Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

DETECÇÃO, IDENTIFICAÇÃO E INFERÊNCIA DE CONGLOMERADOS ESPACIAIS DE FRAUDES BANCÁRIAS EM UMA INSTITUIÇÃO FINANCEIRA NO CENTRO-OESTE DO BRASIL

Lucas Barbosa Fernandes

Samuel Dornelas de Souza Reis

Relatório Final

Orientador: Prof. André Luiz Fernandes Cançado

Brasília

Outubro de 2011

Lucas Barbosa Fernandes

Samuel Dornelas de Souza Reis

DETECÇÃO, IDENTIFICAÇÃO E INFERÊNCIA DE CONGLOMERADOS ESPACIAIS DE FRAUDES BANCÁRIAS EM UMA INSTITUIÇÃO FINANCEIRA NO CENTRO-OESTE DO BRASIL

Relatório apresentado à disciplina Estágio Supervisionado do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. André Luiz Fernandes Cançado

Brasília

Outubro de 2011

AGRADECIMENTOS

Agradeço ao meu pai, Flávio, grande fonte de educação, amizade e força de vontade.

Agradeço a minha irmã Carolina por sempre estar ao meu lado e me apoiar quando eu preciso.

A minha mãe, Ony, por todo seu amor materno e educação, além do apoio e orientação nesse trabalho.

Aos companheiros de curso, essenciais em todas horas de estudo.

Agradeço ao professor Alan Ricardo pela disponibilidade e assistência sobre o *software* SAS.

Ao professor André Cançado por toda a compreensão e orientação. Sem ele esse trabalho não seria possível.

Agradeço ao SAS Institute Brasil por possibilitar a utilização desse software por meio de parceria acadêmica com o Departamento de Estatística da Universidade de Brasília.

Por fim, agradeço ao amigo e companheiro Samuel Dornelas pela paciência e auxílio nesse trabalho.

Lucas Barbosa Fernandes

AGRADECIMENTOS

Agradeço em primeiro lugar a Deus, pois sem ele nada disso seria possível.

Agradeço aos meus pais, Ilda e João, pelo apoio, conselhos e educação.

Aos meus irmãos, Orney e Kleyne, por sempre estarem por perto para ajudarem no que for preciso.

Aos companheiros de curso, essenciais para minha formação academica.

Agradeço ao professor Alan Ricardo pela disponibilidade, e assistência sobre o *software* SAS.

Ao professor André Cançado pela compreensão, assistência e orientação.

Por fim e não menos importante agradeço ao amigo e companheiro Lucas Barbosa pela ajuda e disponibilidade não importando dia ou horário.

Samuel Dornelas de Souza Reis

EPÍGRAFE

Para quem só sabe usar um martelo,

todo problema é um prego.

Abraham Maslow

RESUMO

Esse trabalho utiliza o método de detecção de conglomerados espaciais *Scan* Circular, proposto por Kulldorff (1997), para detectar e identificar uma área de risco de fraudes bancárias na região Centro-Oeste do Brasil, utilizando um banco de dados de uma grande instituição financeira.

O *Scan* Circular é uma técnica de estatística espacial que se utiliza de processos iterativos para detectar os conglomerados, definido como uma área ou região que possui alta incidência de algum fenômeno de interesse, ou incidência maior que o esperado para tal área. São utilizados testes de hipótese e simulação de Monte Carlo para auxiliar na decisão dos testes.

Para a utilização no trabalho, o algoritmo *Scan* Circular foi adaptado com o auxílio do pacote de linguagem matricial *SAS/IML*, presente no *software* SAS 9.2.

Palavras-chaves: Scan Circular, Estatística Espacial, Fraudes, Conglomerados Espaciais.

ABSTRACT

This work uses the Kulldorff's circular spatial scan statistics to detect and identify a risk area of bank fraud in the Midwest region of Brazil, using a real database of a financial institution.

Kulldorff's circular scan is a spatial statistics technique used to detect spatial clusters, defined as an area or region with high incidence of some phenomenon of interest, or a higher-then-expected incidence for this area. A statistical hypothesis test is performed and Monte Carlo simulation is used to assess the cluster's significance.

Circular Scan algorithm has been adapted using the matrix language package SAS / IML, included in the SAS 9.2 software.

Key words: Scan, Spatial Analysis, Fraud, Clustering.

SUMÁRIO

1. INTRODUÇÃO 1

1.1 OBJETIVOS 6

1.2 METODOLOGIA 6

2. O ALGORITMO *SCAN* CIRCULAR 7

2.1 O TESTE DE HIPÓTESES 7

2.2 A ESTATÍSTICA DO TESTE DE RAZÃO DE VEROSSIMILHANÇA 8

2.3 A MATRIZ DE DISTÂNCIAS 11

2.4 ENCONTRANDO CANDIDATOS A CONGLOMERADOS 12

2.5 VERIFICAÇÃO DA SIGNIFICÂNCIA DO CONGLOMERADO 13

2.6 RESUMO DO ALGORITMO 14

2.7 ADAPTAÇÃO PARA DADOS PONTUAIS 15

3. MATERIAL E MÉTODOS 16

3.1 O PROBLEMA 16

3.2 BANCO DE DADOS 16

3.3 FRAUDES 18

3.4 COORDENADAS GEOGRÁFICAS 20

3.5 MACRO %SCANCIRCULAR 21

3.5.1 A busca do conglomerado 22

3.5.2 Simulação 23

3.5.3 Exemplo 24

4. RESULTADOS 26

4.1 DADOS PONTUAIS 26

4.2 DADOS AGREGADOS POR MUNICÍPIOS 30

4.3 CONGLOMERADOS SECUNDÁRIOS 34

4.4 OUTRAS ANÁLISES 35

5. CONSIDERAÇÕES FINAIS 38

5.1 CONCLUSÕES 38

5.2 TRABALHOS FUTUROS 39

6. BIBLIOGRAFIA 40

APÊNDICE A - MACRO %SCANCIRCULAR 42

1. INTRODUÇÃO

Estatística Espacial é o ramo da estatística que congrega um conjunto de métodos exploratórios e inferenciais que visa estudar e analisar padrões em dados distribuídos geograficamente. Sua importância se torna cada dia maior pelo fato de auxiliar na compreensão e resolução de problemas que envolvem dados espaciais, conforme explanado:

> Compreender a distribuição espacial de dados oriundos de fenômenos ocorridos no espaço constitui hoje um grande desafio para a elucidação de questões centrais em diversas áreas do conhecimento, seja em saúde, em ambiente, em geologia, em agronomia, entre tantas outras. (CÂMARA et al., 2004, p. 1)

Os principais tipos de dados que ocorrem em Estatística Espacial são: (a) dados pontuais, onde se conhece a localização geográfica de cada evento (ponto), (b) dados de área, onde os "dados são agregados em unidades de análise, usualmente delimitadas por polígonos fechados" (CÂMARA et al., 2004, p. 3) e (c) dados contínuos ou de superfície, típicos de levantamentos de recursos naturais.

No caso de dados pontuais, o banco de dados constitui-se num conjunto de pares (x_i, y_i) que indicam as coordenadas geográficas de cada ocorrência do evento de interesse (por exemplo, casos de uma determinada doença ou crime). Dados de área são obtidos quando não estão disponíveis as coordenadas de cada ocorrência do evento, mas apenas o número total de ocorrências em cada região - geralmente uma unidade administrativa - do mapa em estudo (por exemplo, o número total de casos de uma doença ou crime em cada município de um estado). Finalmente, os dados de superfície são obtidos ao se realizar medições em determinadas localizações do mapa, sendo então cada elemento do conjunto de dados formado por uma tripla (x_i, y_i, z_i) que corresponde à coordenada geográfica aliada à medição feita

naquela localização (por exemplo, medição de temperatura ou umidade em determinadas localizações).

A análise de dados pontuais e dados de área geralmente envolve questões relativas à distribuição geográfica dos eventos, ao passo que na análise de dados de superfície o pesquisador geralmente deseja prever, baseado nas medições feitas, o valor da medição em pontos que não tenham sido amostrados. Este trabalho envolve a análise de dados pontuais e dados de área.

Uma importante questão que surge ao se analisar dados pontuais e dados de área é se a ocorrência dos eventos se dá de forma completamente aleatória ou se os casos estão distribuídos segundo algum padrão que defina um conglomerado de casos. Para responder a tais questões lançamos mão de uma área da estatística espacial que se presta à detecção e análise de conglomerados espaciais. Pode-se definir conglomerado espacial como um "subconjunto de regiões do mapa em que a ocorrência de casos é discrepante do restante do mapa" (FIGUEIREDO, 2010, p. 3).

A detecção, identificação e inferência de conglomerados espaciais possui grande utilidade prática, principalmente no direcionamento de políticas públicas como controle de criminalidade, de trânsito ou mesmo melhoria da saúde pública. Em estudos de uma doença contagiosa, por exemplo, é possível identificar a área de maior risco em um estado ou região com a detecção de conglomerados, direcionando e otimizando recursos para combater a doença.

Em outras áreas a detecção de conglomerados também é bastante utilizada. Segundo Kulldorff (1997), para encontrar depósitos de urânio, aviões utilizam instrumentos de medição de radiação – denominados contadores Geiger - enquanto voam em linhas paralelas sobre grandes áreas. Um conglomerado seria detectado por um grande número de contagens em uma área específica. A medida seria uniforme nas linhas de voo e zero em outros lugares.

Existem diversos métodos para detecção de conglomerados espaciais, a exemplo dos métodos GAM (*Geographical Analysis Machine)* de Openshaw (1988), Besag e Newell (1991) e o *Scan* Circular de Kulldorff (1997). Além desses métodos, existem também estatísticas específicas que identificam correlações espaciais, mas não os conglomerados propriamente, como a estatística I de Moran (1950).

O método GAM de Openshaw (1988) propunha fixar um raio r e então utilizar um processo iterativo com base em círculos com raio idêntico ao raio pré-fixado para identificar os supostos conglomerados.

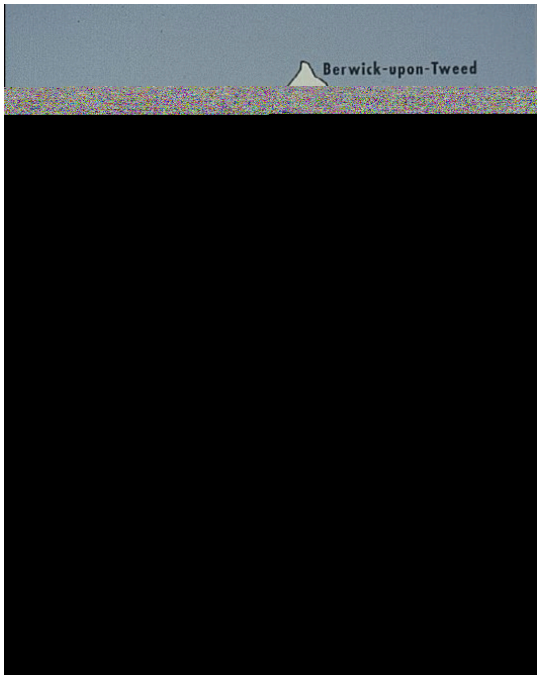

Figura 1 – Representação gráfica do Método GAM.

Fonte: http://www.csiss.org/learning_resources/content/good_sa/

Apesar do método GAM ser de fácil compreensão e apresentar grande apelo visual, possui algumas desvantagens em relação a outros métodos:

> É um método exploratório e não inferencial devido ao problema de muitos testes simultâneos e dependentes implicando em níveis de significância nominais; é intensivo computacionalmente; os círculos não são inteiramente comparáveis entre si pois as variáveis aleatórias envolvidas possuem diferentes distribuições. (ASSUNÇÃO, 2001, p. 97)

Besag e Newell (1991) propuseram outro método que, ao contrário do GAM, expande o círculo até incluir um número pré-estabelecido de casos. A população dentro do círculo é então comparada com a esperada. Apesar de estabilizar as estatísticas de teste locais, o método ainda é exploratório e ineficiente para realizar inferências, pelo mesmo motivo do problema com testes simultâneos.

