PABLO LOURENÇO RIBEIRO DE ALMEIDA

ANÁLISE EXPLORATÓRIA DE DADOS DE ÁREA UTILIZANDO O R

LAVRAS – MG

2018

PABLO LOURENÇO RIBEIRO DE ALMEIDA

ANÁLISE EXPLORATÓRIA DE DADOS DE ÁREA UTILIZANDO O R

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropécuaria, área de concentração em Estatística Espacial, para a obtenção do título de Mestre.

Dr. Renato Ribeiro de Lima

Orientador

LAVRAS – MG

2018

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Almeida, Pablo Lourenço Ribeiro de
ANÁLISE EXPLORATÓRIA DE DADOS DE ÁREA UTILIZANDO O R / Pablo Lourenço Ribeiro de Almeida. – Lavras : UFLA, 2018.
87 p. : il.

Dissertação(mestrado acadêmico)–Universidade Federal de Lavras, 2018.
Orientador: Dr. Renato Ribeiro de Lima.
Bibliografia.

1. Estatística espacial. 2. Autocorrelação espacial. 3. Detecção de clusters. I. Lima, Renato Ribeiro de. II. Título.

PABLO LOURENÇO RIBEIRO DE ALMEIDA

ANÁLISE EXPLORATÓRIA DE DADOS DE ÁREA UTILIZANDO O R

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropécuaria, área de concentração em Estatística Espacial, para a obtenção do título de Mestre.

APROVADA em 19 de Junho de 2018.

Dr. Marcelo Silva de Oliveira	UFLA
Dr. João Domingos Scalon	UFLA
Dr. Denismar Alves Nogueira	UNIFAL

Dr. Renato Ribeiro de Lima
Orientador

LAVRAS – MG
2018

Aos meus pais,
Maria José Lourenço Ribeiro e
Ladimir de Almeida Silva.
A minha irmã,
Poliana Lourenço Ribeiro de Almeida.
A todos meus amigos.
Com todo amor, DEDICO.

AGRADECIMENTOS

Primeiramente gostaria de agradecer a Deus, por ter nos concedido a dádiva da vida e por estar presente em todos os momentos dela. Através dele obtive a coragem e sabedoria necessária para vencer os obstáculos da vida e nunca desistir dos meus sonhos.

A minha mãe, por suprir minhas necessidades, me dando vida, amor, cuidado, conselhos. Me ajudando a levantar nos meus tombos da vida e comemorando comigo minhas conquistas. Mãe, te levarei no meu coração por toda a eternidade.

Ao meu pai, em memória, pois mesmo não estando mais entre nós, me ensinou durante os 13 anos que passamos juntos a ser uma pessoa boa, honesta e altruísta. Levarei esses ensinamentos para toda a vida.

A minha irmã, pelo seu amor, conselhos, incentivos e por sempre apoiar as minhas decisões.

Ao meu orientador Renato Ribeiro de Lima, pela amizade, paciência, incentivos e apoio que foram fundamentais para o desenvolvimento desse trabalho.

A todos meus amigos do Departamento de Estatística da Universidade Federal de Lavras, em especial para Allana Lívia, Carolina Bicalho, Cristian Tiago, Édipo Menezes, Elias Medeiros, Érica Cruz, Henrique José, Lourenço Manuel, Kelly Lima e Victor Ferreira pelos bons momentos que passamos juntos, companheirismo e amizade que levaremos para vida.

Aos meus amigos de casa, Ariel, Bruno, Carlos, Elias, Jeob, Rennan, Rodrigo e Sérgio por toda a amizade e bons momentos que vivemos juntos nesse tempo que morei em Lavras.

A todos os professores e servidores do Departamento de Estatística da Universidade Federal de Lavras, pelo comprometimento em fazer sempre o seu melhor todos os dias para os alunos e instituição. A vocês lhes dou todo o meu respeito e carinho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

O presente trabalho tem como objetivo produzir um material descrevendo detalhadamente a análise de dados de área usando índices de autocorrelação global, autocorrelação local e teste para detecção de *clusters*. A aplicação do estudo foi feita usando um banco de dados dos casos de dengue proveniente da Secretária de Saúde da cidade de Campina Grande-PB. As análises estatísticas foram realizadas com uso do programa estatístico R e do programa SaTScan. Pôde-se observar no estudo que houve presença de autocorrelação espacial global e local para a variável taxa de incidência de dengue nos bairros da cidade. Foi comparado o poder de detecção de áreas de risco do método estatística Scan com o método índice local de autocorrelação espacial (LISA), por meio de mapas temáticos. Dessa comparação, pôde-se concluir que para esse estudo o índice local de autocorrelação espacial (LISA) obteve um menor poder de detecção dos bairros com potencial risco de infecção por dengue.

Palavras-chave: Estatística espacial, autocorrelação espacial, detecção de *clusters.*

ABSTRACT

The present work aims to produce a material describing in detail the spatial data analysis on lattices by considering global autocorrelation indices, local autocorrelation and clusters detection tests. The methods studied in this work were applied in a dataset of classical dengue fever from of the city of Campina Grande, PB, Brazil, obtained from the Municipal Health Office. The statistical analyzes were performed using the statistical softwares R and SaTScan. It was observed that there were global and local spatial autocorrelation for the incidence of the dengue fever. Furthermore, it was compared the power of detecting risk areas by using maps with the information about the local indication of spatial association (LISA) and spatial scan statistics. It was concluded the LISA presented smaller power than the scan statistics to detect the districts with highest risk of dengue fever transmission.

Keywords: Spatial statistics, spatial autocorrelation, cluster detection.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os métodos estatísticos para a análise de dados espaciais desempenham um papel cada vez mais importante no âmbito científico, pois a cada dia que se passa os pesquisadores estão mais cientes da importância de agregar informações espaciais em seus estudos. O ramo da estatística que lida com esse tipo de análise de dados é conhecido como estatística espacial. A teoria da estatística espacial pode ser dividida em três áreas distintas, a depender do tipo de dado a ser analisado, e são intituladas como: padrão de pontos, geoestatística e dados de área.

Neste trabalho o foco será direcionado ao estudo de Dados de Área, em que seus métodos são voltados para análise de dados agregados por regiões (municípios, bairros, setores censitários, etc). Tais métodos foram desenvolvidos para tentar identificar regiões onde a distribuição dos valores da variável em estudo possa apresentar um padrão associado a sua localização espacial. Segundo Câmara et al. (2004), os métodos de análise de Dados de Área foram desenvolvidos para tentar detectar regiões onde a distribuição dos valores pudesse apresentar um padrão específico, ou seja, um padrão aleatório ou não-aleatório associado à sua localização espacial.

Uma característica inerente aos dados agregados por regiões, que é relevante a estudos em análise de dados de área, é a dependência espacial. A dependência espacial pode ser entendida como o fato do valor de uma variável associada a uma localização assemelhar-se ao valor de seus vizinhos. Essa semelhança entre vizinhos é medida por meio de indicadores de autocorrelação espacial que podem ser expressos em escala global ou local. Outra classe de métodos que buscam investigar padrões espaciais em dados georreferenciados são os testes para detecção de *clusters*. Esses testes tem como objetivo investigar a existência de padrões espaciais de uma variável aleatória em uma determinada região de estudo, buscando detectar a presença de conglomerados (*clusters*).

Tanto os métodos de análise de dados de área quanto os métodos para detecção de *clusters* são bastante utilizados em estudos relacionados à investigação epidemiológica. Para este trabalho é proposto fazer uma análise exploratória de dados de casos de dengue da cidade de Campina Grande-PB ocorridos no ano de 2016, utilizando os softwares R (R Core Team, 2017) e SaTScan (KULLDORFF, 2016).

2 OBJETIVOS

2.1 Objetivo Geral

O objetivo principal dessa dissertação é produzir um material didático sobre análise de dados de áreas, que possa servir como base para pesquisadores e estudantes que estejam iniciando seus estudos nessa área de conhecimento.

2.2 Objetivos Específicos

i) Descrever de forma clara e compreensível a teoria de análise de dados de área;

ii) Aplicar a teoria de análise de dados de área aos dados dos casos de dengue na cidade de Campina Grande-PB no ano de 2016;

iii) Instruir com detalhes o passo-a-passo de comandos do Programa R (R Core Team, 2017) e SaTScan (KULLDORFF, 2016) na aplicação do estudo.

3 REVISÃO BIBLIOGRÁFICA

3.1 Estatística espacial

A área da Estatística que integra informações de referências espaciais à análise de dados é conhecida como estatística espacial, cujo seu desenvolvimento passou a ganhar força durante a segunda metade do século XX (DIGGLE, 2013). Para Carvalho et al. (2007), a estatística espacial é o ramo da Estatística que permite analisar a localização espacial de fenômenos. Ao identificar, localizar e visualizar a ocorrência de fenômenos que se materializam no espaço, é possível modelar tais fenômenos por meio da estatística espacial, incorporando ao modelo fatores determinantes como a estrutura de distribuição espacial, podendo assim identificar-se ou não padrões espaciais do fenômeno.

De acordo com Cressie (1993), a estatística espacial pode ser compreendida como um conjunto de métodos, conceitos e técnicas que considera, explicitamente, as coordenadas espaciais, como por exemplo latitude e longitude, que são relevantes para a análise, podendo assim ser modelada como um processo estocástico espacial. Esse processo estocástico espacial pode ser formalmente definido como

$$\left\{Z(s) : s \in D \subset \mathbb{R}^d\right\},$$

em que $Z(s)$ representa uma coleção de variáveis aleatórias indexadas por um conjunto $s \in D$ correspondente à coordenadas no espaço $\mathbb{R}^d$. Esse processo tem como objetivo principal descrever o comportamento de algum fenômeno, em que esse comportamento pode ser caracterizado pelos efeitos de primeira ordem (relacionado ao valor médio do processo) e de segunda ordem (relacionado à dependência espacial).

Com os avanços dos recursos computacionais e o desenvolvimento de novos métodos na estatística espacial, com aplicações em diversas áreas do conhecimento (Geologia, Agronomia, Epidemiologia, etc), teve-se a necessidade de dividir a estatística espacial em áreas distintas. Cressie (1993) dividiu a estatística espacial em três grandes áreas, denominadas de padrões de pontos, geoestatística e dados de área. Segundo Câmara et al. (2004), essas três áreas podem ser definidas da seguinte maneira:

I) Padrões de pontos - são fenômenos expressos por meio de ocorrências identificadas como pontos localizados no espaço, denominados de processos pontuais. Nesses

casos é preciso saber a localização exata da coordenada espacial da ocorrência de cada ponto do estudo no espaço.

II) Geoestatística - são fenômenos caracterizados pela continuidade espacial da variável aleatória de interesse, estimada a partir de um conjunto de amostras de campo, que podem estar regularmente ou irregularmente distribuídas. Normalmente esses tipos de dados são resultantes de levantamentos de recursos naturais, como por exemplo, dados geológicos, ecológicos, fitogeográficos, dentre outros.

III) Dados de área - são dados caracterizados por constituírem valores agregados de uma determinada variável aleatória em uma região de estudo delimitada por sub-regiões (polígonos), como por exemplo, setores censitários, municípios, bairros, etc. Como esses dados são delimitados por sub-regiões, não há a necessidade de conhecer a coordenada geográfica exata, apenas saber se o dado em estudo está localizado dentro de cada sub-região.

Para cada categoria definida anteriormente, existem tipos de dados específicos e métodos estatísticos diferentes a serem aplicados. Para fins deste trabalho, os estudos serão voltados apenas a teoria e os métodos estatísticos de análise de dados de área.

3.2 Análise de dados de área

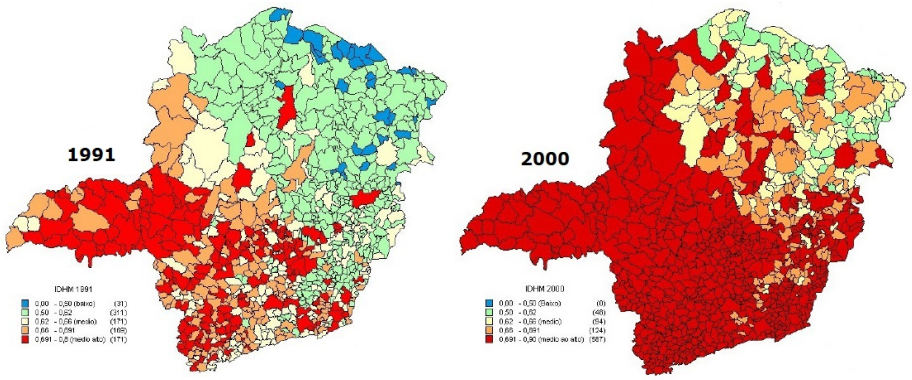
Pode-se dizer que a análise de dados de área é o ramo da estatística espacial em que sua aplicação está associada a determinadas áreas A_i, com $i = 1,2,...,n$, também chamada de polígonos, em que a junção dessas áreas (sub-regiões) A_i formam uma região R.

Segundo Câmara et al. (2004), o modelo de distribuição mais utilizado para dados de área é o modelo de variação espacial discreta. Considerando a existência de um processo estocástico Z_i, $i = 1,2,...,n$, onde Z_i é a realização do processo espacial na área i e n é o total de áreas A_i. O objetivo principal da análise é construir uma aproximação para a distribuição conjunta de variáveis aleatórias $Z = \{Z_1,...,Z_n\}$, estimando sua distribuição, em que Z_i é uma variável aleatória que descreve uma contagem, incidência ou taxa associada à área A_i.

Os dados analisados como dados de área são geralmente oriundos de levantamentos populacionais, tais como: censos, estatísticas de saúde e epidemiologia, estatísticas de criminalidade, taxas de mortalidade, cadastramento de imóveis, etc. Como exemplo de

dados agregados por área, podemos citar Romero (2006), que realizou um estudo sobre análise espacial da pobreza dos municípios do estado de Minas Gerais 1991 a 2000, utilizando como base o Índice de desenvolvimento Humano Municipal (IDHM), como mostra a Figura 3.1.

Figura 3.1 – Distribuição do Índice de Desenvolvimento Humano Municipal (IDHM) do estado de Minas Gerais nos anos 1991 e 2000.



Fonte: Romero (2006).

Na Figura 3.1, pode-se observar uma análise exploratória de dados espaciais da variável Índice de Desenvolvimento Humano Municipal (IDHM) do estado de Minas Gerais. Na análise é possível visualizar um alto crescimento do Índice de Desenvolvimento Humano Municipal (IDHM) entre os anos de 1991 e 2000, principalmente na região sul do estado.

