de GROOT and H. L. F. **Estimation of the gravity equation in the presence of zero flows**. Tinbergen Institute Discussion Paper (2006).

MARSHALL, R. J. **Mapping Disease and Mortality Rates Using Empirical Bayes Estimators**. Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 40, No. 2(1991), pp. 283-294.

MOURA, F. R. **Detecção de Clusters Espaciais via Algoritmo Scan Multi-Objetivo**. Dissertação (Mestrado em Estatística). Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, 2006.

OHMANN, Janet L.; GREGORY, Matthew J.; HENDERSON, Emilie B.; ROBERTS, Heather M. **Mapping gradients of community composition with nearest-neighbour imputation**: extending plot data for landscape analysis, Journal of Vegetation Science, 2011. 10.1111/j.1654-1103.2010.01244.x: 17 p.

R Development Core Team. R: **A language and environment for statistical computing.** R Foundation for Statistics Computing, Vienna, Austria. ISBN 3-900051-07-0.

SIGNORINO, G.; PASETTO, R.; GATTO, E.; MUCCIARDI, M.; ROCCA, M. La; MUSO, P. **Gravity models to classify commuting vs. resident workers**. An application to the analysis of residential risk in a contaminated area. International Journal of Health Geographics, 2011. 10:11, pp. 1-10.

SISTEMA ÚNICO DE SAÚDE. <www.datasus.gov.br>.

SHAO, J. ;WANG, H. **Confidence Intervals Based On Survey Data With Nearest Neighbor Imputation**. Statistica Sinica, 2008.18: 281-297.

SILVA, S. B. **Detecção de Clusters Irregulares Através da Não Conectividade Ponderada de Grafos.** Dissertação (Mestrado em Estatística) – Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.

SMITH, R. L. **Maximum likelihood estimation in a class of non-regular cases**. Biometrika, 1985.72:67–90.

STEPHENSON, A. G. evd: Extreme value distributions. R News, 2002. 2(2):0.

YIANNAKOULIAS, N.; ROSYCHUK, R.J.; HODGSON, J. **Adaptations for finding irregularly shaped disease clusters**. International Journal of Health Geographics, 2007. 6, 28.