Por ser de fácil entendimento e resolver o problema de testes simultâneos dos métodos citados anteriormente, o método *Scan* Circular possui ampla aplicabilidade em diversas áreas. Esse método, proposto por Kulldorff (1997) consiste em uma técnica de detecção e inferência de conglomerados espaciais de um evento de interesse, como uma doença ou um tipo de crime. Parte-se do pressuposto que são conhecidos, para cada região do mapa em estudo:

1) a população sob risco de ocorrência do evento a ser estudado. Por exemplo, no caso de uma doença, todas as pessoas da região que podem, de alguma forma, vir a contrair tal doença.
2) o número de casos do evento. Por exemplo, o número de ocorrências de uma doença registrado na região.
3) As coordenadas geográficas de um ponto arbitrário e pré-estabelecido no interior da região, conhecido como centroide.

O método proposto por Kulldorff consiste na execução de um processo iterativo varre o mapa com janelas circulares em busca dos conglomerados. Os círculos são limitados por uma proporção máxima da população que poderá estar dentro deles, proporção essa que será pré-estabelecida. O objetivo é encontrar um conglomerado onde o número de casos do evento a ser estudado é significativamente maior que o esperado dentro da janela circular.

Após a identificação de um suposto conglomerado, é realizado um teste de hipóteses para verificar se o conglomerado é ou não significativo do ponto de vista estatístico. São utilizadas simulações de Monte Carlo como auxílio para a decisão dos testes.

Visando identificar um conglomerado de fraudes bancárias de uma grande instituição financeira na região Centro-Oeste do Brasil, o trabalho aplicará o algoritmo *Scan* Circular a um banco de dados real com informações sobre fraudes de uma grande instituição financeira do Brasil. Tais fraudes referem-se a operações com cartão de débito no ano de 2010.

O algoritmo encontra-se já implementado no *software* SaTScan, porém sua adaptação em outras linguagens também não é complicada. O trabalho se utilizou da adaptação do algoritmo e confecção de uma macro em linguagem *SAS* para a análise dos resultados.

O objetivo será de verificar a existência ou não de um conglomerado espacial de fraudes bancárias na região centro-oeste, composta pelos estados de Goiás, Mato Grosso, Mato Grosso do Sul e o Distrito Federal.

A técnica Scan Circular já foi aplicada no contexto de criminalidade (BEATO FILHO et. al. 2001) e os seus resultados podem ajudar a instituição financeira a preparar e aplicar planos de prevenção em uma determinada área onde o número de fraudes foi superior ao esperado, podendo assim reduzir prejuízos.

Com a detecção de conglomerados, também podem ser estudadas supostas causas sociais e econômicas para um maior entendimento do motivo de algumas regiões estarem incluídas no conglomerado. Esse estudo pode contribuir ainda mais para entender as verdadeiras causas desses crimes em algumas regiões.

1.1 OBJETIVOS

O objetivo geral desse trabalho é a aplicação da técnica de detecção de conglomerados espaciais denominada *Scan* Circular em um banco de dados real de uma instituição financeira para verificar a existência de conglomerados de fraudes bancárias na região centro-oeste do Brasil.

Os objetivos específicos são:

- Implementação e adaptação da técnica de *Scan Circular* na linguagem SAS/IML.

- Detectar e identificar conglomerados espaciais em um banco de dados real com informações de fraudes bancárias na região centro-oeste do Brasil.

1.2 METODOLOGIA

Os dados foram fornecidos por uma grande instituição financeira e manipulados no software SAS ® 9.2.

A implementação da técnica *Scan Circular* na linguagem SAS foi realizada no software SAS ® 9.2 utilizando principalmente linguagem matricial no procedimento SAS/IML. Para a confecção dos mapas utilizados no trabalho, foi utilizado o pacote SAS/GRAPH 9.2.

2. O ALGORITMO *SCAN* CIRCULAR

Esse capítulo visa detalhar o algoritmo *Scan* Circular, proposto por Kulldorff (1997) para realizar detecção e inferência de conglomerados espaciais.

Em sua versão para dados de área, pressupõe-se que são conhecidos, para cada região do mapa, os seguintes dados: (a) a população, (b) o número de casos do evento de interesse e (c) as coordenadas geográficas do centroide, um ponto arbitrário em seu interior.

O processo iterativo varre o mapa em janelas circulares em busca de possíveis conglomerados. É feito um teste de razão de verossimilhança para verificar a significância do conglomerado, descrito a seguir. Os procedimentos do algoritmo serão explicados posteriormente.

2.1 O TESTE DE HIPÓTESES

A fim de verificar se o número de casos dentro de uma determinada janela é, do ponto de vista estatístico, significativamente alto, é suposto que os dados são variáveis aleatórias que seguem a distribuição de Poisson, geralmente utilizada para descrever ocorrências de casos em um determinado período de tempo. Pode-se utilizar também o modelo multinomial, porém os dois são assintoticamente equivalentes. Sendo assim, o número de casos na *i*-ésima região, r_i, terá a seguinte função de probabilidade:

$$f_i(c) = \begin{cases} \dfrac{e^{-\lambda_i}\lambda_i^c}{c!}, & c \geq 0 \\ 0, & \text{c.c.} \end{cases}$$

Uma zona será qualquer subconjunto conexo de regiões no mapa. Uma zona z possuirá, respectivamente, uma população n_z e um número de casos c_z. Seja p a probabilidade de que um indivíduo nessa zona venha a ser um caso. Logo μ_z, a média de casos na zona z, será proporcional à sua população, isto é:

$$\mu_z = pn_z$$

Como o número de casos em uma determinada zona segue a distribuição Poisson, a função de probabilidade de casos nessa zona será:

$$f_i(c_z) = \begin{cases} \frac{\mu_z^{c_z} e^{-\mu_z}}{c_z!}, & c_z \geq 0 \\ 0, & \text{c. c.} \end{cases}$$

Sejam N e C, respectivamente, a população e o número de casos total do mapa. Considere agora $N_{\bar{z}}$ e $C_{\bar{z}}$ como a população e o número de casos fora da zona z. Então:

$$N_{\bar{z}} = N - N_z \tag{2.1}$$

$$C_{\bar{z}} = C - C_z \tag{2.2}$$

Sob a hipótese de não-existência de conglomerados no mapa, espera-se que a probabilidade de que ocorra um caso seja a mesma independente de estar dentro ou fora da zona *z*, para todo *z*. Logo, são utilizados testes de hipóteses para verificar ou não a presença de conglomerados no banco de dados. São consideradas as seguintes hipóteses:

$$\begin{cases} H_0: p = q \\ H_1: p > q \end{cases}$$

onde, conforme citado anteriormente, p é a probabilidade de ocorrer um caso na zona z e q a probabilidade de ocorrer um caso fora dela. A zona z será considerada um conglomerado caso o teste rejeite a hipótese nula.

O teste a ser usado será o de razão de verossimilhança, utilizado para comparar o ajuste de dois modelos. A obtenção da estatística do teste é detalhada a seguir.

2.2 A ESTATÍSTICA DO TESTE DE RAZÃO DE VEROSSIMILHANÇA

Sob a hipótese nula H_0 tem-se que, para qualquer zona z:

$$\mu_z = pn_z \; ; \; \mu_{\bar{z}} = p(N - n_z)$$

Então, a verossimilhança pode ser escrita como:

$$L_0(z;p) = \frac{\mu_z^{C_z} e^{-\mu_z}}{C_z!} \frac{\mu_{\bar{z}}^{C_{\bar{z}}} e^{-\mu_{\bar{z}}}}{C_{\bar{z}}!}$$

$$= \frac{(pN_z)^{C_z} e^{-pN_z}}{C_z!} \frac{(p(N-N_z))^{(C-C_z)} e^{-p(N-N_z)}}{(C-C_z)!}$$

Utilizando $l_0(z;p) = logL_0(z;p)$, tem-se que:

$$l_0(z;p) = C_z[\log p + \log N_z] - pN_z - \log C_z! + (C-C_z)[\log p + \log(N-N_z)] \\ - p(N-N_z) - \log[(C-C_z)!]$$

Deve-se encontrar o ponto que maximiza a função $l_0(z;p)$. Então:

$$\frac{\partial l_0(z;p)}{\partial p} = \frac{C_z}{p} - N_z + \frac{(C-C_z)}{p} - (N-N_z) = 0 \Rightarrow$$

$$\Rightarrow \frac{C}{p} - N = 0 \Rightarrow p = \frac{C}{N} \qquad (2.3)$$

Substituindo $p = \frac{C}{N}$ em $L_0(z,p)$:

$$L_0(z) = \frac{\left(\frac{C}{N}N_z\right)^{C_z} e^{-\left(\frac{C}{N}N_z\right)}}{C_z!} \frac{\left(\frac{C}{N}(N-N_z)\right)^{C-C_z} e^{-\frac{C}{N}(N-N_z)}}{(C-C_z)!}$$

Considere $\mu_z = c\frac{N_z}{N}$. Logo tem-se que:

$$L_0(z) = \frac{\mu_z{}^{C_z} e^{-\mu_z}}{C_z!} \frac{(C-\mu_z)^{C-C_z} e^{-(C-C_z)}}{(C-C_z)!} = \frac{\mu_z{}^{C_z}(C-\mu_z)^{C-C_z} e^{-C}}{C_z!\,(C-C_z)!}$$

Considerando agora a hipótese alternativa, tem-se que:

$\mu_z = pN_z$ e $\mu_{\bar{z}} = q(N-N_z),\ p > q$

Isto é, a probabilidade p de que um indivíduo na zona z venha a ser um caso é maior que fora dela. Calcula-se então a função de verossimilhança sob H_1:

$$L(z;p;q) = \frac{(pN_z)^{C_z} e^{-pN_z}}{C_z!} \frac{[q(N-N_z)]^{(C-C_z)} e^{-q(N-N_z)}}{(C-C_z)!}$$

Sendo $l(z;p;q) = logL(z;p;q)$, tem-se:

$$l(z;p;q) =$$

$$C_z[\log p + \log N_z] - pN_z - \log C_z! + (C - C_z)[\log q + \log(N - N_z)]$$
$$- q(N - N_z) - \log[(C - C_z)!]$$

Deve-se encontrar os pontos que maximizam $l(z;p;q)$. Então:

$$\frac{\partial l(z;p;q)}{\partial p} = \frac{C_z}{p} - N_z = 0 \Rightarrow p = \frac{C_z}{N_z}$$

$$\frac{\partial l}{\partial q} = \frac{(C - C_z)}{q} - (N - N_z) = 0 \Rightarrow q = \frac{(C - C_z)}{(N - N_z)}$$

Substituindo $p = \frac{C_z}{N_z}$ e $q = \frac{(C-C_z)}{(N-N_z)}$ em $L(z;p;q)$:

$$L(z) = \frac{{C_z}^{C_z} e^{-C_z}}{C_z!} \frac{(C - C_z)^{(C-C_z)} e^{-(C-C_z)}}{(C - C_z)!} = \frac{{C_z}^{C_z} (C - C_z)^{(C-C_z)} e^{-C}}{C_z!\,(C - C_z)!}$$

Então, a razão de verossimilhança (do inglês *likelihood ratio*) para o modelo Poisson será:

$$LR(z) = \frac{L}{L_0} = \begin{cases} \left(\frac{C_z}{\mu_z}\right)^{C_z} \left(\frac{C-C_z}{C-\mu_z}\right)^{C-C_z}, & C_z > \mu_z \\ 1, & \text{c. c.} \end{cases} \tag{2.4}$$

Essa razão maximizada sobre todas as zonas identificará o conglomerado mais verossímil. Então a estatística do teste será:

$$T = \max_z LR(z)$$

Como o valor dessa razão tende a crescer muito rápido, pode-se usar o seu logaritmo, isto é, $LLR(z) = logLR(z)$. Como a função logaritmo é estritamente crescente, então o valor que maximiza $LR(z)$ também maximiza $LLR(z)$. Tem-se então a seguinte equação:

$$LLR(z) = \begin{cases} C_z(logC_z - log\mu_z) + (C - C_z)[\log(C - C_z) - \log(C - \mu_z)], & C_z > \mu_z \\ 0, & \text{c. c.} \end{cases}$$

E a estatística do teste:

$$T = \max_z LLR(z) \tag{2.5}$$

Como não se conhece a distribuição da estatística *T*, será feito um processo de simulação para se obter a distribuição empírica, possibilitando o processo de inferência. Esse processo será detalhado mais a frente.