3.2.1 Análise exploratória de dados espaciais

O processo de modelagem espacial é geralmente precedido de uma fase, chamada de análise exploratória de dados espaciais. Esse tipo de análise busca identificar determinados padrões de associação espacial por meio da apresentação visual dos dados em formas de gráficos e mapas, podendo assim identificar padrões de dependência espacial, como também a presença de *outliers* e agrupamentos no fenômeno em estudo. Dessa forma, a análise exploratória de dados espaciais pode ser compreendida como técnicas para exploração de dados espaciais, resumindo suas propriedades, detectando padrões espaciais, levando à formulação de hipóteses que se referem à distribuição espacial dos dados e identificando casos ou subconjuntos dos casos que são incomuns às áreas de estudo (HAINING, 2003). Para Anselin (1995), a análise exploratória de dados espaciais é composta por uma coleção de técnicas que serve para descrever e visualizar distribuições

espaciais, identificando situações atípicas, descobrindo padrões de associação espacial e formação de *clusters*. Segundo Almeida (2012), a análise exploratória de dados de área trata diretamente dos efeitos decorrentes da dependência e heterogeneidade espacial, buscando identificar determinados padrões de associação espacial agindo sobre um grupo de variáveis sujeita a uma certa matriz de pesos espaciais W. Essa matriz W de pesos espaciais é conhecida na análise de dados de área como a matriz de proximidade espacial ou também como a matriz de vizinhança W. Essa matriz é de suma importância, pois é a partir dela que pode-se calcular os índices de autocorrelação global e local das variáveis em estudo.

3.2.2 Matriz de proximidade espacial

A matriz de proximidade espacial é uma ferramenta básica e útil por realizar uma espécie de ponderação da influência que as regiões exercem entre si. Ela pode ser definida como sendo uma matriz quadrada de dimensão n por n, com pesos espaciais W_{ij} que representam o grau de interação entre as regiões, segundo algum critério de proximidade, mostrando a influência da área A_j sobre a área A_i (ALMEIDA, 2012).

O tipo de conexão expresso na matriz de vizinhança W pode ser classificado de acordo com alguns critérios geográficos. Segundo Câmara et al. (2004), dado um conjunto de n áreas $\{A_1,..,A_n\}$, construímos a matriz de vizinhança W, onde cada um dos elementos W_{ij} representa uma medida de proximidade entre A_i e A_j. Essa medida de proximidade pode ser calculada a partir de um dos seguintes critérios:

I) $W_{ij} = 1$, se o centróide de A_i está a uma certa distância de A_j; caso contrário $W_{ij} = 0$;

II) $W_{ij} = 1$, se A_i compartilha um lado comum com A_j , caso contrário $W_{ij} = 0$;

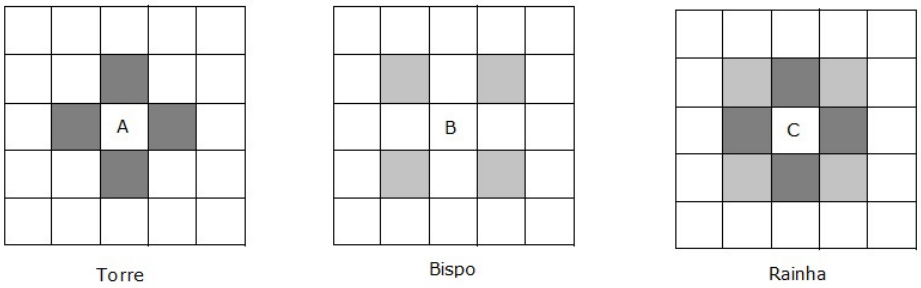
III) $W_{ij} = L_{ij}/L_i$, onde L_{ij} é o comprimento da fronteira entre A_i e A_j e L_i é o perímetro de A_i.

Convencionalmente é presumido que a $W_{ii} = 0$, pois uma área A_i não é considerada como vizinha de si própria, implicando que a matriz de vizinhança W tenha sempre a sua diagonal principal composta por valores nulos.

Observa-se que o critério II) definido acima é construído em consonância com a ideia de vizinhança baseada na contiguidade, ou seja, que duas regiões são vizinhas caso elas partilhem de uma fronteira física em comum. Apesar da aparente simplicidade do

conceito de contiguidade, existem várias possibilidades de convenções entre as fronteiras geográficas. Essas convenções dão alusão ao movimento de peças em um tabuleiro de xadrez. Considerando um *layout* de grade quadrada regular, a convenção de contiguidade é dita torre (*Rook*), quando apenas são considerados os limites comuns diferentes de zero entre as fronteiras. Caso sejam consideradas apenas os vértices entre as fronteiras, a convenção de contiguidade é denominada de bispo (*Bishop*). Se são consideradas ambas, os limites e vértices, a convenção de contiguidade é considerada rainha (*Queen*) (FISCHER; WANG, 2011). Na figura 3.2, podemos observar tipos de convenções em que os vizinhos das regiões A, B e C, respectivamente, estão hachurados.

Figura 3.2 – Diferentes convenções de contiguidade para matriz de vizinhança W.



Fonte: Próprio autor.

Dependendo do critério escolhido, uma área poderá ter quatro (Torre e Bispo) ou oito (Rainha) vizinhos. Esses tipos de convenções normalmente são utilizados em áreas de malha regular, mas mesmo nos casos em que as áreas são de forma irregular, o pesquisador deve tomar uma decisão sobre o tipo de convenção que será utilizada em sua análise.

Um procedimento importante para estudo de análise de dados de área é a normalização da matriz de vizinhança W. Essa importância é justificada quando uma área A_i exerce uma influência diferente sobre a área A_j, ou seja, quando não existe uma uniformidade de vizinhança entre as áreas que formam a região de estudo. Dessa forma, atribui-se diferentes pesos às n áreas que formam essa região, de acordo com a quantidade de vizinhos, de forma que o somatório desse pesos na linha da matriz de proximidade espacial seja igual a 1 (CÂMARA et al., 2004). De acordo com Almeida (2012), a normalização da matriz de proximidade espacial pode ser expressa como:

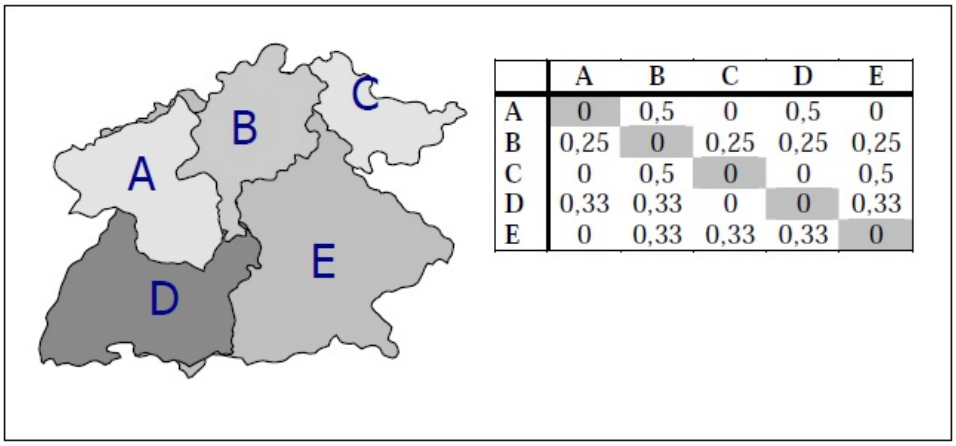
$$W_{ij}^{*} = \frac{W_{ij}}{\sum_{j} W_{ij}}, \tag{3.1}$$

e assim,

$$\sum_j W_{ij}^* = 1. \tag{3.2}$$

As Equações 3.1 e 3.2 informam que para que haja a normalização de uma matriz de vizinhança W, é necessário dividir as células representando os pesos espaciais de cada linha de uma matriz de vizinhança W pela somatória dos pesos das respectivas linhas. Na Figura 3.3, podemos observar um exemplo de uma matriz de proximidade espacial normalizadas pelas linhas.

Figura 3.3 – Matriz de proximidade espacial de primeira ordem, normalizada pelas linhas.



	A	B	C	D	E
A	0	0,5	0	0,5	0
B	0,25	0	0,25	0,25	0,25
C	0	0,5	0	0	0,5
D	0,33	0,33	0	0	0,33
E	0	0,33	0,33	0,33	0

Fonte: Câmara et al. (2004).

A idéia da matriz de proximidade espacial pode ser generalizada para vizinhos de maior ordem (vizinhos dos vizinhos). Com critério análogo ao adotado para a matriz de vizinhança de primeira ordem, pode-se construir as matrizes $W^{(2)}, W^{(3)}, ..., W^{(n)}$, que representam matrizes de 2ª, 3ª,..., n-ésima ordem. Apesar de pouco utilizada, existe ainda a normalização por coluna, em que cada célula da matriz de vizinhança W é dividida pelo total da coluna, e a normalização é feita pelo autovalor da matriz, em que o valor de cada célula é dividido pelo autovalor da matriz de vizinhança W (RIBEIRO; ALMEIDA, 2011).

3.2.3 Média móvel espacial

Uma forma simples de explorar a variação de tendência espacial dos dados é calcular a média dos valores vizinhos,ou seja, a média móvel espacial. Esse método tende a reduzir a variabilidade espacial, pois esse procedimento tende a produzir uma superfície com menor flutuação que os dados originais. Segundo Silva (2010), a média móvel espacial

segue o modelo

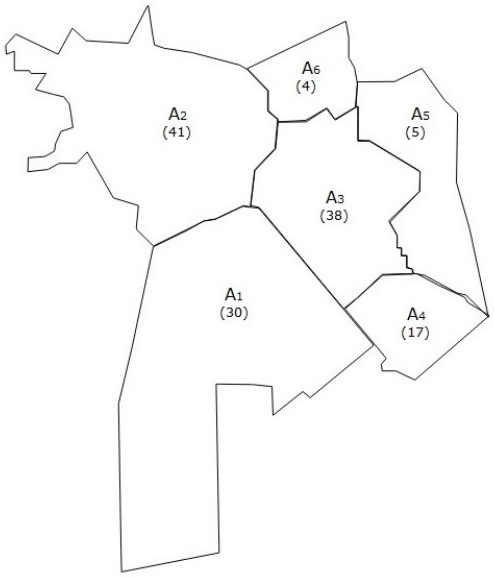
$$Z_i = \mu_i + \varepsilon' + \varepsilon \quad , \tag{3.3}$$

em que, Z_i é o valor da variável aleatória na área i, μ_i é o valor médio Z_i inerente a área i, ε' é a componente espacial estocástica de Z_i, com $E[\varepsilon'] = 0$ e ε é o resíduo, ou seja,variável aleatória i.i.d, com $E[\varepsilon] = 0$ e $VAR[\varepsilon] = \sigma^2$. De acordo com Bailey e Gatrell (1995) a média móvel espacial é definida como

$$\widehat{\mu}_i = \sum_{j=1}^{n} W_{ij}^{*} z_j \tag{3.4}$$

em que $\widehat{\mu}_i$ é o estimador da média móvel na área i, n é o número de áreas (polígonos), W_{ij}^{*} são os elementos da matriz de proximidade espacial normalizada nas linhas e z_j é o valor da variável aleatória na área j. Para que se possa ficar mais claro, vamos fazer um exemplo teórico utilizando os dados dos casos de dengue do município de Campina Grande - PB em 2016. Dos 51 bairros do município, vamos selecionar 6 bairros da zona leste da cidade para o exemplo. São eles: Catolé (A_1), Centro (A_2), José Pinheiro (A_3), Mirante (A_4), Monte Castelo (A_5) e Santo Antônio (A_6), como mostra a Figura 3.4.

Figura 3.4 – Número de casos de dengue nos bairros selecionados da zona leste do município de Campina Grande - PB no ano de 2016.



Fonte: Próprio autor.

Aplicando os valores na equação 3.4, temos que

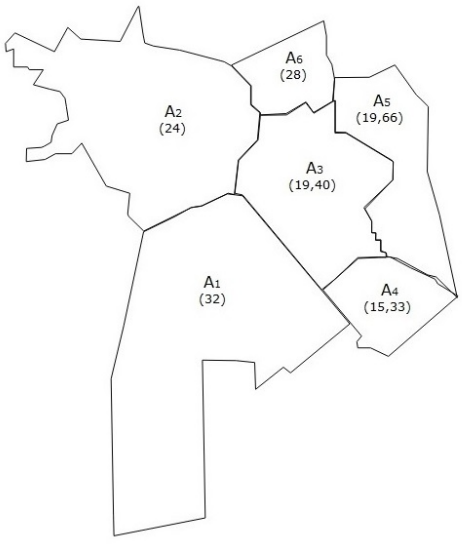
$$\begin{bmatrix} \widehat{\mu}_{A_1} \\ \widehat{\mu}_{A_2} \\ \widehat{\mu}_{A_3} \\ \widehat{\mu}_{A_4} \\ \widehat{\mu}_{A_5} \\ \widehat{\mu}_{A_6} \end{bmatrix} = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 0 & 1/3 \\ 1/5 & 1/5 & 0 & 1/5 & 1/5 & 1/5 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 30 \\ 41 \\ 38 \\ 17 \\ 5 \\ 4 \end{bmatrix}$$

fazendo a multiplicação entre as matrizes, vamos obter as seguintes médias moveis

$$\begin{aligned}
\widehat{\mu}_{A_1} &= (0\text{x}30) + (1/3\text{x}41) + (1/3\text{x}38) + (1/3\text{x}17) + (0\text{x}5) + (0\text{x}4) = 32; \\
\widehat{\mu}_{A_2} &= (1/3\text{x}30) + (0\text{x}41) + (1/3\text{x}38) + (0\text{x}17) + (0\text{x}5) + (1/3\text{x}4) = 24; \\
\widehat{\mu}_{A_3} &= (1/5\text{x}30) + (1/5\text{x}41) + (0\text{x}38) + (1/5\text{x}17) + (1/5\text{x}5) + (1/5\text{x}4) = 19,40; \\
\widehat{\mu}_{A_4} &= (1/3\text{x}30) + (0\text{x}41) + (1/3\text{x}38) + (0\text{x}17) + (1/3\text{x}5) + (0\text{x}4) = 15,33; \\
\widehat{\mu}_{A_5} &= (0\text{x}30) + (0\text{x}41) + (1/3\text{x}38) + (1/3\text{x}17) + (0\text{x}5) + (1/3\text{x}4) = 19,66; \\
\widehat{\mu}_{A_6} &= (0\text{x}30) + (1/3\text{x}41) + (1/3\text{x}38) + (0\text{x}17) + (1/3\text{x}5) + (0\text{x}4) = 28.
\end{aligned}$$

Na Figura 3.5, podemos observar que após a aplicação da média móvel espacial, os valores médios das áreas ficaram mais homogêneos em relação aos valores originais das áreas, causando assim um efeito de suavização entre as áreas do estudo.

Figura 3.5 – Aplicação do método de média móvel espacial nos bairros selecionados da zona leste do município de Campina Grande - PB no ano de 2016.



Fonte: Próprio autor.

3.2.4 Autocorrelação espacial

Um aspecto fundamental do padrão espacial é caracterizado pela avaliação da dependência espacial, o que mostra como os valores estão correlacionados no espaço. A estrutura de dependência entre os valores observados nas várias áreas do fenômeno em estudo é analisada pela função de autocorrelação espacial (CÂMARA et al., 2004).

Embora, de maneira geral, pareça ser fácil definir as propriedades de autocorrelação espacial, é evidente que existem várias formas nas quais os dados podem ser organizados. Portanto, é preciso que se tenha uma definição formal no que se entende como autocorrelação espacial (UPTON; FINGLETON, 1985). Uma definição mais rigorosa foi apresentada por Haining (1980), em que se exige uma independência completa para as variáveis quantitativas. Essa definição pode ser descrita como

$$P(X_1 < x_1, ..., X_n < x_n) = \prod_{i=1}^{n} P(X_i < x_i), \tag{3.5}$$

em que n é o número de áreas na região de estudo. Se as variáveis são categóricas, então uma definição apropriada de independência seria

$$P(X_i = a, X_j = b) = P(X_i = a)P(X_j = b), \tag{3.6}$$

em que a e b são duas categorias possíveis de variáveis, com $i \neq j$.

Uma dificuldade que ocorre nas Equações 3.5 e 3.6 é que as quantidades $P(X_i < x_i)$ e $P(X_i = a)$ requerem estimativas e para isso precisa-se conhecer a forma da distribuição de X. Dessa forma, não basta simplesmente assumir que, por exemplo, a distribuição seja $N(0, \sigma^2)$ com base em evidências fracas, embora isso possa simplificar o trabalho de detecção de presença de autocorreção espacial (HAINING, 1980).

3.2.5 Autocorrelação global

Medidas globais de autocorrelação espacial comparam o conjunto de valor (observação) de similaridade M_{ij} com o conjunto de similaridade espacial W_{ij} , combinando-os em um único índice de um produto cruzado (FISCHER; WANG, 2011), ou seja,

$$\sum_{i}^{n} \sum_{j}^{n} W_{ij} M_{ij}. \tag{3.7}$$

Em outras palavras, o total é obtido aplicando o somatório do produto de cada célula da matriz W com sua entrada correspondente na matriz M. Na literatura são encontrados uma gama de coeficientes que medem a autocorrelação espacial global. Dentre esses, o índice I de Moran se destaca pela sua constante aplicação e por sua facilidade de interpretação. Alguns autores, como Oden (1995) e Assunção e Reis (1999), propuseram uma forma ajustada do índice de I de Moran, que levam em consideração o efeito variado do tamanho da população em cada área da região de estudo. Outros índices globais abordados neste trabalho são a Estatística c de Geary e Estatística G de Getis-Ord.

3.2.5.1 Índice I de Moran global

Moran (1948) propôs um coeficiente de autocorrelação espacial usando a medida de autocovariância na forma de produto cruzado. Dessa forma, surgia o primeiro coeficiente de autocorrelação espacial, denominado de I de Moran. Para WALLER e GOTWAY (2004), na análise de dados de área, o grau de similaridade ou de dependência espacial é avaliada utilizando-se a autocorrelação espacial que pode ser medida por meio do índice de Moran. Portanto, o índice I de Moran global pode ser definido como

$$I = \frac{n}{W_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}^{*}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \tag{3.8}$$

com o fator de normalização

$$W_0 = \sum_{i=1}^{n} \sum_{j=i}^{n} W_{ij}^{*} \quad , \tag{3.9}$$

em que n é o número de áreas, z_i o valor do atributo considerado na área i, $\bar{z}$ é o valor médio do atributo na região de estudo e W_{ij}^{*} os elementos da matriz normalizada de proximidade espacial. Diferente dos demais coeficientes de correlação, como por exemplo os coeficientes de correlação de Pearson e Spearman, que variam entre o intervalo de -1 a 1, o índice de Moran pode variar em qualquer intervalo dos números reais. Porém, se mantém, na maioria dos casos, no intervalo de -1 a 1 (WALLER; GOTWAY, 2004). Quanto mais próximo de 1 o valor da estatística for, maior a semelhança entre vizinhos. Se o valor aproxima-se de zero, indica inexistência de correlação. Se o valor aproxima-se de -1, indica menor semelhança entre vizinhos.

Assumindo que os z_i são observações sobre variáveis aleatórias Z cuja a distribuição é normal, então o índice I de Moran tem distribuição, aproximadamente normal, com os momentos (FISCHER; WANG, 2011)

$$E[I] = -\frac{1}{(n-1)} \quad \text{e} \tag{3.10}$$

$$VAR[I] = \frac{n^2(n-1)W_1 - n(n-1)W_2 - 2W_0^2}{(n+1)(n-1)^2 W_0^2}, \tag{3.11}$$

com W_0 definido na Equação 3.9 e

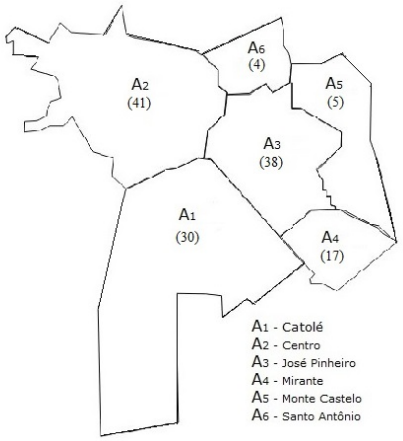
$$W_1 = \sum_{i=1}^{n}\sum_{j\neq i}^{n}(W_{ij} + W_{ji}) \quad \text{e}$$

$$W_2 = \sum_{k=1}^{n}\left(\sum_{j=1}^{n} W_{kj} + \sum_{i=1}^{n} W_{ik}\right)^2.$$

Dessa forma, pode-se testar os valores observados do índice I de Moran versus os percentuais da distribuição Normal. Se um valor excede o valor esperado $E[I]$, há indícios de uma autocorrelação positiva, caso seja inferior, têm-se indícios de uma autocorrelação negativa (FOTHERINGHAM; ROGERSON, 2008).

Para um melhor entendimento, vamos utilizar um exemplo prático com os dados reais dos casos de dengue de seis bairros da zona leste do município de Campina Grande-PB. Os bairros escolhidos para esse exemplo foram: Catolé, Centro, José Pinheiro, Mirante, Monte Castelo e Santo Antônio, como mostra a Figura 3.6.

Figura 3.6 – Bairros selecionados para exemplo prático.



Fonte: Próprio autor.

Inicialmente calculam-se algumas estatísticas básicas relacionadas as áreas do exemplo, ou seja,

- Média:

$$\bar{z} = \frac{30+41+38+17+5+4}{6} = 22,5 \ ;$$

- Variância:

$$\sigma^2 = \frac{\sum_{i=1}^{6}(z_i - \bar{z})^2}{6} = \frac{(30-22,5)^2 + ... + (4-22,5)^2}{6} = 263,5 \ ;$$

Estamos utilizando a variância populacional pelo fato de estarmos considerando que os bairros selecionados para o exemplo prático são o universo que desejamos analisar.

- Desvio padrão:

$$\sigma = \sqrt{\sigma^2} = \sqrt{263,5} = 16,23 \ .$$

Para se calcular o índice I de Moran é necessário construir a matriz de vizinhança W. Dessa forma, iremos construir uma matriz W binária utilizando o critério rainha ($Queen$), logo

$$W = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \end{bmatrix} .$$

A matriz de vizinhança W pode ser normalizada em linhas para que se possa ter uma ponderação nos pesos de cada área em relação aos seus vizinhos, dessa forma a matriz W normalizada será

$$W^* = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 0 & 1/3 \\ 1/5 & 1/5 & 0 & 1/5 & 1/5 & 1/5 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 \end{bmatrix}.$$

Os dados dos casos de dengue podem ser padronizados da seguinte maneira

$$Z_i = \frac{(z_i - \bar{z})}{\sigma}.$$

Logo,

$$\begin{aligned} Z_{A_1} &= \frac{(30-22,5)}{16,23} = 0,4621; \\ Z_{A_2} &= \frac{(41-22,5)}{16,23} = 1,1398; \\ Z_{A_3} &= \frac{(38-22,5)}{16,23} = 0,9550; \\ Z_{A_4} &= \frac{(17-22,5)}{16,23} = -0,3388; \\ Z_{A_5} &= \frac{(5-22,5)}{16,23} = -1,0782; \\ Z_{A_6} &= \frac{(4-22,5)}{16,23} = -1,1398. \end{aligned}$$

Quando normalizamos a matriz de vizinhança W, a expressão do índice I de Moran é simplificada. Isso acontece, pois, o somatório das linhas da matriz de vizinhança W normalizada sempre será igual ao número de áreas n, com isso, o primeiro termo n/W_0 do índice I de Moran sempre será um valor unitário. Logo, a Equação 3.8 pode ser reescrita da seguinte forma

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

Para encontrar o valor do índice I de Moran, vamos resolver a equação por etapas, substituindo os valores do exemplo na forma padronizada do índice.

1ª etapa: Achar a matriz Z_iZ_j,

$$Z_iZ_j = \begin{bmatrix} 0,4621 \\ 1,1398 \\ 0,9550 \\ -0,3388 \\ -1,0782 \\ -1,1398 \end{bmatrix} \begin{bmatrix} 0,4621 & 1,1398 & 0,9550 & -0,3388 & -1,0782 & -1,1398 \end{bmatrix}$$

$$Z_iZ_j = \begin{bmatrix} 0,2135 & 0,5267 & 0,4413 & -0,1565 & -0,4982 & -,5267 \\ 0,5265 & 1,2991 & 1,0885 & -0,3861 & -1,2289 & -1,2991 \\ 0,4413 & 1,0885 & 0,9120 & -0,3235 & -1,0296 & -1,0885 \\ -0,1565 & -0,3861 & -0,3235 & 0,1147 & 0,3652 & 0,3861 \\ -0,4982 & -1,2289 & -1,0296 & 0,3652 & 1,1625 & 1,2289 \\ -0,5267 & -1,2991 & -1,0885 & 0,3861 & 1,2289 & 1,2991 \end{bmatrix};$$

2ª etapa: Achar a matriz $M_{ij} = W_{ij}Z_iZ_j$,

$$M_{ij} = \begin{bmatrix} 0,0000 & 0,1755 & 0,0882 & -0,0521 & 0,0000 & 0,0000 \\ 0,1755 & 0,0000 & 0,2177 & 0,0000 & 0,0000 & -0,4330 \\ 0,1471 & 0,3628 & 0,0000 & -0,1078 & -0,3432 & -0,3628 \\ -0.0521 & 0,0000 & -0,0647 & 0,0000 & 0,1217 & 0,0000 \\ 0,0000 & 0,0000 & -0,2059 & 0,1217 & 0,0000 & 0,4096 \\ 0,0000 & -0,4330 & -0,2177 & 0,0000 & 0,4096 & 0,0000 \end{bmatrix};$$

3ª etapa: Calcular o somatório da matriz M_{ij} e da matriz Z_i^2,

$$\begin{aligned} \sum_{i=1}^{6}\sum_{j=1}^{6} M_{ij} &= -0,0428 \\ \sum_{i=1}^{6} Z_i^2 &= 5,0011; \end{aligned}$$

4ª etapa: Calcular o valor do índice I de Moran,

$$I = \frac{\sum_{i=1}^{6}\sum_{j=1}^{6} M_{ij}}{\sum_{i=1}^{6} Z_i^2} = \frac{-0,0428}{5,0011} = -0,0085.$$

Observamos que para esse exemplo, o valor do índice I de Moran está bem próximo de zero, indicando uma evidência de independência espacial entre as áreas do estudo.

3.2.5.2 Empirical Bayes Index

Assunção e Reis (1999) analisaram o efeito do uso de taxas de prevalência com base em populações de diferentes tamanhos comparando o poder do teste de autocorrelação espacial do índice I de Moran com um índice proposto pelos próprios autores, chamado de Empirical Bayes Index (EBI). Segundo os autores, o índice EBI é mais poderoso que o índice I de Moran quando aplicado diretamente a taxas. Sua estatística é calculada por meio do teste de permutação aleatória. Para cada mapa permutado, é calculado o valor da EBI. O valor-p é obtido pela proporção de vezes que os EBI permutados excedem o EBI observado calculado a partir do mapa real. O Empirical Bayes Index pode ser definido como

$$EBI = \frac{n}{\sum W_{ij}^*} \frac{\sum W_{ij}^*(p_i - \bar{p})(p_j - \bar{p})}{\sum (p_i - \bar{p})^2}$$

em que n é o número de áreas, W_{ij}^* os elementos da matriz normalizada de proximidade espacial e $p_i = m_i/t_i$ é a taxa de prevalência com m_i sendo o número de casos na área i e t_i o total da população de risco na área i.

O valor esperado e a variância de estatística EBI sob hipótese nula podem ser estimados da seguinte forma. Seja $\theta_1, ..., \theta_n$ uma taxa desconhecida e possivelmente diferente nas áreas. Suponha que o número de eventos observados durante um determinado período tenha uma distribuição de Poisson com média condicional $E(m_i \mid \theta_i) = t_i\theta_i$. A taxa estimada p_i tem uma média condicional $E(p_i \mid \theta_i) = \theta_i$ e variância $VAR(p_i \mid \theta_i) = \theta_i/t_i$. Adotando uma abordagem de mistura, suponha que a taxa θ_i tenha sua expectativa e variância iguais a β e α, respectivamente. Portanto, a expectativa marginal de p_i é β e a

variância marginal é $\alpha+\beta/t_i$, de maneira que apenas as variâncias diferem entre as áreas e aumentam à medida que a população diminui.

3.2.5.3 Estatística *c* de Geary global

Geary (1954) propôs uma outra estatística global de autocorrelação. Ela foi construída com base em uma medida de covariância diferente, ou seja, o quadrado da diferença entre pares dos valores do atributo em estudo. Dessa forma, o índice c de Geary pode ser definido como

$$c = \frac{(n-1)}{2W_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}^{*}(z_i - z_j)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \ , \tag{3.12}$$

em que W_0 é o fator de normalização definido na Equação 3.9, n é o número de áreas, z_i é o valor do atributo considerado na área i, $\bar{z}$ é o valor médio do atributo na região de estudo e W_{ij}^{*} são os elementos da matriz normalizada de proximidade espacial. O valor de seu índice varia aproximadamente entre 0 e 2. Quando a autocorrelação espacial é positiva, os pares de regiões próximas umas das outras tendem a ter valores semelhantes, de modo que a sua estatística tenha um valor próximo de zero. Quando a autocorrelação espacial é negativa, as regiões próximas umas das outras tendem a ter valores bastante diferentes, de modo que sua Estatística se aproxime do outro valor extremo 2. A não existência de autocorrelação espacial resultaria em um valor próximo de 1 (ROGERSON; YAMADA, 2008). Assumindo que os z_i são observações de variáveis aleatórias Z cuja a distribuição é normal, então a Estatística c de Geary tem distribuição aproximadamente normal com os momentos

$$E(c) = 1 \quad \text{e}$$

$$VAR(c) = \left\{ (n-1) \left\{ W_0 + \sum_{i}^{n} \left[\frac{\sum_j^n W_{ij}^{*} \left(\sum_j^n W_{ij}^{*} - 1 \right)}{2} \right] \right\} - \frac{(W_0)^2}{2} \right\} \left\{ (n+1) \left(\frac{W_0}{2} \right)^2 \right\} \ ,$$

em que W_0 é o fator de normalização definido na Equação 3.9.