De posse do conhecimento de como calcular a estatística *T*, pode-se então iniciar o algoritmo, que será explicado a seguir.

2.3 A MATRIZ DE DISTÂNCIAS

O primeiro passo do algoritmo é a obtenção da matriz de distâncias.

Considere um mapa dividido em n regiões, cada qual com uma população n_i e um número de casos c_i, $i = 1,2,\dots,n$. Cada região r_i possui também um centroide com coordenadas geográficas (x_i, y_i).

A distância entre dois centroides será dada pela equação de distância entre dois pontos. Então para dois centroides i e j quaisquer, teremos a seguinte expressão:

$$D_{i,j} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}$$

A matriz de distâncias entre os centroides será uma matriz quadrada com n linhas e n colunas, onde n é o número de regiões do banco de dados estudado. Seja $d_{i,j}$ o elemento da *i*-ésima linha e *j*-ésima coluna dessa matriz. Então:

$$d_{i,j} = \begin{cases} D_{i,j}, & i \neq j \\ 0, & \text{c. c.} \end{cases}$$

Isto é, $d_{i,j}$ será a distância entre os centroides i e j. Logo a matriz de distâncias será:

$$D = \begin{bmatrix} 0 & D_{1,2} & \cdots & D_{1,j} & \cdots & D_{1,n} \\ D_{2,1} & 0 & \cdots & D_{2,j} & \cdots & D_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ D_{i,1} & D_{i,2} & \cdots & 0 & \cdots & D_{i,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ D_{n,1} & D_{n,2} & \cdots & D_{n,j} & \cdots & 0 \end{bmatrix} \tag{2.6}$$

O próximo passo será encontrar as zonas candidatas a conglomerados, calculando a estatística LLR para cada uma delas, como será visto a seguir.

2.4 ENCONTRANDO CANDIDATOS A CONGLOMERADOS

Esse processo se inicia selecionando uma região e selecionando as regiões mais próximas para formar uma zona. Por exemplo, iniciando pela região r_1, temos o seguinte vetor coluna correspondente às distâncias para as outras regiões:

$$\begin{bmatrix} 0 \\ D_{2,1} \\ D_{3,1} \\ \vdots \\ D_{n,1} \end{bmatrix}$$

Seja $\mathrm{D}_{(\mathrm{j}),\mathrm{i}}$ a distância entre a região r_i e a *j*-ésima região mais próxima de *i,* denotada por $r_{(j)}$. Isto é, $D_{(n),i} > \cdots > D_{(3),i} > D_{(2),i}$. Tem-se então o vetor coluna anterior ordenado em ordem crescente de distâncias:

$$\begin{bmatrix} 0 \\ D_{(2),1} \\ D_{(3),1} \\ \vdots \\ D_{(n),1} \end{bmatrix}$$

A primeira zona selecionada será formada somente pela região r_1, ou seja, $z_1 = \{r_1\}$. Calcula-se então o valor inicial da estatística *T*, representada por $T = LLR(1)$, conforme mostrado anteriormente na seção **2.2**.

A segunda zona pela região r_1 e também pela região mais próxima da região r_1, isto é, a região correspondente à distância $D_{(2),1}$. Essa zona será representada por $z_2 = \{r_1, r_{(2)}\}$. Com os dados dessa zona selecionada, calcula-se a estatística $LLR\{1,(2)\}$. Como visto anteriormente, a estatística *T* é o valor máximo de $LLR(z)$. Então, se o valor da estatística dessa zona supera o da anterior, esta será considerado, ao menos temporariamente, o valor da a estatística *T*.

O processo iterativo se repete para outras zonas até atingir um tamanho máximo de população[1], sempre comparando a estatística $LLR(z)$ da zona atual com o máximo obtido até o momento. A Figura 2 mostram esse processo, com a janela circular varrendo o mapa em busca das zonas candidatas a conglomerados.

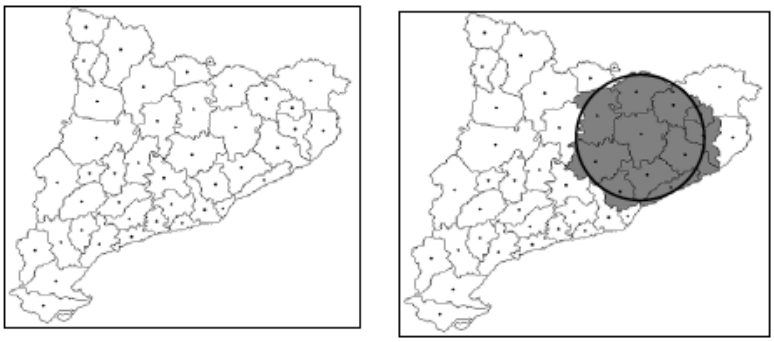

Figura 2: Visualização gráfica do *Scan* Circular

Os procedimentos anteriores são então repetidos, iniciando em cada região. O conglomerado mais verossímil será a zona z^* que corresponder ao maior valor de *LLR(z),* obedecendo o limite de população máxima, e o valor da estatística de teste será $T = \mathrm{LLR}(z^*)$.

2.5 VERIFICAÇÃO DA SIGNIFICÂNCIA DO CONGLOMERADO

Após a execução do algoritmo para a detecção do conglomerado mais verossímil, é necessário testar sua significância. Entretanto, como não se conhece a distribuição da estatística *T*, será utilizada uma simulação de Monte Carlo para isso.

Segundo Figueiredo (2010, p.5), "essa simulação consiste em construir milhares de réplicas do mapa original em que o número total de casos *C* está fixo e os casos em cada região são distribuídos aleatoriamente sob H_0."

[1] Não faz sentido obter um conglomerado que tenha uma grande proporção da população total. A literatura recomenda que o conglomerado obtido tenha no máximo 50% da população total do mapa.

Sob H_0, a probabilidade de que um indivíduo seja um caso é igual em todo o mapa e assume valor $p = \frac{C}{N}$. Então o número de esperado de casos em uma região r_i qualquer é proporcional à sua população, isto é, $\mu_i = pn_i$, onde n_i é a população dessa região.

Sendo assim, são gerados *C* casos em cada réplica, distribuindo-os aleatoriamente nas regiões, considerando a hipótese nula de proporcionalidade quanto às populações.

Esse procedimento será repetido um grande número de vezes, sendo que em cada réplica também será calculada a estatística *T* como explicado na seção **2.4**. Segundo Figueiredo (2010, p.5), "para cada réplica teremos um valor da estatística *T* e o conjunto delas obtido gerará uma distribuição empírica da estatística de teste".

A partir da distribuição empírica pode-se então calcular o *p-valor* utilizando a teoria estatística. Utilizando nível de significância em 0,05, rejeita-se a hipótese nula de que a zona encontrada não é um conglomerado se

$$P(X > T) \leq 0{,}05$$

onde *X* tem a distribuição empírica gerada pela simulação. Neste caso, o conglomerado encontrado é significativo.

2.6 RESUMO DO ALGORITMO

O algoritmo *Scan* Circular pode ser resumido nos seguintes passos, explicados com detalhes anteriormente:

1. Encontrar a matriz de distâncias, de acordo com a seção **2.3.**
2. Partindo da *i*-ésima região, ordenar as regiões segundo a distância até a região r_i.
3. Encontrar as zonas candidatas e calcular a estatística $LLR(z)$ de acordo com a seção **2.4**, registrando a zona com o maior valor de $LLR(z)$ até o momento e respeitando o limite de proporção da população.

4. Repetir os passos 2 e 3 para todas as regiões do mapa, isto é, para $i = 1,2, \dots, n$.
5. Utilizar a simulação descrita na seção **2.5** para avaliar a significância do teste.
6. Rejeitando a hipótese nula no passo anterior, deve-se se armazenar essa zona, confeccionando um mapa que a destaque.

2.7 ADAPTAÇÃO PARA DADOS PONTUAIS

O algoritmo *Scan* Circular se aplica de forma similar em análise com dados pontuais. Nesse caso, ao invés das coordenadas dos centroides, são conhecidas as coordenadas de cada ocorrência do evento de interesse (ponto) e uma zona é a região delimitada pela janela circular.

Defina o_i como a *i*-ésima ocorrência no mapa. *A* matriz de distâncias apresentará a distância entre cada ponto. Sendo assim, a distância definida por $D_{(j),i}$ será *j*-ésima ocorrência mais próxima da ocorrência o_i. Observando essa diferença, o algoritmo para dados pontuais pode ser resumido nos seguintes passos:

1. Encontrar a matriz de distâncias, de acordo com a seção **2.3.**
2. Partindo da *i*-ésima ocorrência, ordenar as ocorrências segundo a distância até a ocorrência o_i.
3. Encontrar as zonas candidatas e calcular a estatística $LLR(z)$ de acordo com a seção **2.4**, registrando a zona com o maior valor de $LLR(z)$ até o momento e respeitando o limite de proporção da população.
4. Repetir os passos 2 e 3 para todas as ocorrências no mapa isto é, para $i = 1,2, \dots, n$.
5. Utilizar a simulação descrita na seção **2.5** para avaliar a significância do teste.
6. Rejeitando a hipótese nula no passo anterior, deve-se se armazenar essa zona, confeccionando um mapa que a destaque.

3. MATERIAL E MÉTODOS

Esse capítulo servirá de base para a aplicação que será feita no capítulo seguinte. Serão detalhados o problema e conceitos relacionados, o banco de dados a ser utilizado e também a macro implementada no *software* SAS 9.2 que será aplicada diretamente ao problema no próximo capítulo.

3.1 O PROBLEMA

A teoria vista anteriormente servirá como base para a aplicação da técnica *Scan* Circular em um banco de dados real com objetivo de detectar um conglomerado espacial de fraudes bancárias na região Centro-Oeste do Brasil, correspondente aos estados de Mato Grosso, Mato Grosso do Sul, Goiás e o Distrito Federal.

A técnica *Scan* Circular já foi utilizada em criminologia (BEATO FILHO et. al. 2001), e nesse caso específico pode apresentar resultados que ajudariam a instituição financeira a preparar e aplicar planos de prevenção em uma determinada área onde o número de fraudes foi superior ao esperado, podendo assim reduzir prejuízos.

3.2 BANCO DE DADOS

O banco de dados utilizado foi fornecido por uma grande instituição financeira brasileira e possui dados de 187 agências na região Centro-Oeste do Brasil, com as respectivas quantidades de contas e quantidades de fraudes relacionadas à operações com cartão de débito durante o ano de 2010. As fraudes identificadas no banco correspondem somente a dispositivos físicos.