3.2.5.4 Estatística G de Getis-Ord global

Getis e Ord (1992) também propuseram na literatura uma medida de autocorrelação espacial de natureza global. Diferente dos índices de Moran e Geary, que detectam autocorrelação espacial positiva e negativa, a estatística G de Getis-Ord só detecta autocorrelação espacial positiva. Segundo Getis e Ord (1992), seu índice pode ser definido como

$$G(d) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(d) z_i z_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j} \ , \text{ para } i \neq j \ ,$$

em que n é o número de áreas, z_i o valor do atributo considerado na área i e $W_{ij}(d)$ os elementos da matriz binária de proximidade espacial. A estatística $G(d)$ mede a associação espacial utilizando todos os pares de valores da variável de interesse (z_i, z_j), de modo que as localizações i e j estão dentro de uma distância d entre si. Ela só identifica associação espacial positiva e o valor do seu índice varia aproximadamente entre 0 e 1. Semelhante ao índice I de Moran global, é comum usar a estatística G de Getis-Ord padronizada e, então, assumir normalidade para testar a hipótese nula de não agrupamento. Se existe um padrão dominante de valores altos próximos de outros valores altos, a estatística G é alta. Caso contrário, quando existe uma tendência geral de agrupamento de valores baixos, a estatística G é baixa(ROGERSON; YAMADA, 2008). Considerando sua estatística padronizada, a média e a variância podem ser definidas por

$$Z[G(d)] = \frac{G(d) - E[G(d)]}{\sqrt{VAR[G(d)]}} \ ,$$

$$E[G(d)] = \frac{W}{n(n-1)}, \quad W = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(d) \quad i \neq j,$$

$$E[G^2(d)] = \frac{1}{(n_1^2 - n_2)}\left[B_0 n_2^2 + B_1 n_4 + B_2 n_1^2 n_2 + B_3 n_1 n_3 + B_4 n_1^4\right] \quad \text{e}$$

$$VAR[G(d)] = E[G^2(d)] - [E[G(d)]]^2 \ ,$$

em que

$$\begin{aligned} B_0 &= (n_2 - n_3 - 3)S_1 - nS_2 + 3W^2; \\ B_1 &= -[(n^2 - n)S_1 - 2nS_2 + 6W^2]; \\ B_2 &= -[2nS_1 - (n+3)S_2 + 6W^2]; \\ B_3 &= 4(n-1)S_1 - 2(n+1)S_2 + 8W^2; \\ B_4 &= S_1 - S_2 + W^2; \end{aligned}$$

$$\begin{aligned} S_1 &= \frac{1}{2}\sum_i \sum_j (W_{ij} + W_{ji})^2, \quad j \neq i; \\ S_2 &= \sum_i (W_{i.} + W_{.i}), \quad W_{i.} = \sum_j W_{ij}, \quad j \neq i; \end{aligned}$$

$$n_j = \sum_{i=1} Y_i^j, \quad j = 1,2,3,4; \qquad n^{(r)} = n(n-1)(n-2)(n-r+1).$$

Numa outra versão dessa estatística proposta por Getis e Ord (1995), os autores procuram relaxar algumas condições em busca de torná-la mais aplicável. Nessa nova versão, a estatística G pode ser calculada para valores negativos, nulos ou positivos da variável de interesse. Além disso, essa nova versão permite usar matrizes de ponderação espacial com pesos não binários e assimétricos. Neste trabalho, não estudaremos essa nova versão da estatística G de Getis-Ord.

Existe ainda uma outra medida de autocorrelação espacial, chamada de índice de autocorrelação local, utilizada para estudos em localidades específicas dentro de uma área global, que será descrita na próxima seção.

3.2.6 Autocorrelação local

Os indicadores globais de autocorrelação espacial fornecem um único valor como medida da associação espacial para todo o conjunto de dados, o que é útil apenas na caracterização da região de estudo como um todo (CÂMARA et al., 2004). Muitas vezes estamos interessados em investigar a presença de aglomerados (*clusters)* em determinadas localidades específicas e para isso utilizamos indicadores locais de autocorrelação espacial.

Getis e Ord (1992), pioneiramente propuseram duas famílias de indicadores locais, denotados como indicadores locais G_i e G_i^*. Anselin (1995) sugeriu um indicador capaz de capturar padrões locais de autocorrelação espacial, o chamado indicador LISA (*Local Indicator of Spatial Association*). Esse indicador também apresenta duas estatísticas testes, que são o Índice I de Moran local e a estatística c de Geary local.

Quando lidamos com um grande número de áreas, é muito provável que ocorram diferentes regimes de associação espacial e que apareçam máximos locais de autocorrelação espacial, onde a dependência espacial é ainda mais pronunciada. Dessa forma, os indicadores locais produzem um valor específico para cada área, permitindo assim a identificação de agrupamentos. Neste trabalho, vamos estudar três desses indicadores locais: o Índice I de Moran local, a estatística c de Geary local e a estatística G de Getis-Ord local.

3.2.6.1 Índice I de Moran local

O Índice I de Moran local é usado para determinar a existência de autocorrelação espacial local em torno de uma área especificada i $(i = 1,...,n)$. Dessa forma, Anselin (1995) definiu o índice I de Moran local como sendo

$$I_i = (z_i - \bar{z}) \sum_{j \in J_i}^{n} W_{ij}^*(z_j - \bar{z})^2 \ ,$$

em que z_i é o valor do atributo na área i, W_{ij}^* são os elementos da matriz normalizada de proximidade espacial, J_i denota um conjunto de áreas vizinhas a área i, o somatório das j áreas percorre apenas nas áreas pertencentes a J_i e $\bar{z}$ indica a média das áreas observadas. Segundo Fischer e Wang (2011), os momentos para I_i, sob a hipótese nula, podem ser descritos por

$$E[I_i] = -\frac{W_i}{(n-1)}, \quad W_i = \sum_{j=1}^{n} W_{ij}^*,$$

com variância

$$VAR[I_i] = \frac{1}{(n-1)} W_{i(2)} + \frac{2}{(n-1)(n-2)} W_{i(kh)}(2b_2 - n) - \frac{1}{(n-1)^2} W_i^2 \ ,$$

em que

$$W_{i(2)} = \sum_{j \neq i}^{n} (W_{ij}^{*})^2 \quad ,$$

$$2W_{i(kh)} = \sum_{k \neq i}^{n} \sum_{h \neq i}^{n} W_{ik} W_{ih} \quad ,$$

com $b_2 = m_4 \; m_2^{-2}$, $m_2 = \sum_i (z_i - z)^2 n^{-1}$ como o segundo momento e $m_4 = \sum_i (z_i - z)^4 n^{-1}$ como o quarto momento. Um teste de associação espacial local pode ser baseado nesses momentos, embora a distribuição exata da estatística ainda seja desconhecida (ANSELIN, 1995). A significância estatística do uso do índice de Moran local é computada de forma similar ao caso do índice global. Para cada área, calcula-se o índice local, e depois permuta-se aleatoriamente o valor das demais áreas, até obter uma pseudo-distribuição para a qual possa-se computar os parâmetros de significância (FISCHER; WANG, 2011).

3.2.6.2 Estatística *c* de Geary local

Em Anselin (1995) é apresentada uma outra estatística local baseada na estatística c de Geary global. Usando a mesma notação do Índice I de Moran local, a estatística c de Geary local para as observações (áreas) $i = 1, ..., n$, é definida como

$$c_i = \sum_{j \in J_i}^{n} W_{ij}^{*} (z_i - z_j)^2 \quad ,$$

em que z_i é o valor do atributo na área i, z_j é o valor do atributo na área j, W_{ij}^{*} são os elementos da matriz normalizada de proximidade espacial e J_i denota um conjunto de áreas vizinhas as áreas i. Seu critério de decisão é dado de maneira similar estatística global. Se o valor de c_i for próximo de 1, não existe autocorrelação espacial. Se o valor de c_i for menor que 1, existe evidências de autocorrelação espacial positiva. Caso o valor de c_i for maior que 1, existe evidências de autocorrelação espacial negativa (ALMEIDA, 2012).

3.2.6.3 Estatística G de Getis-Ord local

Getis e Ord (1992) propuseram uma forma de analisar localmente a associação espacial. Diferente do índice I de Moran, que mede a autocorrelação entre valores de atributo em áreas vizinhas, a estatística local de Getis-Ord é um indicador que mede a concentração local de uma variável de atributo distribuída espacialmente (PFEIFFER et al., 2008). Segundo Getis e Ord (1995), a Estatística G de Getis-Ord local pode ser definida como

$$G_i(d) = \frac{\sum_j W_{ij}(d) z_j}{\sum_j z_j}, \text{ com } j \neq i \ , \tag{3.13}$$

em que $W_{ij}(d)$ são os elementos da matriz binária de proximidade espacial, z_j é o valor do atributo considerado na área j e o somatório em j significa que apenas os valores dos vizinhos próximos da área i serão usados no cálculo da estatística para se obter o numerador da Equação 3.13, segundo algum critério de vizinhança dado pela matriz de proximidade espacial. Essa matriz de proximidade espacial é baseada num raio construído em torno de uma área i com base em uma distância de coorte (d).

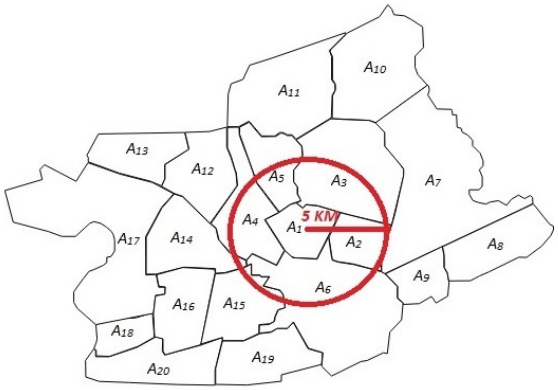
Getis e Ord (1995) também propuseram nesse mesmo artigo uma segunda forma de se calcular a estatística G de Getis-Ord local, que foi definida como

$$G_i^*(d) = \frac{\sum_j W_{ij}(d) z_j}{\sum_j z_j}, \quad \text{para qualquer } j, \tag{3.14}$$

em que a única diferença entre a Equação 3.13 e a Equação 3.14 está no somatório do denominador, que agora está definido para qualquer j, inclusive para $j = i$. Essa leve diferença entre as duas equações podem ser justificadas pelos seguintes fatos: se não incluir a observação sob consideração i, temos $G_i(d)$, caso contrário, tem-se $G_i^*(d)$.

Para um melhor entendimento, vamos exemplificar o cálculo das estatísticas $G_i(d)$ e $G_i^*(d)$, mostrando como se aplicam as Equações 3.13 e 3.14. Suponha uma cidade composta por 20 bairros, em que os valores da variável de interesse z_i são representados em cada bairro por A_i, com $i = 1, ..., 20.$), como mostra a Figura 3.7.

Figura 3.7 – Ilustração de localizações para o calculo das estatísticas $G_i(d)$ e $G_i^*(d)$.



Fonte: Próprio autor.

Para computar o $G_i(d)$ do bairro A_1, em primeiro lugar é especificada uma matriz de proximidade espacial binária, baseada numa distância em torno do bairro A_1, que para esse exemplo, foi definido um raio de 5 Km. Em seguida, foram computadas todos os valores para variável z_i dos bairros situados dentro do raio de 5 Km para o numerador da Equação 3.13, ao passo que o somatório de todos os valores de z_i constam no denominador, exceto o valor z_1. Logo, a estatística $G_i(d)$ é calculada da seguinte maneira

$$G_i(d) = \frac{A_2 + A_3 + ... + A_6}{A_2 + A_3 + ... + A_{20}}$$

Já para a estatística $G_i^*(d)$, inclui-se o valor de Y_A no denominador da Equação 3.14 e é calculada da seguinte maneira

$$G_i^*(d) = \frac{A_2 + A_3 + ... + A_6}{A_1 + A_2 + A_3 + ... + A_{20}}$$

Uma observação importante é que a estatística G de Getis-Ord local só pode ser calculada para valores positivos da variável de interesse X, ou seja, só são captados nessa estatística os padrões espaciais Alto-Alto (AA) e Baixo-Baixo (BB). De acordo com Getis e Ord (1995), a estatística G de Getis-Ord local tem distribuição apropriadamente normal com os momentos

$$E[G_i(d)] = \frac{W_i}{(n-1)}, \quad W_i = \sum_j w_{ij}(d) \quad \text{e}$$

$$VAR[G_i(d)] = \frac{W_i(n-1-W_i)}{(n-1)^2(n-2)} \left[\frac{s(i)}{\overline{Y}(i)}\right]^2.$$

em que

$$\overline{Y}(i) = \frac{\sum_j Y_j}{(n-1)} \quad \text{e} \quad s^2(i) = \frac{\sum_{i=1}^{n} Y_i^2}{(n-1)} - \left[\overline{Y}(i)\right]^2.$$

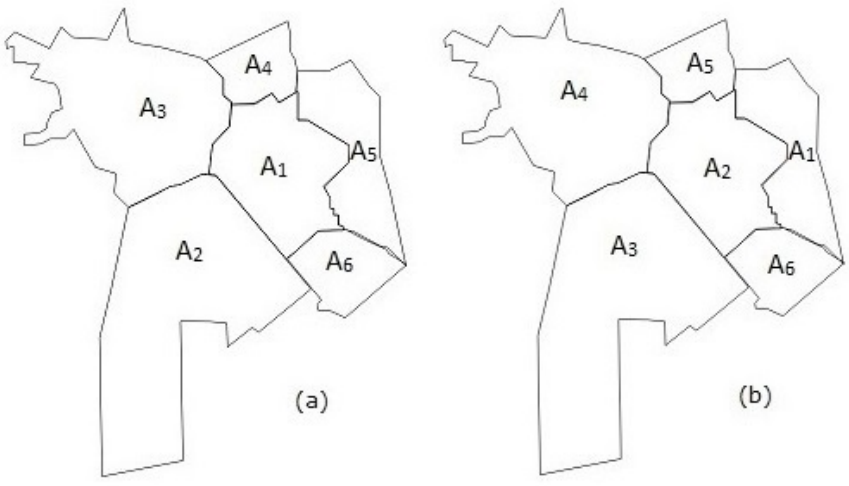
3.2.7 Inferência associada a coeficientes de autocorrelação espacial

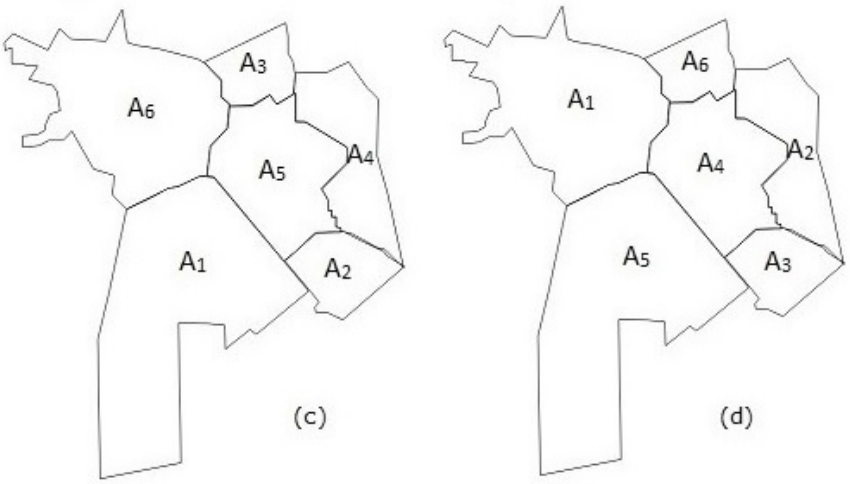
Um dos aspectos mais relevantes para os testes em uma análise espacial é estabelecer sua validade estatística, ou seja, obter uma segurança para valores medidos, a partir de um determinado nível de significância estabelecido. Para calcular os índices de autocorrelação espacial é necessário associar a estes uma distribuição de probabilidade. Para isso, duas abordagens são possíveis no estudo em análise de dados de áreas (WALLER; GOTWAY, 2004).