Foi também informado o CEP de cada agência. Esse dado foi utilizado para se obter as coordenadas geográficas necessárias para a aplicação do algoritmo *Scan* Circular. No caso desse problema, foram obtidos os dados de latitude e longitude das agências. As coordenadas seguem o padrão *Global Positioning System* (GPS), ou Sistema de Posicionamento Global, que será detalhado posteriormente.

Uma primeira análise foi realizada com os dados pontuais de agências bancárias. A tabela 3.1 mostra a quantidade de agências por município.

Tabela 4.10 – Número de agências por município na região Centro-Oeste

Município	**Nº de agências**
Brasília	49
Goiânia	21
Campo Grande	10
Cuiabá	9
Anápolis	5
Aparecida de Goiânia	2
Outros	91

A quantidade de fraudes por agência será considerada o número de casos e o número de contas em cada agência funcionará como uma variável de controle.

É importante notar que em casos como esse, a variável que seria considerada como “população” pode ter valores menores que a variável considerada como “caso” em cada observação. Por esse motivo o número de contas não será chamada de “população em risco”, como no caso de utilização para detecção de conglomerados de doenças, e sim de controle.

Essa variável será essencial para a obtenção do valor esperado de casos em cada agência, sob a hipótese nula de que não existem conglomerados.

Para uma segunda análise, foram obtidos dados geográficos referentes aos 446 municípios da região Centro-Oeste possibilitando a confecção de um

novo banco de dados. Os dados anteriores foram agregados por município, possibilitando uma melhor visualização de um suposto conglomerado.

Na análise por dados agregados, o número de contas por município foi considerada a variável de controle e a quantidade de fraudes por município foi considerada o número de casos. Os centroides serão os centros políticos de cada município.

3.3 FRAUDES

Com a grande competitividade atual, as instituições financeiras brasileiras aprimoram seus produtos e serviços e colocam à disposição de seus clientes tecnologia de ponta, criando facilidades e conveniência no relacionamento e minimizando custos operacionais na medida em que o cliente utiliza o autoatendimento, deixando de ir a uma agência bancária para usar o caixa eletrônico em suas transações.

Os dispositivos eletrônicos utilizados no autoatendimento trazem muitas facilidades aos clientes, mas também proporcionam facilidades aos fraudadores, especialmente em relação aos crimes contra o patrimônio.

Fraude é todo artifício empregado com o fim de enganar uma pessoa e causar-lhe prejuízo e traduz a intenção de procurar uma vantagem indevida, patrimonial ou não. Entende-se por fraude bancária o ato de se obter, por meios ilícitos, dados sigilosos de clientes de uma instituição financeira e com isso efetuar movimentação financeira sem autorização.

Existem diversas técnicas para realizar a fraude bancária, podendo ocorrer tanto fisicamente em agências e terminais eletrônicos, quanto por meio da internet. Com o desenvolvimento de novas tecnologias de segurança, os fraudadores também evoluem suas técnicas e utilizam conceitos avançados de programação de sistemas para realizar os crimes.

No caso de operações com cartões, os fraudadores utilizam sistemas físicos para obter as trilhas e senhas dos clientes e a partir daí fabricam cartões clonados e passam a efetuar movimentações financeiras nas contas dos clientes.

A clonagem da trilha do Cartão de Débito em um terminal adulterado é a modalidade mais utilizada e a que gera maiores prejuízos às instituições financeiras. Consiste na instalação de dispositivos no terminal do banco ou de sua rede conveniada que gravam os dados do cliente durante um acesso autêntico.

A adulteração dos terminais dos bancos pode ocorrer internamente através de dispositivos conhecidos como “chupa-cabras” que são instalados entre a CPU, a leitora magnética e o teclado do equipamento, mas também podem ser colados na parte externa do equipamento, com o uso de dispositivos conhecidos como “luvas”.

Outra forma de adulterar um terminal consiste na modalidade denominada “golpe do notebook”, em que o criminoso cria uma máscara idêntica à carenagem do terminal da instituição financeira contendo um notebook com aplicação que simula a aplicação do banco e cola na parte externa do terminal do banco. Neste caso o cliente acredita estar utilizando um terminal do seu banco, mas está somente fornecendo seus os dados, uma vez que a transação não é concluída, mas suas senhas foram capturadas pela aplicação do fraudador.

Diariamente são noticiadas ações dos fraudadores na imprensa brasileira, indicando que a atuação das quadrilhas ocorre em todo o país. Em muitos casos há sucesso nas operações policiais e são noticiadas prisões de fraudadores.

Além da ação policial, as instituições financeiras buscam constantemente melhorar a segurança das transações dos clientes e implantam soluções de segurança a exemplo da senha denominada identificação positiva (“sopa de letrinhas” ou a sílaba), *tokens* com senha aleatórias e a substituição do cartão de débito comum pelo cartão com chip.

Com a implementação de soluções de segurança o número de fraudes tende a reduzir, mas ainda persiste, pois há constante alteração do perfil da fraude, o que exige das instituições financeiras a atuação em busca de

melhorias para o seu combate, sendo imprescindível conhecer a fundo as ocorrências de fraude.

Sendo a fraude um risco operacional que as instituições financeiras estão expostas e devido ao potencial nível de perdas, o Banco Central do Brasil determinou na Resolução BACEN 3380/2008 que os bancos implementem estrutura de gerenciamento do risco operacional para identificar, avaliar, monitorar, controlar e mitigar o risco operacional de fraudes e outros eventos de interesse.

Assim, os resultados desse trabalho poderão auxiliar para que a instituição financeira que cedeu os dados da fraude obtenha informações sobre as áreas de maior risco e possa implementar medidas preventivas e minimizar os prejuízos decorrentes das fraudes.

3.4 COORDENADAS GEOGRÁFICAS

Para referências no sistema de coordenas geográficas, são traçados, em torno da Terra, linhas imaginárias conhecidas como paralelos e meridianos. Os paralelos são linhas paralelas à linha do Equador (que também é um paralelo), e os meridianos são linhas semicirculares que conjuntamente com seus antemeridianos formam circunferências. O meridiano principal é o meridiano de *Greenwich,* estipulado por convenção.

Para a localização de um ponto no globo terrestre, são utilizados como base os paralelos e os meridianos, e a partir deles são estipuladas as coordenadas de latitude e longitude. As latitudes podem variar de 0 a 90 graus para o Norte ou para o Sul, e as longitudes de 0 a 180 graus para Leste ou Oeste, utilizando como referência, respectivamente, o Equador e o meridiano de *Greenwich*.

Existem diversas maneiras de escrever as coordenadas geográficas. Uma das mais tradicionais é o sistema de graus – minutos – segundos. Nesse caso, por exemplo, o morro do Pão de Açúcar, localizado no Rio de Janeiro, possui as seguintes coordenadas: 22º 56' 48.2136"S 43º 09' 21.456" W.

As coordenadas também podem ser escritas em forma decimal, onde cada grau é dividido em frações decimais e a escrita também é diferenciada. Nesse caso, a latitude é abreviada como *lat* e a longitude como *long*. Os pontos ao Sul, no caso da latitude, e ao Oeste, no caso da longitude, recebem valores negativos.

Para esse sistema, a localização do morro do Pão de Açúcar é *lat* -22.946697, *long* -43.15596. Os dados geográficos dos centroides obtidos para o trabalho também seguem o padrão decimal.

3.5 MACRO %SCANCIRCULAR

O algoritmo *Scan* Circular foi adaptado utilizando o *software* SAS 9.2, utilizando o procedimento IML (*Interactive Matrix Language*) que permite desenvolver programas em linguagem matricial. Para algumas análises foram utilizados os procedimentos UNIVARIATE e o GPLOT.

Também foi utilizado o procedimento GMAPS, presente no pacote SAS/GRAPH, para a confecção dos mapas utilizados na análise.

Foi criada a seguinte macro para a realização da detecção de conglomerados pelo método *Scan* Circular:

```
%scancircular(data=,pop=,cases=,x=,y=,nsim=);
```

Essa macro realiza todo os passos do algoritmo *Scan* Circular, incluindo a verificação da significância do conglomerado indicado. Para executá-la, é necessário a indicação dos seguintes parâmetros:

Data é o banco de dados a ser utilizado; **pop** é o nome da variável no banco de dados referente à informação da população considerada; **cases** é o nome da variável no banco de dados referente à informação do número de casos; **x** é o nome da variável no banco de dados referente à informação de longitude; **y** é o nome da variável no banco de dados referente à informação de latitude; **nsim**, cujo valor refere-se ao número de simulações de Monte Carlo.

A macro pode ser dividida em duas partes: a busca do conglomerado e seus respectivos cálculos e a simulação de Monte Carlo para verificar a significância do conglomerado encontrado.

3.5.1 A busca do conglomerado

Considerando **a** uma matriz de tamanho $n\ x\ 4$, onde n é o número de observações do banco de dados e o parâmetro **pmax** correspondente a 50% da população total do banco, a primeira parte da macro pode ser resumida com os seguintes passos:

1. Faça $a \leftarrow data$
2. Faça $llr^* \leftarrow 0$ e $z^* \leftarrow \{\}$
3. Calcular a matriz $D_{n\ x\ n}$ onde cada elemento $D_{i,j}$ corresponde a distância da região r_i à região r_j.
4. Para a região r_i, faça $z \leftarrow r_i$.
5. Ordenar o *i*-ésimo vetor coluna da matriz *D* em ordem ascendente, onde o *j*-ésimo elemento corresponde à distância da *j*-ésima região, $r_{(j)}$, mais próxima de r_i.
6. Definir $z \leftarrow z \cup \{r_{(j)}\}$.
7. Se $\{llr(z) > llr^*$ e $pop(z) \leq pmax\}$ então $\{llr^* = llr(z)$ e $z^* = z\}$.
8. Repetir os passos 6 e 7 para todo *j,* até encontrar o conglomerado com maior *LLR* dentre todos os possíveis conglomerados, respeitando a condição $pop(z) \leq pmax$.
9. Volte ao passo 4 para cada região r_i, *i*=1,...,n.

A macro apresenta o conglomerado mais verossímil entre todos os possíveis, com o resultado das seguintes variáveis:

Population_total, a população total; **population_max**, a população máxima; **total_cases**, o número de casos total; **t**, estatística $T = \max_z LLR(z)$; **cluster_population**, a população presente no conglomerado; **cluster_size**, o número de regiões presentes no conglomerado; **expected_cases**, o número de casos esperados no conglomerado; **observed_cases**, o número de casos presentes no conglomerado.

Para efetuar os cálculos, foi utilizado o procedimento IML.

Além disso, é confeccionado um gráfico primário relativo ao mapa, para uma auxiliar primeira visualização. As regiões incluídas no conglomerado são destacadas das demais e representadas por círculos. Para isso, foi utilizado o procedimento GPLOT.

3.5.2 Simulação

A segunda parte da macro é mais intensiva computacionalmente. São criados *nsim* novos bancos de dados, e em cada um é calculado o valor de *LLR*. O valor de *nsim* deverá ser razoavelmente grande. Geralmente é utilizada cerca de 1000 repetições. Ao final do processo, é obtida a distribuição empírica dessa estatística.

É importante relembrar as hipóteses do teste:

$$\begin{cases} H_0: p = q \\ H_1: p > q \end{cases}$$

Onde, p é a probabilidade de ocorrer um controle na zona z e q a probabilidade de ocorrer fora da zona *z*. Uma zona será considerada um conglomerado caso o teste rejeite a hipótese nula.