I) Teste de pseudo-significância (Teste de permutação aleatória);
II) Distribuição aproximada (Hipótese de normalidade).

Para realizar o teste de pseudo-significância, são geradas diferentes permutações dos valores de atributos associados às áreas do estudo, em cada permutação é produzido um novo arranjo espacial, onde os valores são redistribuídos às áreas. No exemplo da figura 3.8, podemos observar como são feitos os arranjos espaciais dessas diferentes permutações.

Figura 3.8 – Exemplo de permutação aleatória.

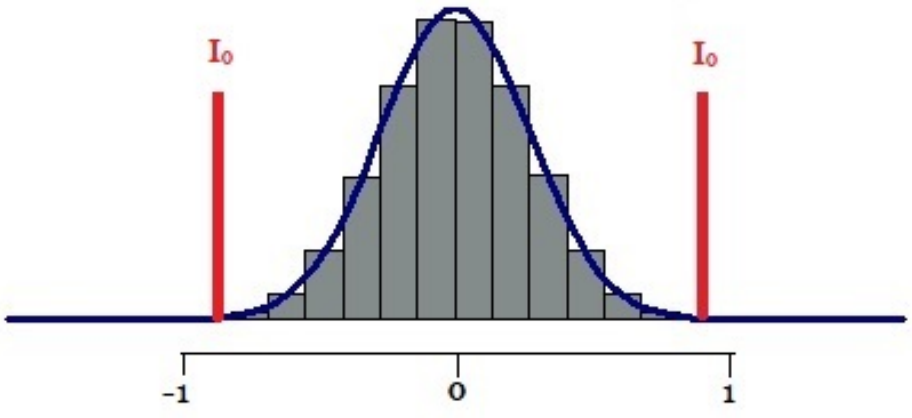




Fonte: Próprio autor.

Na Figura 3.8, cada A_i, com $i = 1, ..., 6$, representa o valor de uma variável aleatória de uma área na região de estudo. A hipótese nula é que as variáveis aleatórias $A_1, ..., A_6$ são permutáveis, ou seja, se não existe autocorrelação espacial então toda possível alocação das variáveis $A_1, ..., A_6$ às áreas do mapa seria igualmente provável. O mapa da Figura 3.8(a), mostra a distribuição original do valor da variável aleatória, enquanto que os demais mapas mostram os valores permutados. Por meio de uma distribuição simulada, podemos exemplificar a validade estatística de um índice de autocorreção espacial através do teste de pseudo-significância, como mostra a Figura 3.9.

Figura 3.9 – Distribuição simulada para o teste de pseudo-significância.



Fonte: Próprio autor.

A interpretação da Figura 3.9 se dá da seguinte forma: se o índice de autocorrelação efetivamente medido, representado na figura pela linha vermelha, corresponder a um

extremo da distribuição simulada, então o índice de autocorrelação possui significância estatística, ou seja, existe uma dependência espacial entre as áreas do estudo.

Teoricamente, no teste de pseudo-significância podem ser produzidas $n!$ combinações diferentes de permutações aleatórias dos valores do atributo de interesse (UPTON; FINGLETON, 1985). No exemplo da Figura 3.8, com um mapa que consta apenas 6 bairros, ou seja, com $6!$ pode-se obter 720 combinações diferentes de permutação para esse mapa. No entanto, essa teoria pode se torna inviável, pois dependendo da quantidade de áreas na região de estudo, o cálculo de todas permutações pode se tornar intratável. Um exemplo dessa situação seria estudar os casos de dengue no estado de Minas Gerais, que é composto 853 municípios. Logo, no teste de pseudo-significância teriam que ser feitas um total de $853!$ combinações diferentes de permutações aleatórias para essa região de estudo, número esse inviável de ser calculado.

Uma maneira encontrada de contornar esse problema foi selecionar aleatoriamente um número menor de possíveis permutações (amostra) dentro da região de estudo e atribuir ao teste um número fixo de permutações a serem executadas. Esse número fixo de permutações é definido a critério do pesquisador, onde geralmente são usadas pelo menos 999 permutações. Nos programas de análise de dados de área é padrão existir uma função onde atribui-se o número de permutações a serem calculadas no estudo.

A segunda abordagem é feita por distribuição aproximada, ou seja, se existe um certo número n de áreas na região de estudo e suas respectivas observações, logo é possível construir uma distribuição de probabilidade. Se esse número n de áreas na região é grande (tende ao infinito), então sua distribuição amostral tende a ser aproximadamente normal. Sua esperança e variância serão atribuídas a depender de cada índice utilizado. Por exemplo, o índice I de Moran terá sua distribuição normal padronizada dada por

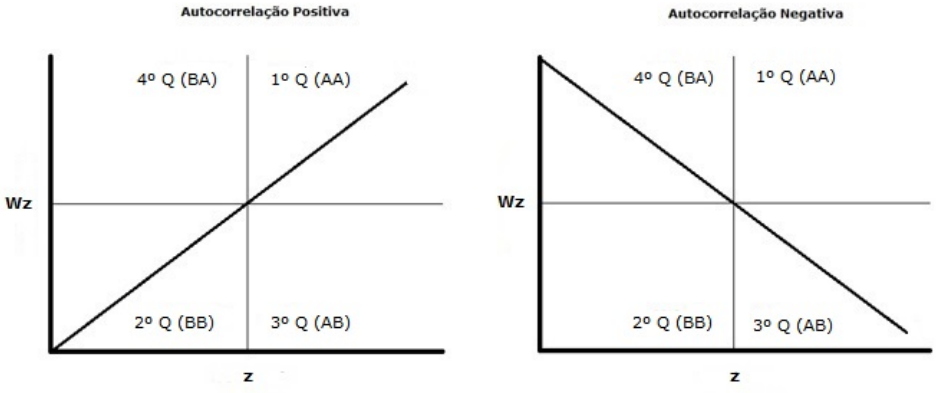
$$Z_I = \frac{I - E[I]}{\sqrt{VAR[I]}} \ , \tag{3.15}$$

com $E[I]$ e $VAR[I]$ definidas, respectivamente, nas Equações 3.10 e 3.11. O valor de Z_I obtido na Equação 3.15, corresponde a um determinado quantil da distribuição normal padronizada, que corresponde a um determinado valor-p. O índice I de Moran será considerado significativamente diferente de zero (hipótese nula) se o valor-p for inferior ao nível nominal de significância α previamente estabelecido.

3.2.8 Diagrama de dispersão de Moran

O diagrama de dispersão de Moran é uma maneira adicional de visualizar a dependência espacial. Construído com base nos valores normalizados, ele permite analisar o comportamento da variabilidade espacial. Segundo Marconato et al. (2017), o diagrama de dispersão mostra a associação espacial entre cada área da região de estudo com os seus vizinhos, construindo um gráfico bidimensional de z (valores normalizados) por wz (média dos vizinhos). O diagrama de dispersão de Moran é dividido em quatro quadrantes: alto-alto (AA), baixo-baixo (BB), alto-baixo (AB) e baixo-alto (BA), como mostra a Figura 3.10.

Figura 3.10 – Diagrama de dispersão de Moran.



Fonte: Próprio autor.

De acordo com Almeida (2012), esses quadrantes podem ser interpretados da seguinte maneira:

I) O primeiro quadrante, Alto-Alto (AA), significa que as regiões exibem valores altos da variável de interesse, rodeados por regiões que apresentam valores também altos;

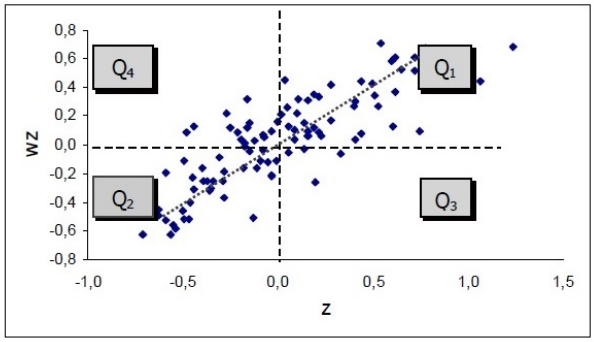
II) O segundo quadrante, Baixo-Baixo (BB), significa que as regiões exibem valores baixos da variável de interesse, rodeados por regiões que apresentam valores também baixos;

III) O terceiro quadrante, Alto-Baixo (AB), significa que as regiões exibem valores altos para variável de interesse, rodeadas por regiões que apresentam valores baixos;

IV) O quarto quadrante, Baixo-Alto (BA), significa que as regiões exibem valores baixos para variável de interesse, rodeadas por regiões que apresentam valores altos.

Segundo Anselin (1995), o diagrama de dispersão de Moran é uma das formas de interpretar a estatística I de Moran e pode ser representada por um coeficiente de regressão que permite visualizar a correlação linear entre z e wz por meio de um gráfico de duas variáveis. No gráfico da Figura 3.10, pode-se observar que quando o I de Moran é positivo a reta de regressão apresenta inclinação ascendente e os dados tendem a estar agrupados no primeiro e segundo quadrante. Quando o I de Moran é negativo, a reta de regressão é descendente e os dados tendem a estar agrupados no terceiro e quarto quadrante. Como exemplo, temos o diagrama de dispersão de Moran para índice de exclusão/inclusão social de São Paulo, censo de 1991 (CÂMARA et al., 2004), como mostra a Figura 3.11.

Figura 3.11 – Diagrama de dispersão de Moran para o índice de exclusão/inclusão social de São Paulo, censo de 1991.



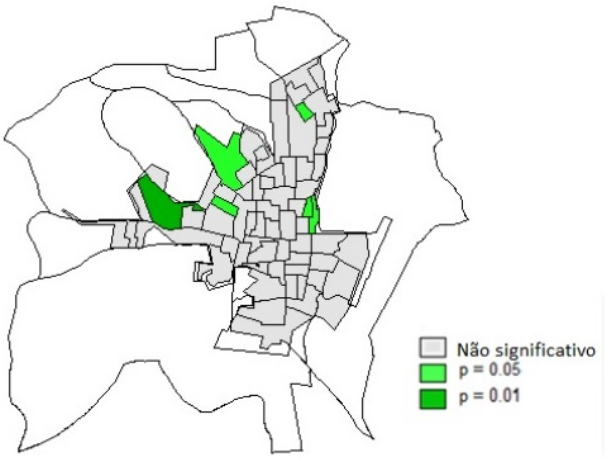
Fonte: Câmara et al. (2004).

Pode-se observar na Figura 3.11 que a nuvem de pontos concentra-se em sua maioria nos quadrantes um e dois, onde a reta de regressão foi traçada nos mesmos quadrantes, indicando dessa forma uma autocorrelação espacial positiva.

3.2.9 Mapa LISA

Uma vez determinada a significância estatística do índice local de Moran, é útil gerar um mapa indicando as regiões que apresentam correlação local significativamente diferente do restante dos dados (CÂMARA et al., 2004). Esse mapa foi denominado por Anselin (1995) como *"LISA MAP"*. Na geração desse mapa, os índices locais I_i são classificados nos seguintes grupos: não significativos, significativos a 5%, 1% e 0,1%. Como exemplo, podemos observar o estudo sobre a análise de mortalidade infantil no município de Alfenas-MG, realizado por Manuel (2011). A Figura 3.12, mostra *"LISA MAP"* para o número de óbitos com menos de 1 ano de idade.

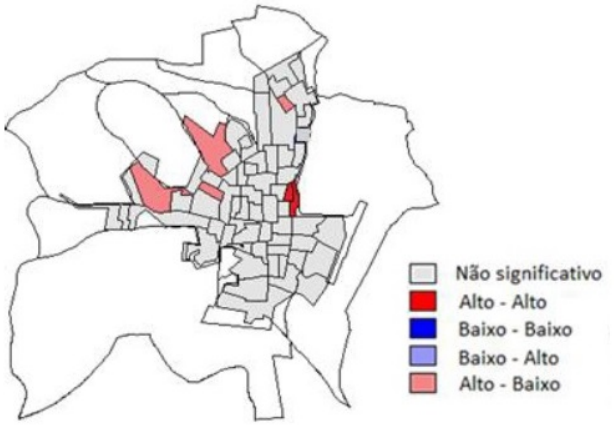
Figura 3.12 – *"LISA MAP"* para o número de óbitos com menos de 1 ano de idade na área urbana do município de Alfenas.



Fonte: Manuel (2011).

No mapa da Figura 3.12, observa-se que as áreas que apresentam coloração verde, possuem um valor do índice de Moran local estatisticamente significativo. Nesse estudo, apresentou-se uma área em verde mais escuro, com uma significância à 1% e cinco outras áreas com o índice significativo à 5%. Não houve áreas significativas à 0,1% e as demais áreas do estudo foram não significativas. Uma outra maneira de gerar o *"LISA MAP"* é quando as áreas para quais os índices de Moran locais foram significativos são classificados em grupos, conforme o diagrama de dispersão de Moran. Na Figura 3.13, podemos observar um exemplo do *"LISA MAP"* classificados de acordo como o diagrama de dispersão de Moran.

Figura 3.13 – *"LISA MAP"* para o número de óbitos com menos de 1 ano de idade na área urbana do município de Alfenas.



Fonte: Manuel (2011).