Considerando **total_cases** como o número de casos total do banco de dados utilizado, a segunda parte da macro pode ser resumida com os seguintes passos:

1. Calcule a função de distribuição acumulada da população para cada observação. Considere k_i como o valor de distribuição acumulada até a observação i.
2. Gere um novo banco de dados onde todas as observações possuem 0 casos.
3. Gere **total_cases** números aleatórios no intervalo [0,1]
4. Considere w um número gerado no passo anterior. Atribuir um caso para a observação i se $k_i < w \leq k_{i+1}$.
5. Com o novo banco de dados gerado, executar o algoritmo *Scan* Circular para calcular o *LLR* como na seção **3.5.1**. Guardar esse valor.

6. Repetir os passos 2 até 5 um número *nsim* de vezes e construir a distribuição empírica de *LLR*.
7. Rejeita-se a hipótese com 95% de confiança se $T > P_{95}$ onde *T* é a estatística de teste obtida na seção **3.5.1** e P_{95} é o 95º percentil da distribuição empírica de LLR.

Ao final do processo, a macro apresentará o valor da estatística *T*, os percentis de interesse, a média da distribuição empírica e a decisão do teste de hipóteses com 95% de confiança. Além disso, é gerado um histograma da distribuição.

Para os cálculos da simulação, foi utilizado o procedimento IML, enquanto a análise da distribuição empírica e a confecção do histograma foi realizada com auxílio do procedimento UNIVARIATE.

3.5.3 Exemplo

As Figuras 4, 5, 6 e 7 apresentam as saídas da macro *%scancircular* utilizando dados reais. Nesse caso, foram utilizados dados referentes a casos de doença de Chagas em mulheres que tiveram filhos no período de Julho a Setembro de 2006, no estado de Minas Gerais, dividido em 853 municípios. A população de risco é de 63519 mulheres e foram notificados 803 casos. Foram feitas 200 simulações para a construção da distribuição empírica.

Nesse caso em específico a hipótese nula do teste foi rejeitada, então a zona mais verossímil encontrada (identificada por regiões com círculos) na primeira parte da macro constitui uma aglomeração anormal de casos.

```
%scancircular(data=chagasmg,pop=pop,cases=cas,x=x,y=y,nsim=200) ;
```

population_total	population_max	total_cases
63519	31759.5	803

t	cluster_population	cluster_size
255.66818	17400	92

expected_cases	observed_cases
219.96883	528

Figura 4: Saída da macro *%scancircular* – Informações de interesse.

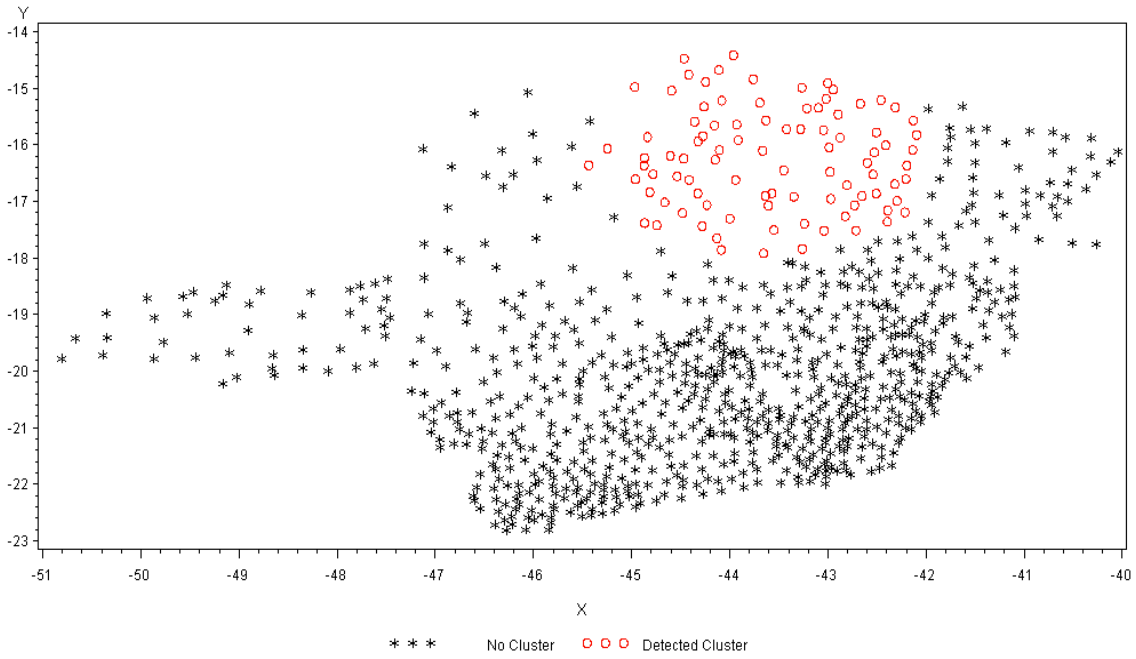

Figura 5: Saída da macro *%scancircular* – Mapa simples do estado de Minas Gerais, com o conglomerado detectado em destaque.

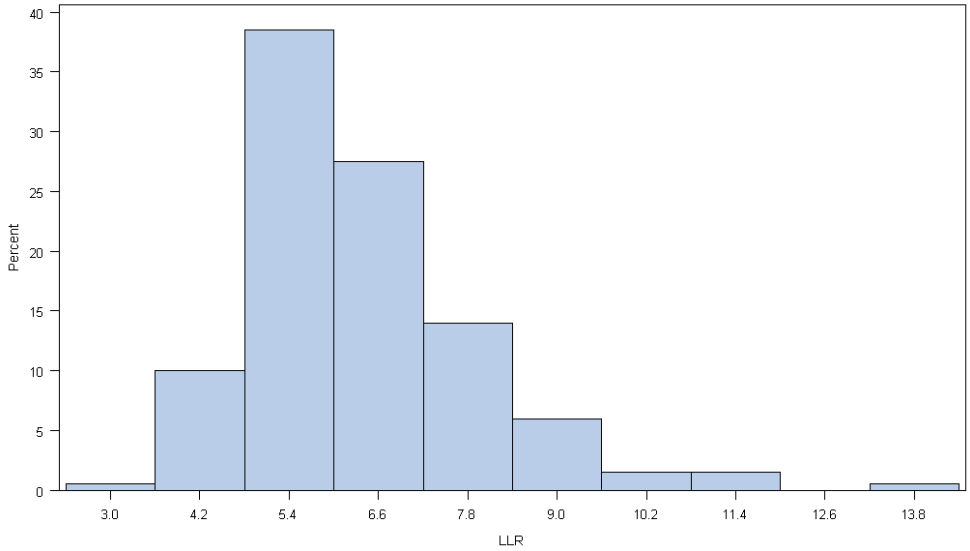

Figura 6: Saída da macro *%scancircular* – Histograma da distribuição empírica para dados de doença de Chagas, utilizando 200 repetições Monte Carlo.

t	p95	decision
255.668	9.31726	Reject

Figura 7: Saída da macro *%scancircular* – Informações da distribuição empírica e decisão do teste de hipótese.

4. RESULTADOS

Nesse capítulo serão apresentados os resultados da aplicação do algoritmo *Scan* Circular para os dados de fraude no Centro-Oeste do Brasil, além de outras análises auxiliares para a conclusão.

O banco é composto por 187 agências, com um número total de 2.986.049 contas e 6.051 fraudes. Os dados são referentes a fraudes em cartões de débito no ano de 2010.

4.1 DADOS PONTUAIS

Utilizando dados pontuais das 187 agências bancárias na região Centro-Oeste, o algoritmo indicou uma zona como um suposto conglomerado bastante concentrado em agências da cidade de Brasília e no seu entorno. A zona engloba 62 agências e possuiu uma estatística $T = \max_z LLR(z) = 839{,}555$.

Foram confeccionados dois mapas. A Figura 8 apresenta todos os estados da região Centro-Oeste e a Figura 9 somente o estado de Goiás e Distrito Federal. As agências foram marcadas nos mapas, e os pontos que foram incluídos no suposto conglomerado aparecem em destaque na forma de estrela. Por ser bastante concentrado, o conglomerado é melhor visualizado no segundo mapa (Figura 9).

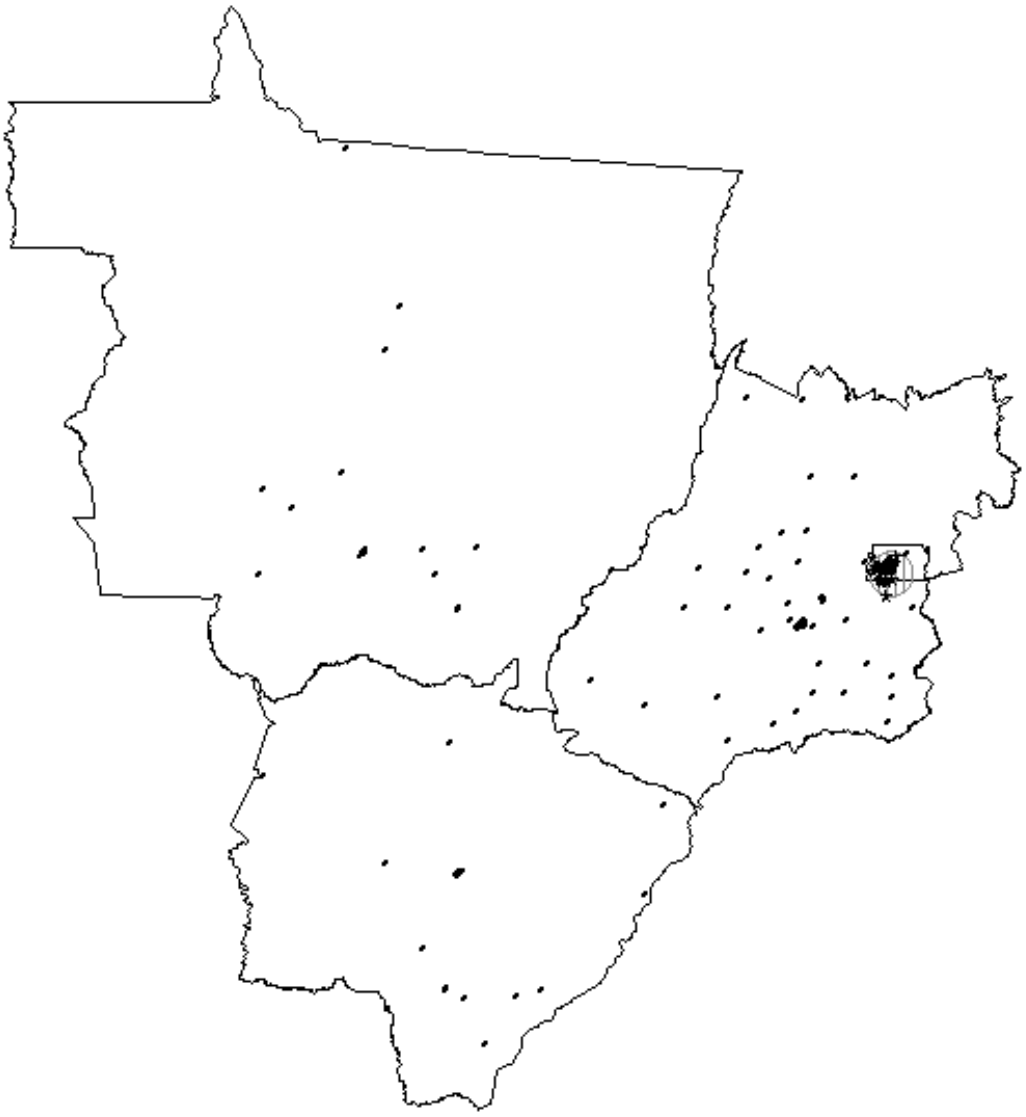

Figura 8: Mapa da região Centro-Oeste do Brasil com o suposto conglomerado identificado em destaque.