No mapa da Figura 3.13, pode-se observar que das seis áreas que apresentaram um valor do índice de Moran local estatisticamente significativo, duas correspondem ao primeiro quadrante (AA) e as demais correspondem ao quarto quadrante (AB). Para o caso das áreas pertencentes ao primeiro quadrante (AA), significa que a dependência espacial entre estas áreas é positiva, isto é, os valores da mortalidade infantil nessas áreas são similares entre si. Para o caso das áreas correspondentes ao quarto quadrante (AB), a dependência espacial nessas áreas em relação aos seus vizinhos é negativa, ou seja, os valores da mortalidade infantil nessas áreas tendem a ser dissimilares em áreas vizinhas.

3.3 Teste para detecção de *clusters*

A análise de *clusters* espaciais tem recebido uma atenção considerável em diversas áreas do conhecimento, dentre elas, em estudos epidemiológicos. O objetivo básico em um problema de detecção de *clusters* é determinar automaticamente áreas do espaço onde tenha ocorrido uma mudança não esperada no padrão espacial do processo estocástico observado (LIMA, 2011). Essas mudanças no padrão espacial podem corresponder a uma variedade de fenômenos, dependendo do campo de aplicação, como por exemplo, a detecção de *clusters* dos casos de dengue do município de Campina Grande-PB.

Segundo Rogerson e Yamada (2008), essa classe de testes de detecção de*clusters* realizam os testes locais sob a hipótese nula, ou seja, a ausência de *clusters* na região de estudo. Para Assunção e Costa (2004), essa hipótese de ausência de *clusters* pode ser expressa da seguinte forma. Suponha que uma região de estudo seja dividida em n áreas, sendo associado a cada área o número observado de casos y_i e o número esperado de casos E_i, com $i = 1, ..., n$. Sendo λ especificado como a taxa de ocorrência de casos e N_i como o número total de pessoas em risco na área i, o modelo nulo de aleatoriedade ou ausência de conglomerados pode ser escrito como

$$H_0 : y_i \sim Poisson(\lambda N_i) \tag{3.16}$$

em que $E_i = \lambda N_i$ e o estimador da taxa de ocorrência de casos λ é definido por

$$\widehat{\lambda} = \frac{\sum_i y_i}{\sum_i N_i}.$$

Segundo Lima (2011), os estudos de *clusters* podem ser realizados tanto em análise de processo pontual (localizações de eventos), como em análise de dados de área (agregado de eventos), uma vez que um processo pontual pode ser transformado em dados de área. Existem vários métodos para a detecção de *clusters* espaciais e com diferentes propósitos. Alguns avaliam a existência de um *cluster* global, sem especificar sua localização, enquanto outros, determinam a localização e avaliam a significância estatística do *cluster*. Estes métodos usam técnicas computacionalmente intensivas como permutação aleatória, Monte Carlo, etc. Desses vários métodos para a detecção de *clusters* espaciais encontrados na literatura, o mais usual deles é a estatística Scan.

3.3.1 Estatística Scan

Considere a situação em que uma região geográfica em estudo seja dividida em n áreas. Suponha que para cada uma dessas áreas sejam conhecidas as coordenadas geográficas, suas populações e o número de casos de um fenômeno observável. Pode-se ter como exemplo de fenômeno observável, o número de casos de dengue ou número homicídios de uma cidade. A partir dessas informações, pode-se levantar os seguintes questionamentos: Os casos do fenômeno estão distribuídos aleatoriamente nessas áreas? Existe alguma área da região com valores discrepantes das demais? Partindo desses questionamentos, a estatística Scan testa as seguintes hipóteses:

$$\begin{cases} H_0: & \text{Não há } \textit{cluster} \text{ na região de estudo} \\ H_1: & \text{Há } \textit{cluster} \text{ na região de estudo} \end{cases}$$

Segundo Kulldorff e Nagarwalla (1995), o objetivo da estatística Scan é identificar *clusters* em regiões cuja a ocorrência de um fenômeno é significantemente mais provável dentro de uma área do que fora dela. Atualmente, a estatística Scan é o método mais empregado em abordagens de detecção e inferência de *clusters* espacial, temporal e espaço-temporal (OLIVEIRA, 2017). Essa estatística usa diferentes modelos de probabilidades, a depender da natureza dos dados, sendo os mais usuais os modelos da distribuição Bernoulli e Poisson. A distribuição Bernoulli é usada quando se tem um estudo de caso-controle, já a distribuição Poisson é usada quando existe um fator de risco no estudo. Por motivo do tipo de dados em análise neste trabalho, só será usado a distribuição Poisson. Segundo

Kulldorff (1997), sob hipótese nula, o teste da razão de verossimilhança da distribuição Poisson é dada por:

$$L(Z)\begin{cases} \left(\frac{c_Z}{E[c_Z]}\right)^{c_Z}\left(\frac{C-c_Z}{C-E[c_Z]}\right)^{C-c_Z}, & \text{se } c_Z > E[c_Z] \\ 1 & , \text{ caso contrário.} \end{cases}$$

em que C é o número total de casos na região em estudo, c_Z é o total de casos na zona Z e $E[c_Z]$ é o número esperado de casos na zona Z. Para detectar a zona como sendo o conglomerado mais provável é escolhido a zona $\widehat{Z}$ para o qual a função do teste da razão de verossimilhança é maximizada. A distribuição de $L(Z)$, sob hipótese nula, é obtida via simulação de Monte Carlo e seu algoritimo circular para detecção de *clusters* é realizado seguindo os seguintes passos:

i) Escolhe-se um centroide de uma área da região de estudo;

ii) Calcula-se a distância entre o centroide escolhido no passo i) e os demais centroides da região, ordena-os em ordem crescente e guarda-os em um vetor;

iii) A partir do centroide escolhido no passo i), cria-se um círculo centrado nesse ponto e continuamente aumenta-se seu raio de acordo com as distâncias encontradas no passo ii). Para cada novo centroide que entra no círculo, atualiza-se o número de casos e população dentro do círculo Z. Calcula-se o $L(Z)$ para cada par de casos e população dentro do círculo Z e registra-se o círculo com maior $L(Z)$;

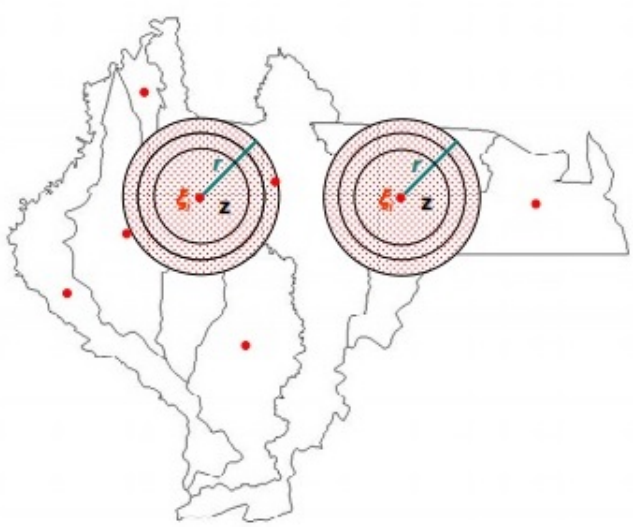
iv) Repete-se os passos i), ii) e iii) para cada centroide da região de estudo;

v) Aplica-se simulações de Monte Carlo para avaliação da significância do teste:
(1) Gera-se X conjuntos de casos independentes, em que cada réplica possui o mesmo número de casos C do conjunto de dados original. Cada novo conjunto de dados C é distribuído ao acaso entre as m as áreas da região de estudo, sob a hipótese nula;
(2) Para cada um dos X conjuntos de dados gerados, calcula-se a estatística do teste da razão de verossimilhança, obtendo-se $L_1, L_2, L_3, ..., L_X$;
(3) Ordena-se os valores de L dos X conjuntos simulados e observados no conjunto de dados original. Determine o posto da estatística L associado ao conjunto de dados original por Θ. Se Θ estiver entre os $100\alpha\%$ maiores postos, rejeite a hipótese nula

ao nível de significância de α. O valor-p associado a este teste é $1-\Theta/(X+1)$;
(4) Se a hipótese nula for rejeitada, então a zona $\widehat{Z}$ associada com a máxima verossimilhança do modelo alternativo é o *cluster* mais provável.

O raio dessa zona Z varia entre zero e um número real maior que zero, em que normalmente assume valor em até 50% do tamanho da população de cada área, seguindo o modelo de Poisson. Na Figura 3.14 é possível observar um exemplo hipotético de como se comporta essas zonas no mapa em uma região de estudo.

Figura 3.14 – Exemplo hipotético de varredura espacia da estatística Scan.



Fonte: Balieiro (2008).

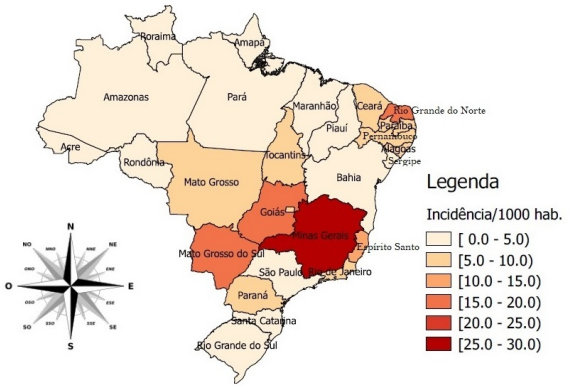
Os estudos da estatística Scan neste trabalho são voltados para a forma circular do teste, mas muitos outros métodos tem sido desenvolvidos para detecção de ***clusters*** de forma não circular. Um desses testes foi desenvolvido por Kulldorff et al. (2006), que propôs um novo método de análise da estatística Scan com janelas em formas de elipses. Uma elipse pode ser definida pelas as coordenadas (x,y) de seu centroide, com o comprimento de seus eixos maior e menor e o ângulo entre seu eixo maior e o eixo das abscissas (OLIVEIRA, 2017). O método elíptico só apresenta vantagens quando o verdadeiro ***cluster*** em estudo possui forma alongado, podendo assim a forma elíptica captar maior números de casos. Portanto, em geral, a estatística Scan circular ainda tem-se mostrado mais eficiente que os demais testes não circulares.

3.4 Problemática da dengue

Nas últimas décadas a dengue tornou-se uma das doenças mais importantes transmitida por mosquito e que afetam os seres humanos no mundo. Ela é uma doença viral que se espalha rapidamente, principalmente em países subtropicais e tropicais, onde as condições ambientais favorecem o desenvolvimento e a proliferação do mosquito. Atualmente não existe uma vacina ou uma terapia etiológica eficaz para dengue, de maneira que a única forma de combate dessa doença é por meio de um controle vetorial. Esse controle se torna uma tarefa difícil, devido às várias complexidades de configurações urbanas, como falhas em programas de controle da doença, saneamento básico deficiente, condições ambientais favoráveis para o mosquito (temperatura, altitude e umidade do ar), disponibilidade de potenciais locais de reprodução (esgotos, lixos, etc) e fatores socioeconômicos envolvidos.

No Brasil, a dengue é considerada um sério problema de saúde pública, devido principalmente às epidemias constantes e contínuas a cada ano, sendo possível encontrar a presença do mosquito em todo território nacional. Segundo o Ministério da Saúde (MINISTÉRIO DA SAÚDE, 2017), no ano de 2016 foram registrados 1.500.535 casos confirmados de dengue no país. A região Sudeste registrou o maior número de casos prováveis (858.273 casos; 57,2%) em relação ao total do país, seguida das regiões Nordeste (324.815 casos; 21,6%), Centro Oeste (205.786 casos; 13,7%), Sul (72.650 casos; 4,8%) e Norte (39.011 casos; 2,6%). Ainda com os dados do Ministério da Saúde, foi possível construir um mapa de incidência dos casos de dengue no país em 2016, como mostra a Figura 3.15.

Figura 3.15 – Mapa de incidência dos casos de dengue no Brasil no ano de 2016.



Fonte: Próprio autor.

Observar-se no mapa da Figura 3.15, que os estados da região nordeste, sudeste e centro-oeste foram os que tiveram maior incidência do número de casos no ano de 2016. Destaque para o estado de Minas Gerais, que teve a maior incidência da doença no país. Diante do exposto, pode-se inferir que o uso da estatística espacial, mais especificamente da análise de dados de área pode ajudar a entender o comportamento espacial da incidência dos casos de dengue, indicando as áreas de maior risco da doença, para que os órgãos públicos responsáveis possam tomar medidas de combate e prevenção.

4 MATERIAL E MÉTODOS

Neste trabalho foi feita uma revisão teórica sobre análise de dados de área e suas principais ferramentas estatísticas, utilizando os métodos de autocorrelação global, autocorrelação local e teste para detecção de *clusters*. Para aplicação prática da teoria foi usado uma banco de dados dos casos de dengue da cidade de Campina Grande-PB. O cálculo para obter a taxa de incidência de dengue em cada bairro no ano de 2016 foi feita da seguinte forma:

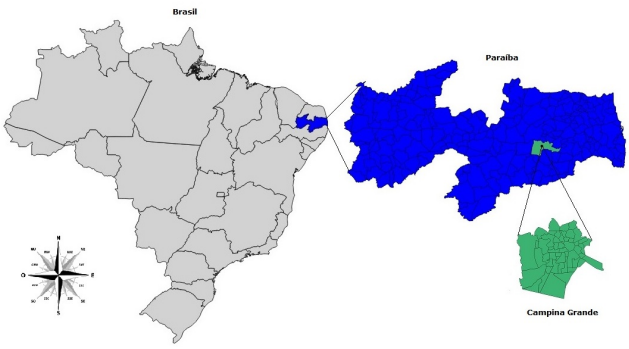
$$\text{Incidência} = \frac{\text{Nº de casos de dengue no bairro}}{\text{Total da população residente no bairro}} / 1000 \text{ habitantes}$$

Na avaliação do nível de significância dos índices globais e locais foram geradas 999 permutações. O modelo probabilístico utilizado nas análises foi o de Poisson. As análises foram realizadas pelo programa R (R Core Team, 2017) e SaTScan (KULLDORFF, 2016).

4.1 Área de estudo

O estudo será desenvolvido com dados da cidade de Campina Grande - PB, descrito no Anexo A deste trabalho. Essa cidade está situada entre o alto sertão e a zona litorânea, na mesorregião do agreste paraibano. Campina Grande é conhecida por seus moradores como sendo a Rainha da Borborema, por estar localizada na serra da Borborema. Localizada na região nordeste do país, com latitude $7^\circ\ 13'\ 50''$ sul, por longitude $35^\circ\ 52'\ 52''$ oeste e estando a uma altitude de 551 metros, possuindo uma área de $593{,}026\ \text{km}^2$. Pode-se observar na Figura 4.1, a localização da cidade no estado da Paraíba.

Figura 4.1 – Localização da área de estudo.



Fonte: Próprio autor.