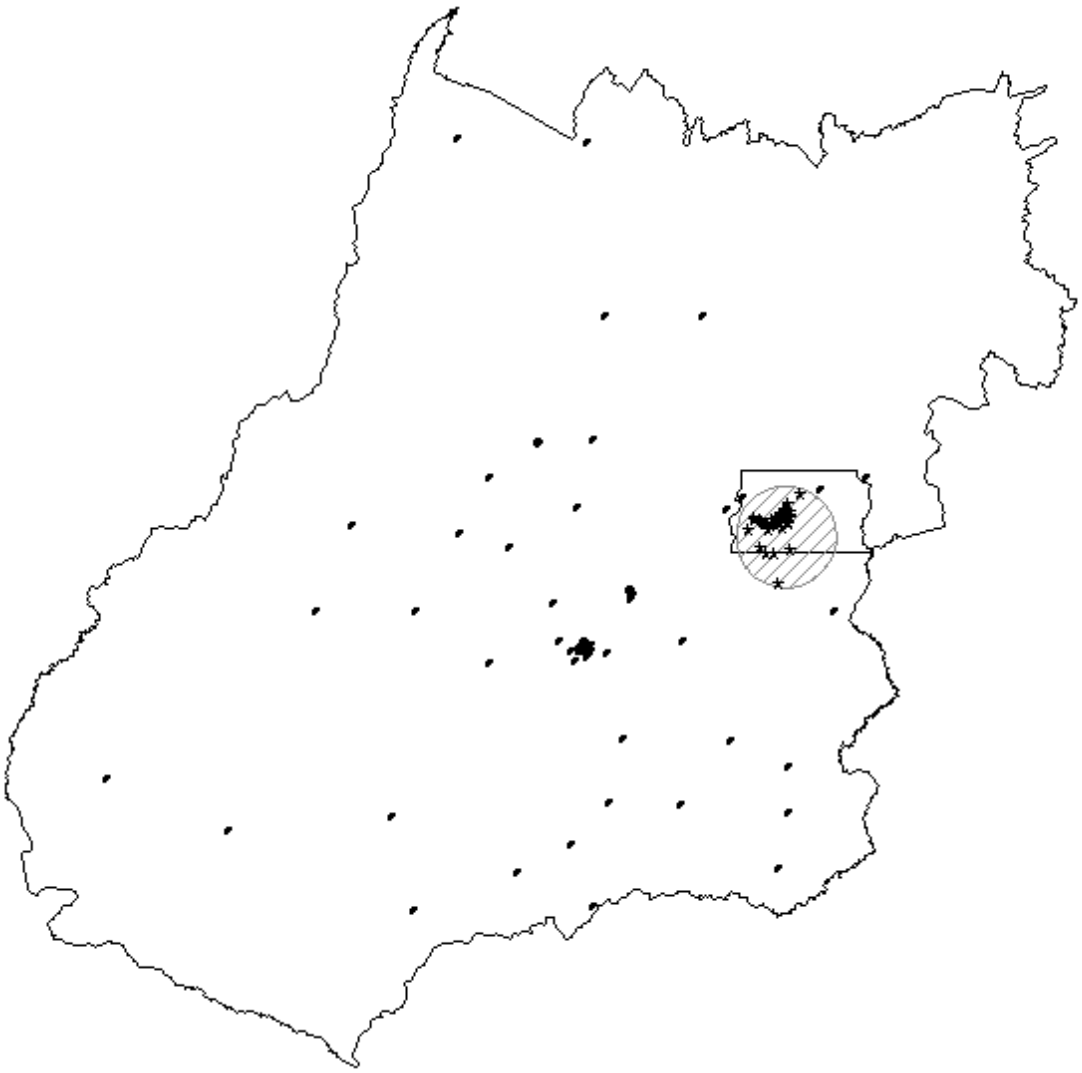

Figura 9: Mapa do estado de Goiás e Distrito Federal, com o suposto conglomerado identificado em destaque.

A Tabela 4.1 apresenta os dados de número de contas para essa análise. Foi delimitado um máximo de 50% do número de contas(1.493.024,5) dentro do conglomerado. O suposto conglomerado obtido possui 863.426 contas.

Os dados de casos de fraudes para dados pontuais são mostrados na Tabela 4.2. Verificou-se que o número observado de fraudes no suposto conglomerado detectado é 87% maior do que o esperado sob a hipótese nula

de que não existem conglomerados, mostrando uma evidência, a ser comprovada, de que a zona detectada é um conglomerado atípico.

Tabela 4.1 – Número de contas para dados pontuais

Número de contas (Total)	Número máximo de contas	Número de contas na zona detectada
2.986.049	1.493.024,5	863.426

Tabela 4.2 – Número de fraudes para dados pontuais

Número de fraudes (total)	Número esperado de fraudes na zona	Número observado de fraudes na zona
6.051	1.749,7	3.277

Para verificar a significância desse suposto conglomerado, foi gerada uma distribuição empírica da estatística LLR. A Figura 10 mostra o histograma da distribuição gerada por meio de uma simulação de Monte Carlo com 1000 réplicas.

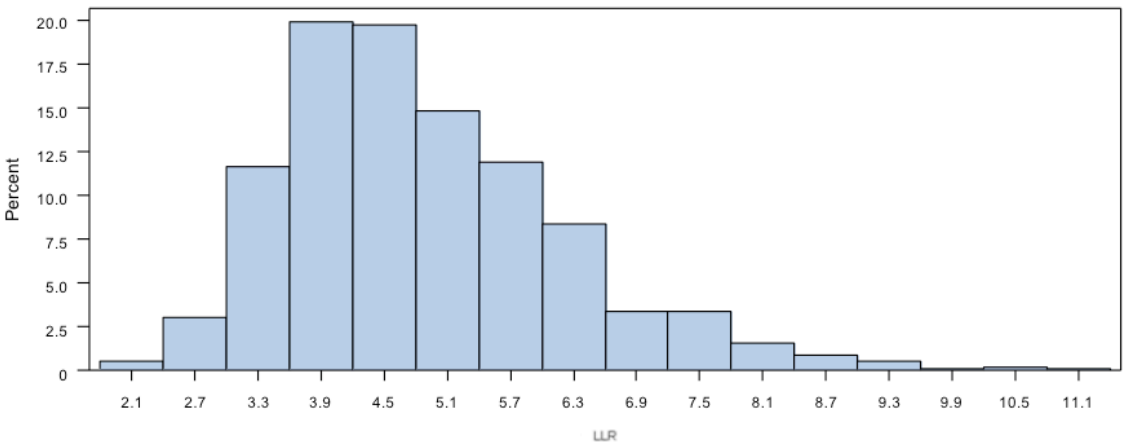

Figura 10: Histograma da distribuição empírica de *LLR* para dados pontuais de fraude.

A Tabela 4.3 apresenta os valores da estatística *LLR* observada nos dados e o percentil 95 da distribuição empírica, para efeitos de comparação.

Tabela 4.3 – Estatísticas para dados pontuais

LLR Observada	Percentil 95
839,555	7,49678

Esses valores mostram que, a um nível de significância de 5%, a zona detectada é um conglomerado atípico, já que $LLR_{obs} > P_{95}$.

Para um melhor entendimento e visualização do conglomerado, foi feita uma segunda análise utilizando dados de todos os municípios do Centro-Oeste, como mostrado na seção 4.2.

4.2 DADOS AGREGADOS POR MUNICÍPIOS

Os dados pontuais de agências bancárias foram agregados por município, gerando um novo banco de dados com informações referentes aos 446 municípios da região Centro-Oeste.

O algoritmo indicou uma zona como um suposto conglomerado nos seguintes municípios: Brasília, Águas Lindas de Goiás, Luziânia, Valparaíso de Goiás, Santo Antônio do Descoberto, Novo Gama e Planaltina. A estatística de teste para essa zona foi $T = \max_z LLR(z) = 757{,}844$

Assim como na análise anterior, foram confeccionados dois mapas, um contendo todos os estados da região Centro-Oeste e outro somente com o estado de Goiás e Distrito Federal, apresentados nas Figuras 11 e 12. Os limites municipais foram explicitados nos dois mapas. No segundo mapa (Figura 12) é possível ver também os centroides atribuídos a cada município.

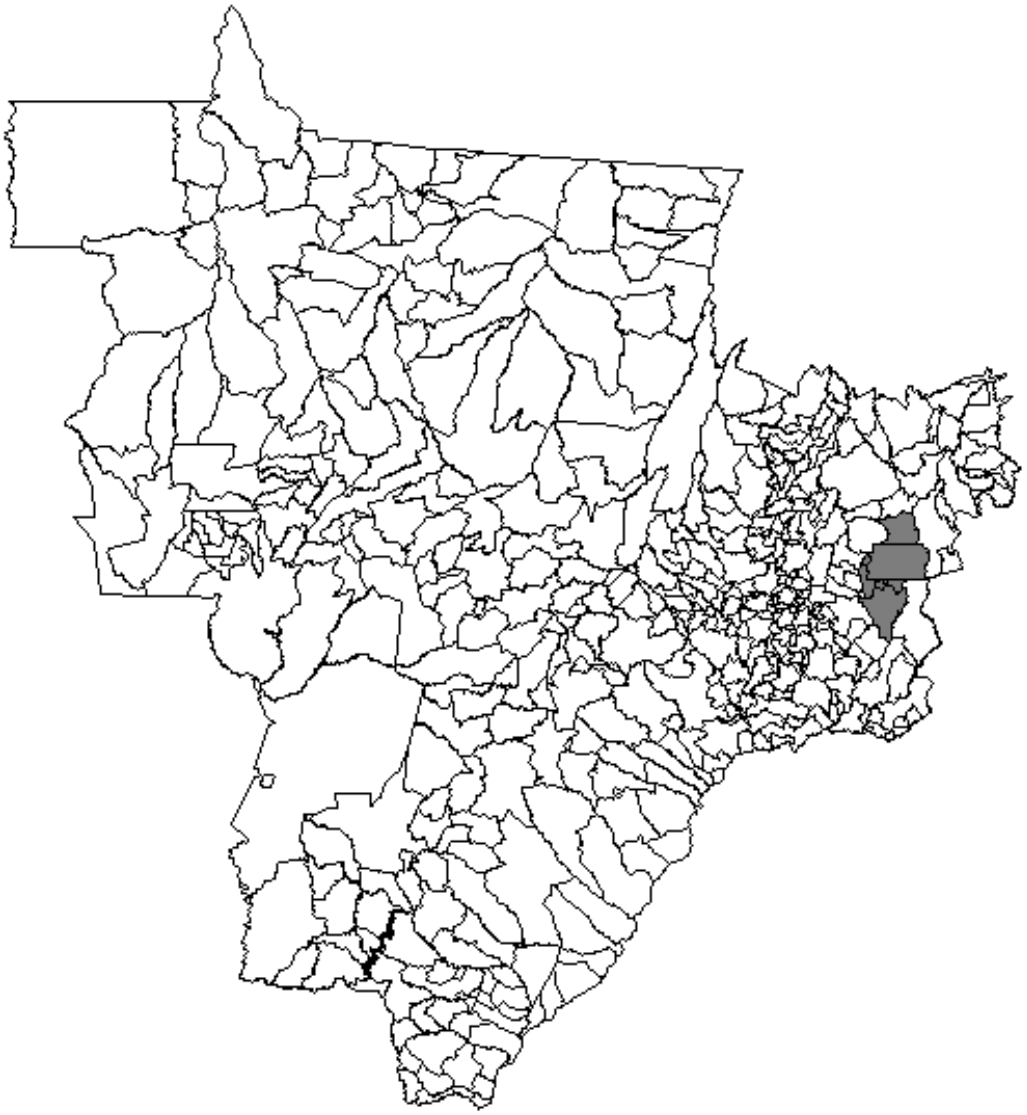

Figura 11: Mapa da região Centro-Oeste do Brasil com o suposto conglomerado identificado em destaque.