No último censo 2010, a cidade de Campina Grande apresentava uma população de 385.213 pessoas, com uma densidade demográfica de $648,31 \ \ \text{hab/km}^2$ e uma população estimada em 2017 de 410.332 pessoas. Campina Grande apresenta 84,1 % de domicílios com esgotamento sanitário adequado e 19,4 % de domicílios urbanos em vias públicas com urbanização adequada (presença de bueiro, calçada, pavimentação e meio-fio). O clima de Campina Grande, segundo a classificação climática de Köppen-Geiger, é Aw, considerado como seco sub-úmido. Possui temperatura máxima média anual de $28,7°\text{C}$ e mínima de $19,8°\text{C}$, com período chuvoso entre os meses de março a julho (JÚNIOR, 2006).

4.2 Banco de dados

O banco de dados utilizado neste estudo é proveniente da Secretária de Saúde da cidade de Campina Grande - PB. Trata-se do número de casos confirmados de dengue no ano 2016, cujas informações foram obtidas por meio de prontuários médicos, onde são colhidas informações dos pacientes com a doença. Nesse período foram notificados um total de 430 casos confirmados da doença na cidade. Todos os dados estão descritos no Anexo B.

4.3 Programa

As análises e modelagens referente ao estudo de análise de dados de área foram desenvolvidas no programa R (R Core Team, 2017), utilizando-se os seguintes pacotes:

- *maptools*, desenvolvido por Bivand e Lewin-Koh (2017), é utilizado para manipulação de mapas e dados geográficos;
- *sp*, desenvolvido por Pebesma e Bivand (2005), é utilizado para representação de dados espaciais no R, como exemplo, conversão de coordenadas geográficas (lat/long) para o formato de coordenadas UTM (Universal Transversa de Mercator);
- *spdep*, desenvolvido por Bivand (2017), é utilizado para análise de dados de áreas, com as funções para calcular as estatísticas I de Moran e c de Geary, gerar os mapas de autocorrelação global e local, além do diagrama de dispersão de Moran;
- *smerc*, desenvolvido por French (2015),é utilizado para detecção de *clusters* e possui a função para cacular a estatística Scan.

A análise referente ao teste de detecção de *clusters* com a estatística Scan foi realizada com o programa SaTScan (KULLDORFF, 2016), versão 9.4.

5 RESULTADOS E DISCUSSÕES

Nessa seção serão mostrados os passos necessários para realizar uma análise de dados de áreas e o teste Estatística Scan. Além disso, serão mostrados e discutidos os resultados obtidos dessas análises. O estudo foi aplicado ao banco de dados dos casos de dengue da cidade de Campina Grande-PB no ano de 2016. Os resultados e discussões serão divididos em duas partes: na primeira parte serão mostradas as análises de dados de áreas realizadas no programa R (R Core Team, 2017) e na segunda parte serão feitas as análises da Estatística Scan nos programas SaTScan e programa R (R Core Team, 2017).

5.1 Análise de dados de área no programa R

Essa análise será composta pelas seguintes etapas: análise exploratória de dados espaciais, construção da matriz de vizinhança W, cálculo dos índices de autocorrelação global, cálculo dos índices de autocorrelação local, construção do mapa de probabilidade de Moran local, construção do diagrama de dispersão de Moran, construção do mapa LISA e cálculo da estatística Scan.

Antes de iniciar qualquer procedimento no programa R é preciso limpar sua memória. Essa limpeza é feita por meio da função $rm(list = ls())$, que apaga todos os objetos listados e salvos na área de trabalho. Essa limpeza garante que não fique nenhuma informação residual de análises anteriores que possam influenciar em análises futuras feitas no programa. A função $setwd()$ serve para mudar o diretório de trabalho, ou seja, informa ao R onde encontra-se a pasta com o *Script*, o arquivo *"Shapefile"* e o arquivo de dados do estudo em seu computador. Neste trabalho as informações estão contidas na pasta Análises, no local ("$C:/Users/Pablo/Desktop/Analises$"). Para certificar se o diretório de trabalho está no endereço correto, utiliza-se a função $getwd()$. As linhas de comando do programa estão descritas a seguir:

```
> rm(list=ls())
> setwd ("C:/Users/Pablo/Desktop/Análises")
> getwd()
```

Para iniciar as análises no R, primeiramente precisa-se instalar os pacotes necessários para o andamento da análise. Essa instalação pode ser feita pelo seguinte comando descrito a seguir:

```
> install.packages(c("maptools", "sp", "spdep", "classInt",
+                    "RColorBrewer"), dep=T)
```

Após a instalação dos pacotes é necessário requeri-los por meio da função *require*. Cada pacote possui funções específicas para cada etapa da análise. O pacote *maptools* possui funções para importação e manipulação de mapas e dados geográficos. A função *gpclibPermit*() do pacote *maptools* serve para habilitar licença de uso em pacotes de análise espacial. O pacote *sp* possui as funções para representação de dados espaciais. O pacote *spdep* possui as funções para análises de dados de áreas. O pacote *classInt* é utilizado nessa análise para facilitar a divisão de dados em classes por critérios e o pacote *RColorBrewer* é utilizado aqui para criar palhetas de cores nas visualizações em mapas. As linhas de comando para requiri os pacotes estão descritas a seguir:

```
> require(maptools)
> gpclibPermit()
> require(sp)
> require(spdep)
> require(classInt)
> require(RColorBrewer)
```

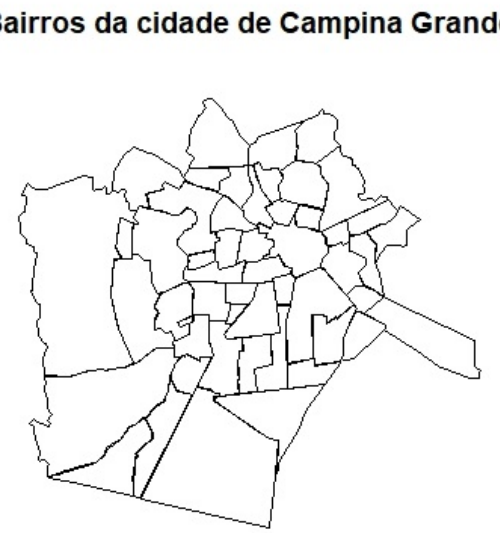
O primeiro passo da análise é importar o mapa da cidade de Campina Grande-PB para o programa R. A importação do mapa é feito por meio da função *readShapePoly* do pacote *sp*. No primeiro argumento da função *readShapePoly* é necessário colocar o nome que foi atribuído a camada *Shapefile* com extensão *.shp*, que nesse caso foi chamado de *Mapa_Campina_Grande*. O segundo argumento, *IDvar* =, refere-se à informação da variável que contém os IDs que dão formas aos bairros da cidade (Polígonos). Essa informação é encontrada na primeira coluna, chamada de *Id*, da planilha denominada por

Mapa_Campina_Grande que possui extensão *.dbf*. A função *readShapePoly* foi atribuído a um objeto chamado de *ShapeCG*. Portanto, para visualizar o mapa de Campina Grande basta fazer o *plot* do objeto *ShapeCG*. As linhas de comando para plotar o mapa estão descritas a seguir:

```
> ShapeCG<- readShapePoly("Mapa_Campina_Grande.shp",IDvar="Id")
> plot(ShapeCG, axes=F)
> title("Bairros da cidade de Campina Grande")
```

Na Figura 5.1 pode ser observado o mapa com os 51 bairros da cidade de Campina Grande-PB.

Figura 5.1 – *"Shapefile"* com os bairros da cidade de Campina Grande-PB.



Para saber a dimensão do mapa, pode-se usar função $dim(ShapeCG)$. Seu resultado mostrará um vetor com o toral de 51 linhas e 2 colunas que compõe o arquivo *Shapefile*. Com a função $head(ShapeCG@data, 51)$ pode-se visualizar um resumo do arquivo *Shapefile*, onde o seu resultado mostrará as seis primeiras linhas e as duas colunas do vetor, ou seja, mostrará os seis primeiros bairros em ordem alfabética e suas duas colunas que representam o nome das variável no arquivo *Shapefile*. As linhas de comando para visualização estão descritas a seguir:

```
> dim(ShapeCG)
[1] 51  2
> head(ShapeCG@data,51)
Id                  Nome
1   1    Acácio Figueiredo
2   2          Alto Branco
3   3                Araxá
4   4   Bairro das Cidades
5   5    Bairro das Nações
6   6 Bairro Universitário
```

É bastante comum o aparecimento de variáveis em formato de texto em banco de dados, dessa forma o programa precisa entender que aquela variável é composta por caracteres. A variável *Nome* do arquivo *Shapefile* possui os nomes dos 51 bairros que compõe o mapa da cidade. Dessa forma é necessário que o programa R reconheça esses bairros como caracteres e isso é feito por meio da função *as.character*. As linhas de comando para reconhecer os bairros como caracteres estão descritas a seguir:

```
> ShapeCG@data$Nome <- as.character(ShapeCG@data$Nome)
> ShapeCG@data$Nome
[1] "Acácio Figueiredo"    "Alto Branco"          "Araxá"
[4] "Bairro das Cidades"   "Bairro das Nações"    "Bairro Universitário"
[7] "Bela Vista"           "Bodocongó"            "Castelo Branco"
[10] "Catolé"               "Centenário"           "Centro"
[13] "Conceição"            "Cruzeiro"             "Cuités"
[16] "Dinamérica"           "Distrito Industrial"  "Estação Velha"
[19] "Itararé"              "Jardim Continental"   "Jardim Paulistano"
[22] "Jardim Quarenta"      "Jardim Tavares"       "Jeremias"
[25] "José Pinheiro"        "Lauritzen"            "Liberdade"
[28] "Louzeiro"             "Malvinas"             "Mirante"
[31] "Monte Castelo"        "Monte Santo"          "Nova Brasília"
[34] "Novo Bodocongó"       "Palmeira"             "Pedregal"
[37] "Prata"                "Presidente Médice"    "Quarenta"
[40] "Ramadinha"            "Sandra Cavalcante"    "Santa Cruz"
[43] "Santa Rosa"           "Santa Terezinha"      "Santo Antônio"
[46] "São josé"             "Serrotão"             "Tambor"
[49] "Três Irmãs"           "Velame"               "Vila Cabral"
```

Após organizar os dados do arquivo *Shapefile* é preciso importar o arquivo de dados com os casos de dengue da cidade de Campina Grande-PB no ano de 2016. Esses dados foram obtidos na Secretária de Saúde da cidade de Campina Grande - PB e estão disponíveis no Anexo C deste trabalho. A importação dos dados é feita pela função *read.table*. No primeiro argumento da função *read.table* é necessário colocar o nome que foi atribuído a base dados, que nesse caso foi chamado de *Dengue*2016. Essa base de dados foi salva em um arquivo de texto com extensão *.txt*. No entanto, a função *read.table* também aceita o formato de planilha com extensão *.csv*. O segundo argumento, $header = TRUE$, indica ao R que a base de dados possui na primeira linha cabeçalho, que corresponde ao nome das variáveis. A base de dados *Dengue*2016 é composta por 4 variáveis como mostra a função *str*(*dengue*). As linhas de comando para importar e visualizar a base de dados estão descritas a seguir:

```
> dengue <- read.table("Dengue2016.txt",header=TRUE)
> str(dengue)
'data.frame': 51 obs. of  4 variables:
$ Id   : int  50 17 1 4 49 38 47 40 29 42 ...
$ Casos: int  4 1 3 5 7 3 1 3 20 9 ...
$ Pop  : int  6755 2819 10408 6762 13663 4810 7734 2428 43323 10536 ...
$ Incid: num  0.592 0.355 0.288 0.739 0.512 ...
```

Após ter importado os dados do arquivo *Shapefile* e os dados do arquivo *Dengue*2016 é preciso ordenar os dados corretamente para compatibilizar a ordem dos bairros do arquivo *Shapefile* com a ordem dos bairros do arquivo *Dengue*2016. A ordenação desses dois arquivos de dados é feita pela função *match*(*ShapeCG@data*Id, dengue*Id*), por meio da variável Id que está presente nos dois arquivos. As linhas de comando para ordenação dos dados estão descritas a seguir:

```
ind <- match(ShapeCG@data$Id, dengue$Id)
> ind
[1]  3 41 31  4 34 30 37 13 47 23 14 45 42 16 32 11  2 22 21 35 17 51
[23] 44 33 48 43 19 40  9 28 49 36 50 29 39 15 38  6 18  8 25 10 12 27
[45] 46 24  7 20  5  1 26
```

Após ordenar os dois arquivos de dados é preciso concatena-los para que possam se tornar um único arquivo de dados. Isso é feito pela função *spCbind*(*ShapeCG*, *dengue*) e pode ser visualizada por meio da função *head*(*ShapeCG@data*). As linhas de comando para concatenar o arquivo de dados estão descritas a seguir:

```
> dengue <- dengue[ind,]
> row.names(dengue) <- ShapeCG$Id
> ShapeCG <- spCbind(ShapeCG, dengue)
> head(ShapeCG@data)
Id                  Nome Id.1 Casos   Pop Incid
1  1    Acácio Figueiredo    1     3 10408 0.288
2  2          Alto Branco    2     2  9904 0.202
3  3                Araxá    3     4  1960 2.041
4  4   Bairro das Cidades    4     5  6762 0.739
5  5    Bairro das Nações    5     2  1573 1.271
6  6 Bairro Universitário    6     2  4176 0.479
```

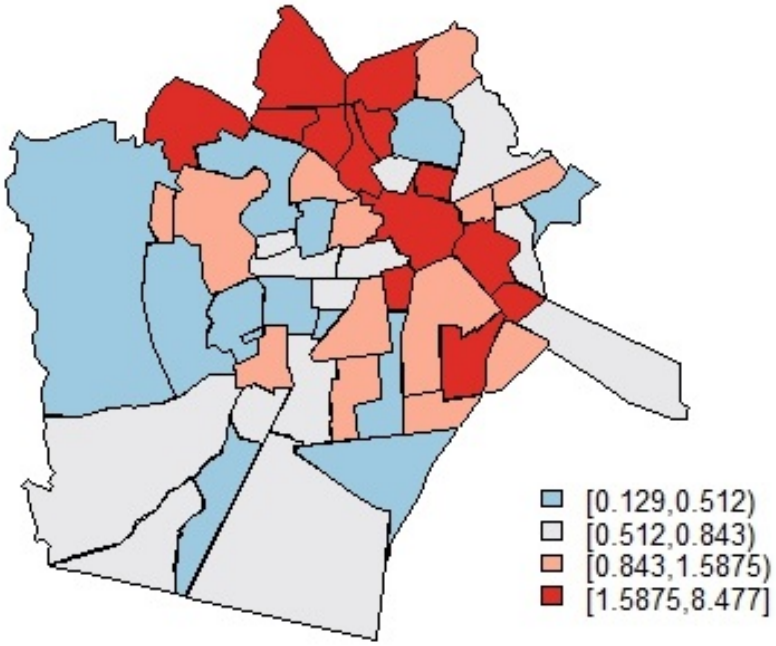
Agora que o arquivo de dados está pronto, pode-se prosseguir para etapa seguinte, que é a análise exploratória de dados espaciais. Essa análise busca identificar determinados padrões de dependência espacial por meio de mapas temáticos. Para construção desses mapas é necessário dividir os valores da variável taxa de incidência de dengue em quantis. Essa divisão é feita pela função *classIntervals*, com o argumento *style* = "*quantile*". As linhas de comando para criação do mapa de atributos estão descritas a seguir:

```
> INT1 <- classIntervals(ShapeCG$Incid, n=4, style="quantile")
> CORES.1 <- c(rev(brewer.pal(3, "Blues")), brewer.pal(3, "Reds"))[-1]
> COL1 <- findColours(INT1, CORES.1)
> plot(ShapeCG, col=COL1)
> TB1 <- attr(COL1, "table")
> legtext <- paste(names(TB1))
> legend("bottomright", fill=attr(COL1, "palette"), legend=legtext,
+        bty="n",cex=0.8)
```

No mapa de atributos, Figura 5.2, pode-se observar uma certa tendência nos bairros de coloração vermelha, ou seja, o mapa nos mostra indícios da existência de uma possível dependência espacial nesses bairros. No entanto, até o momento não foi calculada nenhuma estatística que confirme tal suposição. Logo, essa suposição só poderá ser

confirmada com a aplicação dos índices de autocorrelação espacial na sequência desta análise.