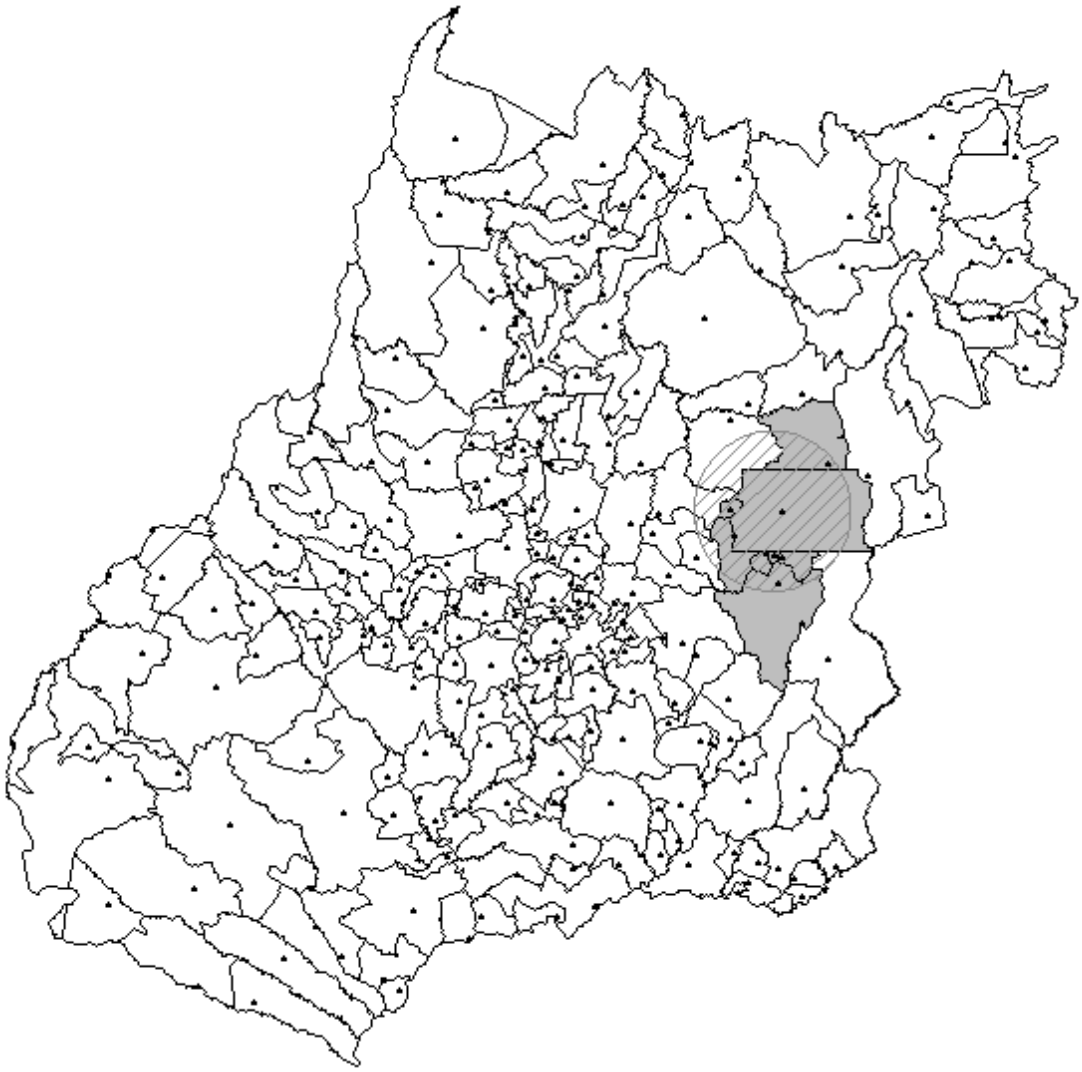

Figura 12: Mapa do estado de Goiás e Distrito Federal, com o suposto conglomerado identificado em destaque.

A Tabela 4.4 apresenta os dados do número de contas para essa análise. Novamente, foi delimitado um máximo de 50% do número de contas(1.493.024,5) dentro do conglomerado. O suposto conglomerado detectado possuiu 974.697 contas.

Os dados de casos de fraudes para dados agregados são mostrados na Tabela 4.5. Verificou-se nesse caso que o número observado de fraudes no suposto conglomerado detectado é cerca de 74% maior do que o esperado sob

a hipótese nula de que não existem conglomerados, mostrando uma evidência, a ser comprovada, de que a zona detectada é um conglomerado atípico.

Tabela 4.4 – Número de contas para dados agregados

Número de contas (Total)	Número máximo de contas	Número de contas na zona detectada
2.986.049	1.493.024,5	976.497

Tabela 4.5 – Número de fraudes para dados agregados

Número de fraudes (total)	Número esperado de fraudes na zona	Número observado de fraudes na zona
6.051	1.978,7965	3.456

Para verificar a significância desse suposto conglomerado, foi gerada uma distribuição empírica da estatística $T = \max_z LLR(z)$. A Figura 13 mostra o histograma da distribuição gerada por meio de uma simulação de Monte Carlo com 1000 réplicas.

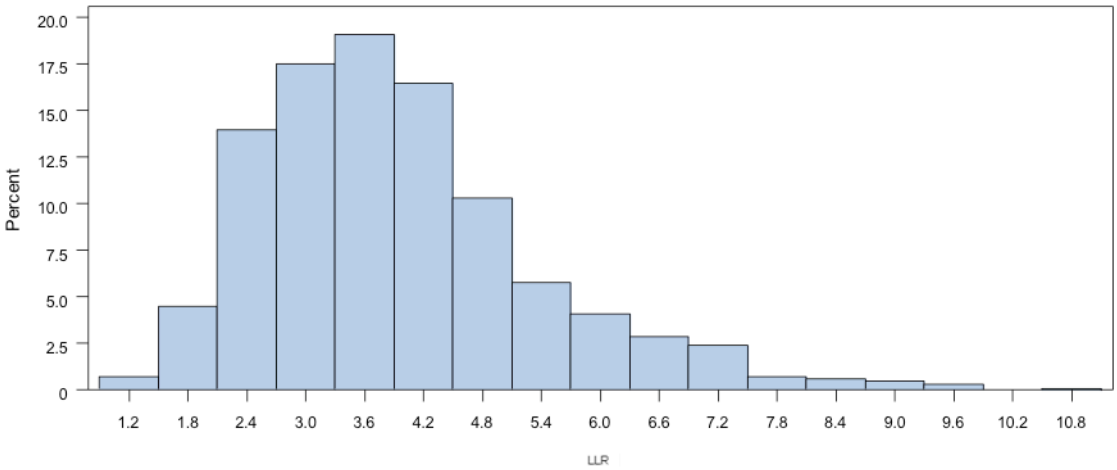

Figura 13: Histograma da distribuição empírica de *LLR* para dados gerais de fraude.

A Tabela 4.6 apresenta os valores da estatística *LLR* observada nos dados e do percentil 95 da distribuição empírica, para efeitos de comparação.

Tabela 4.6 – Estatísticas para dados agregados

LLR	Percentil 95
757,844	6,75871

Esses valores mostram que, a um nível de significância de 5%, a zona detectada é um conglomerado atípico, já que $LLR_{obs} > P_{95}$.

Percebe-se que nas duas análises, tanto por dados agregados quanto para dados gerais, a localização do conglomerado foi muito próxima, abrangendo principalmente o Distrito Federal e entorno.

4.3 CONGLOMERADOS SECUNDÁRIOS

Algumas implementações do algoritmo *Scan* Circular permitem detectar, além do conglomerado mais verossímil, conglomerados secundários significativos. Tais conglomerados também podem ser de interesse do estudo. A maior dificuldade de sua implementação é a realização de inferências de múltiplos conglomerados, conforme discuto em outros artigos (ZHANG et. al. 2010; LI et. al. 2011).

Utilizando o *software* SaTScan com o banco de dados de fraudes bancárias no Centro-Oeste anteriormente utilizado em 4.2, verificou-se a existência de um conglomerado secundário formado somente pelo município de Goiânia-GO. A Tabela 4.9 apresenta os dados de contas e fraudes para o município, e a Tabela 4.10 mostra as estatísticas de interesse[2].

Como o *p*-valor da estatística *LLR* é menor que 0,05, pode-se afirmar, a um nível de significância de 5%, que o município de Goiânia também é um

2 O *software* SaTScan apresenta o *p*-valor da estatística *LLR* diretamente.

conglomerado. Esse foi o único conglomerado secundário significativo detectado.

Tabela 4.9 – Dados de contas e número de fraudes no município de Goiânia

Número de contas	Número esperado de fraudes	Número observado de fraudes
486.568	985,99	1153

Tabela 4.10 – Estatísticas de interesse relativas ao município de Goiânia

LLR	P-Valor
16,1908	0,0000046

4.4 OUTRAS ANÁLISES

A detecção do conglomerado mais verossímil contribui para o início de uma análise dos motivos pelos quais os municípios de Brasília, Águas Lindas de Goiás, Luziânia, Valparaíso de Goiás, Santo Antônio do Descoberto, Novo Gama e Planaltina formam uma área onde a probabilidade de ocorrer uma fraude é maior do que o esperado.

Para essa análise posterior, Brasília será considerada um município, de acordo com o padrão utilizado por institutos oficiais de pesquisa, como IBGE e IPEA, que não fazem a distinção de Brasília e Distrito Federal e considera a cidade como um município.

Brasília, cidade presente no conglomerado mais verossímil, possui o maior PIB dentre todos os municípios da região Centro-Oeste, incluindo as capitais. A Tabela 4.7 mostra os valores do PIB para o ano de 2008.

Sabe-se que os fraudadores buscam locais onde existe uma maior concentração e movimentação de dinheiro, logo a presença de Brasília já era de certa forma esperada.

É interessante notar também que os 4 municípios com menores PIB *per capita* da região Centro-Oeste estão presentes no conglomerado, como mostrado na Tabela 4.8.

Tabela 4.7 – Os 5 maiores PIB da região Centro-Oeste no ano de 2008

Município	**PIB**
Brasília	117.571.952
Goiânia	19.457.328
Campo Grande	10.462.086
Cuiabá	9.014.929
Anápolis	6.265.480

Fonte: www.ibge.gov.br/cidadesat/topwindow.htm?1

Tabela 4.8 – Os 4 menores PIB *per capita* da região Centro-Oeste no ano de 2008

Município	**PIB *per capita***
Águas Lindas de Goiás	3.270,76
Novo Gama	3.599,32
Santo Antônio do Descoberto	3.638,15
Cidade Ocidental	3.777,92

Fonte: www.ibge.gov.br/cidadesat/topwindow.htm?1

Isso mostra que, apesar da presença de Brasília, a distribuição de renda em praticamente todo o conglomerado é bastante ruim.

Sabe-se que esse é um dos motivos para que tais municípios possuam um alto índice de criminalidade, incluindo alguns dos maiores de toda a região Centro-Oeste, e acreditamos que esse fato também contribui para a significativa concentração de fraudes na região identificada pelo conglomerado.

5. CONSIDERAÇÕES FINAIS

5.1 CONCLUSÕES

Nesse trabalho a técnica *Scan* Circular de Kulldorff(1997) foi empregada para detecção, identificação e inferência de conglomerados espaciais. O motivo para a escolha de tal técnica foi a vantagem de se poder realizar inferências, além da ampla utilização em diversas áreas e fácil adaptação e implementação do algoritmo em diversas linguagens de programação.

Utilizando-nos dessa técnica, buscamos detectar e identificar um conglomerado de fraudes em cartão de débito na região Centro-Oeste, utilizando dados de uma grande instituição financeira. Para a realização do estudo, foi utilizado o *software* estatístico SAS, que se mostrou adequado para a obtenção dos conglomerados através da macro implementada.

Uma das dificuldades que encontramos para a utilização do banco de dados foi o fato de possuirmos somente a informação do CEP de cada agência, e não sua localização geográfica. Foi utilizada uma macro em linguagem SAS que acessa o banco de dados do *site Google Maps* e, a partir do CEP, obtém as informações necessárias.

Além disso, existe dificuldade na implementação do algoritmo que permite a detecção de conglomerados secundários na macro *%scancircular*, bem como a integração do procedimento GMAPS que permite a confecção automática dos mapas. Para o a utilização no trabalho o referido procedimento necessitou ser executado em separado, após a execução da macro.

Os dados obtidos dessa instituição bancária revelam que o conglomerado mais verossímil é formado pelo Distrito Federal e municípios da região conhecida como Entorno. Tal resultado foi obtido tanto por meio de análise para dados pontuais de agências bancárias quanto por meio de dados agregados por município, e pode ser explicado pelo fato de Brasília possuir alta concentração de renda e movimentação de grandes quantias de dinheiro, atraindo assim os fraudadores.

Além disso, acreditamos que a presença de municípios do Entorno pode ser explicada pelo baixo PIB per capita e altos índices de criminalidade da região, que podem contribuir para um aumento no número de pessoas envolvidas em casos de fraudes.

Com base nessas informações obtidas no estudo, seria possível para a instituição financeira planejar investimentos para combater as fraudes na região do Distrito Federal e Entorno, utilizando sistemas que dificultam a ação dos fraudadores, como identificações biométricas (retina ou impressão digital), e diminuindo o grande prejuízo causado pelas fraudes.

5.2 TRABALHOS FUTUROS

Uma possibilidade de estudo futuro seria a aplicação do algoritmo *Scan* Circular para detecção e identificação de conglomerados de fraudes bancárias em cartão de débito para as demais regiões, possibilitando um combate mais efetivo por parte da instituição financeira. Podem ser estudadas também outras formas de fraudes e também em outros produtos, como em cartão de crédito.

Além disso, pode-se realizar um estudo mais aprofundado na macro implementada para realizar melhorias como: otimizar o processo de simulação; possibilitar a detecção e inferência de conglomerados secundários; integrar o procedimento GMAPS para a confecção automática dos mapas.

6. BIBLIOGRAFIA

- ASSUNÇÃO, Renato. **Estatística Espacial com Aplicações em Epidemiologia, Economia e Sociologia**. São Carlos, Associação Brasileira de Estatística, 2001. Disponível em: <http://www.dpi.inpe.br/~miguel/cursoemilioribas/MaterialRef/Livro_Renato.pdf>. Acesso em: 01 setembro 2011.
- BEATO FILHO, Cláudio Chaves. et al. **Conglomerados de homicídios e o tráfico de drogas em Belo Horizonte, Minas Gerais, Brasil, de 1995 a 1999.** Caderno de Saúde Pública, Rio de Janeiro, vol.17, no.5, p. 1163-1171, Outubro 2001.
- BESAG, Julian; NEWELL, James. **The Detection of Clusters in Rare Diseases**. Journal of the Royal Statistical Society Series A (1991) Volume: 154, Issue: 1, Pages: 143-155. Disponível em: http://www.mendeley.com/research/the-detection-of-clusters-in-rare-diseases/. Acesso em: 20 de junho de 2011
- CARVALHO, M.S; DRUCK, S. **Análise Espacial de Dados Geográficos**. Câmara, G.; Monteiro, A.V.M. (eds) Brasília, EMBRAPA, 2004
- Febraban. **Federação brasileira de bancos**. Disponível em: <http://www.febraban.org.br/>. Acesso em: 10 junho 2011.
- FIGUEIREDO, Ronaldo Lage. **Detecção de clusters usando a Estatística Scan Espacial Circular em conjuntos seletivos e um fator de penalização:** a ocupação circular**.** 2010. 66 f. Dissertação (Mestrado em Estatística) - Curso de Pós-graduação em Estatística, Universidade Federal de Minas Gerais, Minas Gerais, 2010.
- GOODCHILD, Michael F.. **Pre-Conference Seminar SPATIAL ANALYSIS and GIS,** In: **ESRI USER CONFERENCE 2001,** Califórnia. Disponível em: http://www.csiss.org/learning_resources/content/good_sa/ . Acesso em: 01 setembro 2011.

- GOODCHILD, Michael F.. **SPATIAL ANALYSIS and GIS.** In: ESRI USER CONFERENCE . Pre-Conference Seminar , 2001, Santa Barbara CA. Disponível em: <http://www.csiss.org/learning_resources/content/good_sa/>. Acesso em: 20 de junho de 2011.
- Horus. **Soluçoes em prevenção a fraudes**. Disponível em: < http://www.horusprevention.com.br/2009v1/index.php >. Acesso em: 10 junho 2011.
- KULLDORFF, Martin. **A spatial scan statistic, Communications in Statistics – Theory and Methods,** v. 26, n.6, p.1481 - 1496. Disponível em: <http://dx.doi.org/10.1080/03610929708831995>. Acesso em: 08 abril 2011.
- LI, Xiao-Zhou. et al. **A spatial scan statistic for multiple clusters.** Mathematical Biosciences (July 2011). Disponível em: <http://www.citeulike.org/article/9691837>. Acesso em: 20 de junho de 2011.
- OPENSHAW, Stan. **Mark 1 Geographical Analysis Machine for the automated analysis of point data sets.** International Journal of Geographical Information Systems, Vol. 1, No. 4. (1987), pp. 335-358. Disponível em: <http://www.citeulike.org/user/ianturton/article/5207314>. Acesso em: 20 de junho de 2011
- SAS. Support SAS. **Online SAS support documentation**. Disponível em: <http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_91/iml_ug_7306.pdf >. Acesso em: 08 abril 2011.
- ZHANG, Zhenkui; ASSUNÇÃO, Renato; KULLDORFF, Martin. **Spatial Scan Statistics Adjusted for Multiple Clusters.** Journal of Probability and Statistics, vol. 2010, Article ID 642379. Disponível em: <http://www.hindawi.com/journals/jps/2010/642379/cta/>. Acesso em: 20 de junho de 2011.

APÊNDICE A - MACRO %SCANCIRCULAR

```
%MACRO scancircular (data=, pop=,cases=,x=,y=,nsim=);
proc iml;
      use &data;
      read all;
      pop=&pop;
      cases=&cases;
      x=&x;
      y=&y;
      a= pop || cases || x || y ;
      n=nrow(a);
/*distance matrix */
            dist=j(n,n);
            do i=1 to n;
                  do h=1 to n;
                        dist[i,h]=sqrt((a[i,3]-a[h,3])**2+(a[i,4]-
a[h,4])**2);
                  end;
            end;
/*aux. variables */


p=sum(a[,1]);
c=sum(a[,2]);
pmax=0.5*p;
llrot=.M;
cluster=j(1,n,0);
da=j(n,2);
pz=0;
cz=0;

population_total=p;
population_max=pmax;
total_cases=c;

print population_total population_max total_cases;
/*encontrando os candidatos a clusters */
do i=1 to n;
      col1=dist[,i];
      col2=1:n;
      da[,1]=col1;
      da[,2]=col2`;
      call sort(da,1);
      cluster=j(1,n,0);
      pz=0;
      cz=0;
            do k=1 to n;
            b=da[k,2];
                  pzl=pz+a[b,1];
                  pz=pzl;
                  if pzl<=pmax then do;
                        cluster[1,da[1,2]]=1;
                        cluster[1,da[k,2]]=1;
```

```
                        czl=cz+a[b,2];
                        cz=czl;
                        mz=(c*pzl)/p;
                              if cz>mz then llr=cz*(log(cz)-
log(mz))+(c-cz)*(log(c-cz)-log(c-mz)); else  llr=0;
                              if llr>llrot & pzl<=pmax then do;
                                    clusterot=cluster;
                                    llrot=llr;
                                    bestcz=czl;
                                    bestpz=pzl;
                                    bestsize=sum(clusterot[1,]);
                                    clustera=loc(clusterot);
                              end;
                  end;
            end;
end;

t=llrot;
popcluster=bestpz;
casclus=bestcz;

clustam=ncol(clustera);
casespe=0;
clusterv=j(n,1,0);
do i=1 to clustam;
      x= a[clustera[1,i],1]/p;
      casespe= casespe+(x*c);
      clusterv[clustera[1,i],1]=1;
end;

cluster_population=     popcluster;
cluster_size=clustam;
expected_cases=casespe;
observed_cases=casclus;
print t ' ' cluster_population ' ' cluster_size;
print expected_cases ' ' observed_cases;


create t from t;
append from t;

a=a||clusterv;
create dados2 from a;
append from a;

/*Simulation*/
do i=1 to n;
      a[i,2]=0;
      a[i,5]=a[i,1]/p;
      end;

do i=2 to n;
      a[i,5]=a[i,5]+a[i-1,5];
      end;

nsim=&nsim;
llrvetor=j(nsim,1);

do yy=1 to nsim;
```

```
      do i=1 to n;
            a[i,2]=0;
      end;

do k=1 to c;
      x=rand('uniform');
      i=1;
      do while (a[i,5]<x);
            i=i+1;
      end;
      a[i,2]=a[i,2]+1;
end;

llrot=.M;

cluster=j(1,n,0);
da=j(n,2);

pz=0;

cz=0;

/*empirical llr calculation */
do i=1 to n;
      col1=dist[,i];
      col2=1:n;
      da[,1]=col1;
      da[,2]=col2`;
      call sort(da,1);
      cluster=j(1,n,0);
      pz=0;
      cz=0;
            do k=1 to n;
            b=da[k,2];
                  pzl=pz+a[b,1];
                  pz=pzl;
                  if pzl<=pmax then do;
                        cluster[1,da[1,2]]=1;
                        cluster[1,da[k,2]]=1;
                        czl=cz+a[b,2];
                        cz=czl;
                        mz=(c*pzl)/p;
                              if cz>mz then llr=cz*(log(cz)-
log(mz))+(c-cz)*(log(c-cz)-log(c-mz)); else  llr=0;
                              if llr>llrot & pzl<=pmax then do;
                                    clusterot=cluster;
                                    llrot=llr;
                                    bestcz=czl;
                                    bestpz=pzl;
                                    bestsize=sum(clusterot[1,]);
                                    clustera=loc(clusterot);
                              end;
                  end;
            end;
end;

llrvetor[yy,1]=llrot;

end;
```

```
create llrsim from llrvetor;
append from llrvetor;


quit;

SYMBOL1 V=star C=black I=none;
SYMBOL2 V=circle C=red   I=none;

data dados2;
set dados2;
rename COL4= Y;
rename COL3=X;
run;


legend label=none value=('No Cluster' 'Detected Cluster' );

PROC GPLOT DATA=work.dados2;
     PLOT y*x=COL5/legend=legend1;
RUN;
quit;

data llrsim;
set llrsim;
rename COL1=LLR;
run;

proc univariate data=llrsim noprint;
      histogram;
      output out=simulaest p95=p95;
run;

data t;     set t;
rename COL1=t;
run;

data simulaest;
merge t simulaest ;
if (t>p95) then decision='Reject';
else decision='Not Reject';
run;

proc print data=simulaest noobs;
run;

      %MEND scancircular;
```