Figura 5.2 – Taxa de incidência dos casos de dengue da cidade de Campina Grande-PB.



Antes de partir para os cálculos dos índices de autocorrelação espacial é necessário construir a matriz de vizinhança W. Para construir a matriz é preciso encontrar os pontos das coordenadas do centóide de cada bairro pela função *coordinates*(*ShapeCG*) e atribuir suas coordenadas ao objeto *points*. Em seguida, por meio da função *dnearneigh*, cria-se o objeto denominado de *dnb*, que contém a lista de vizinhos das áreas do mapa. Por fim, construí-se a matriz de vizinhança W com a função *nb2mat*. No primeiro argumento da função *nb2mat* é atribuído o objeto *dnb* com a lista de vizinhos e no segundo argumento é atribuído o tipo da matriz de vizinhança que deseja-se obter. Se nesse segundo argumento for colocado *style* = "*B*" a matriz será binária, caso seja colocado *style* = "*W*" a matriz será normalizada. As linhas de comando para construção da matriz de vizinhança W estão descritas a seguir:

```
> ccods = coordinates(ShapeCG)
> points = cbind(ccods[,1],ccods[,2])
> dnb = dnearneigh(points,0,1807)
> W.Bin= nb2mat(neighbours = dnb, style = "B")
> W.Normal= nb2mat(neighbours = dnb, style = "W")
```

Agora que a matriz de vizinhança W está pronta, pode-se calcular os índices de autocorrelação espacial. Os índices de autocorrelação espacial global serão calculados pelos testes de normalidade e permutação. O índice de Moran global é calculado pela função *moran.test* do pacote *spdep*, em que no seu primeiro argumento é atribuído a taxa de incidência de dengue, no segundo argumento a matriz de vizinhança W e no terceiro argumento o tipo do teste. Se no terceiro argumento for colocado $randomisation = FALSE$ o teste será por normalidade, caso seja colocado $randomisation = TRUE$ o teste será por permutação. As linhas de comando para calcular o índice global de Moran pelos testes de normalidade e permutação estão descritas a seguir:

```
> moran.test(ShapeCG$Incid, listw=nb2listw(dnb, style = "W"),
             randomisation= FALSE)

Moran I test under normality

data:  ShapeCG$Incid
weights: nb2listw(dnb, style = "W")

Moran I statistic standard deviate = 2.2397, p-value = 0.01256
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
0.163939212      -0.020000000       0.006744909
```

```
> moran.test(ShapeCG$Incid,listw=nb2listw(dnb, style = "W"),
             randomisation= TRUE)

Moran I test under randomisation

data:  ShapeCG$Incid
weights: nb2listw(dnb, style = "W")

Moran I statistic standard deviate = 2.5002, p-value = 0.006206
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
0.163939212      -0.020000000       0.005412427
```

De forma análoga ao índice de Moran global é calculada a estatística c de Geary. Sua estatística é calculada pela função *geary.test* do pacote *spdep*, onde os argumentos de sua função são iguais aos do índice de Moran global definidos anteriormente. As linhas de comando para calcular a estatística c de Geary pelos testes de normalidade e permutação estão descritas a seguir:

```
> geary.test(ShapeCG$Incid, listw=nb2listw(dnb, style = "W"),
             randomisation= FALSE)

Geary C test under normality

data:  ShapeCG$Incid
weights: nb2listw(dnb, style = "W")

Geary C statistic standard deviate = 3.3474, p-value = 0.0004078
alternative hypothesis: Expectation greater than statistic
sample estimates:
Geary C statistic       Expectation          Variance
0.704243292       1.000000000       0.007806362
```

```
> geary.test(ShapeCG$Incid, listw=nb2listw(dnb, style = "W"),
             randomisation=TRUE)

Geary C test under randomisation

data:  ShapeCG$Incid
weights: nb2listw(dnb, style = "W")

Geary C statistic standard deviate = 2.5962, p-value = 0.004713
alternative hypothesis: Expectation greater than statistic
sample estimates:
Geary C statistic       Expectation          Variance
0.70424329        1.00000000        0.01297726
```

Um outro índice de autocorrelação global bastante difundido é a estatística G de Getis-Ord. Essa estatística pode ser calculada pela função *globalG.test* do pacote *spdep*, em que no seu primeiro argumento é atribuído a taxa de incidência de dengue e no segundo argumento a matriz de vizinhança W. Essa estatística é calculada apenas para o teste de normalidade com matriz de vizinhança W binária. As linhas de comando para calcular a estatística G de Getis-Ord global estão descritas a seguir:

```
globalG.test(ShapeCG$Incid, nb2listw(dnb, style="B"))

Getis-Ord global G statistic

data:  ShapeCG$Incid
weights: nb2listw(dnb, style = "B")

standard deviate = 0.94762, p-value = 0.1717
alternative hypothesis: greater
sample estimates:
Global G statistic        Expectation           Variance
0.174894035        0.149803922        0.000701026
```

O último teste global dessa análise é o Empirical Bayes Index (EBI). Diferente dos demais testes globais já analisados, o teste para o EBI é exclusivo para taxas e tem-se apenas a opção de teste da permutação. O índice EBI é calculado pela função *EBImoran.mc*

do pacote *spdep*, em que no seu primeiro argumento é atribuído a quantidade de casos de dengue por bairro, no segundo argumento a população por bairro, no terceiro argumento a matriz de vizinhança W e no quarto argumento o número de permutações. As linhas de comando para calcular o índice EBI estão descritas a seguir:

```
> EBImoran.mc(ShapeCG$Casos,ShapeCG$Pop, nb2listw(dnb, style="W"),
+             nsim=999)

Monte-Carlo simulation of Empirical Bayes Index (mean subtracted)

data:  cases: ShapeCG$Casos, risk population: ShapeCG$Pop
weights: nb2listw(dnb, style = "W")
number of simulations + 1: 1000

statistic = 0.14504, observed rank = 989, p-value = 0.011
alternative hypothesis: greater
```

Na Tabela 5.1 encontram-se todas as estimativas dos índices de autocorrelação global calculados anteriormente e seus respectivos valores-p. É observado na tabela que os índices I de Moran e c de Geary foram significativos pelo teste por normalidade e pelo teste por permutação. A estatística EBI também mostrou-se significativa para variável taxa de incidência de dengue. A estatística G de Getis-Ord foi a única não significativa entre os testes globais. Logo, pelos índices I de Moran, c de Geary e estatística EBI é possível afirmar que existe uma autocorrelação espacial sobre na de incidência de dengue na cidade de Campina Grande-PB no ano de 2016.

Tabela 5.1 – Estimativas dos índices de autocorrelação global.

Tipo de teste	Índice	Estatística	Valor-p
Normalidade	Moran	0,1639	$0,0125^{*}$
	Geary	0,7042	$0,0004^{*}$
	Getis-Ord	0,1748	$0,1717^{NS}$
Permutação	Moran	0,1639	$0,0062^{*}$
	Geary	0,7042	$0,0129^{*}$
	EBI	0,1450	$0,0110^{*}$

Todas as análises feitas até o momento foram de escala global. No entanto, é necessário que seja feita também uma análise local do estudo. Essa análise pode ser feita pelo índice local de autocorrelação espacial (LISA). Para isso é preciso calcular o índice de

Moran local, que é feito pela função *localmoran* do pacote *spdep*. Nessa função, em seu primeiro argumento é atribuído a taxa de incidência de dengue e no segundo argumento a matriz de vizinhança W. As linhas de comando para calcular o índice de Moran local estão descritas a seguir:

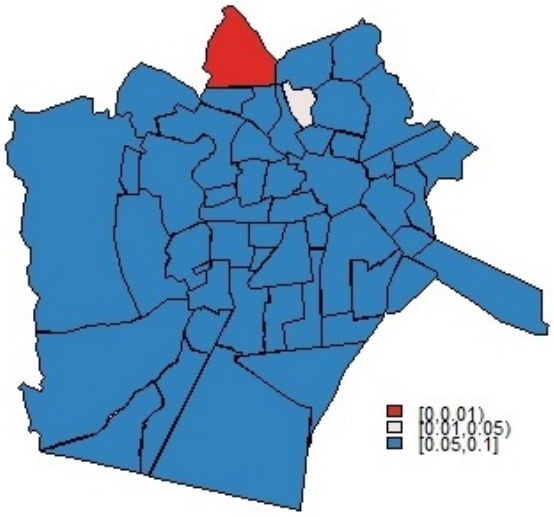
```
> ShapeCG.mloc <- localmoran(ShapeCG$Incid,
+                 listw=nb2listw(dnb, style="W"))
> head(ShapePB.mloc)
               Ii      E.Ii   Var.Ii       Z.Ii    Pr(z > 0)
       1  0.316728611 -0.02 0.25180958  0.67103303 0.2510997
       2 -0.259768537 -0.02 0.06758998 -0.92225503 0.8218022
       3  0.380867861 -0.02 0.10142623  1.25871118 0.1040673
       4  0.237068802 -0.02 0.38339501  0.41517017 0.3390087
       5 -0.005769156 -0.02 0.18601687  0.03299548 0.4868391
       6 -0.011084722 -0.02 0.08732779  0.03016884 0.4879662
```

Por meio dos valor-p do índice de Moran local é possível construir um mapa de probabilidades. As linhas de comando para construção do mapa estão descritas a seguir:

```
> INT4 <- classIntervals(ShapePB.mloc[,5], style="fixed",
+                        fixedBreaks=c(0,0.01, 0.05, 0.10))
> CORES.4 <- c(rev(brewer.pal(3, "Reds")), brewer.pal(3, "Blues"))
> COL4 <- findColours(INT4, CORES.4)
> plot(ShapeCG, col=COL4)
> TB4 <- attr(COL4, "table")
> legtext <- paste(names(TB4))
> legend("bottomright", fill=attr(COL4, "palette"), legend=legtext,
+        bty="n", cex=0.7, y.inter=0.7)
```

De acordo da palheta de cores do mapa de probabilidade, Figura 5.3, pode-se observar que apenas dois bairros foram significativos no estudo considerando o nível de significância até 5%. O bairro Cuité foi significativo até 1% e o bairro Louzeiro foi significativo até 5%. Isto significa que existe uma dependência espacial em relação a incidência de dengue nesses dois bairros. Esse resultado confirma a suposição da existência de dependência espacial nos bairros de coloração vermelha que foi levantada no gráfico de atributos da Figura 5.3.

Figura 5.3 – Mapa de significâncias com os valores-p para o índice de Moran local.

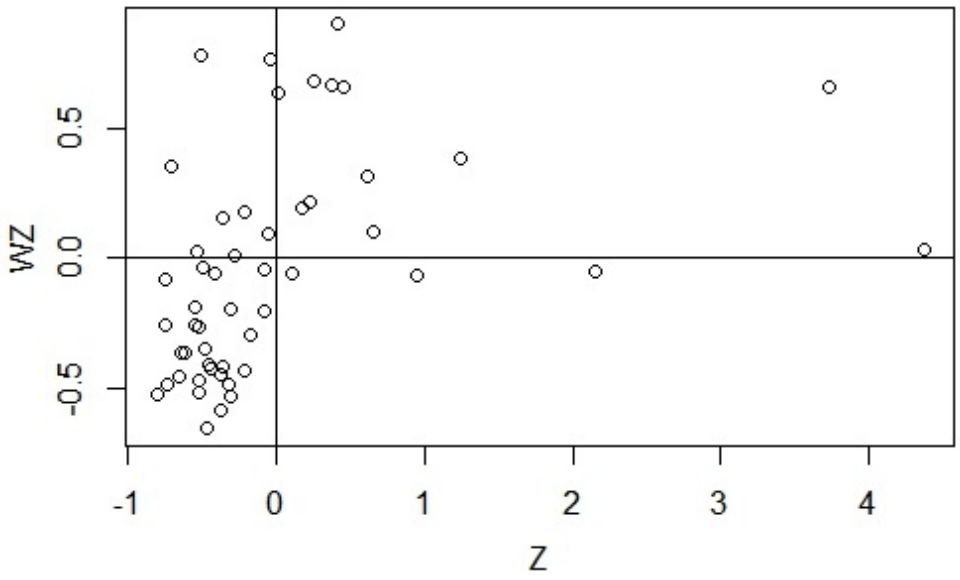


O próximo passo da análise é construir o diagrama de dispersão de Moran. A ideia é comparar os valores normalizados da incidência de dengue em um bairro com a média dos seus vizinhos, construindo um gráfico bidimensional de z (valores normalizados) por wz (média dos vizinhos normalizados), onde os valores de z são normalizados pela função *scale*. As linhas de comando para construção do diagrama de dispersão estão descritas a seguir:

```
> # Montando matrix W de vizinhança
> ShapeCG.nb1.mat <- nb2mat(dnb)
> # Incidência padronizado
> Dengue_SD <- scale(ShapeCG$Incid)
> # Média padronizada nos vizinhos
> Dengue_W <- ShapeCG.nb1.mat %*% Dengue_SD
> # Diagrama de espalhamento de Moran
> plot(Dengue_SD, Dengue_W,xlab="Z",ylab="WZ")
> abline(v=0, h=0)
```

No diagrama de dispersão de Moran, Figura 5.4, pode-se observar que a maioria dos valores estão concentrados no segundo quadrante, ou seja, existe um grande número de bairros com uma baixa incidência de casos de dengue, rodeados por bairros que apresentam também uma baixa incidência de casos.

Figura 5.4 – Diagrama de dispersão de Moran.

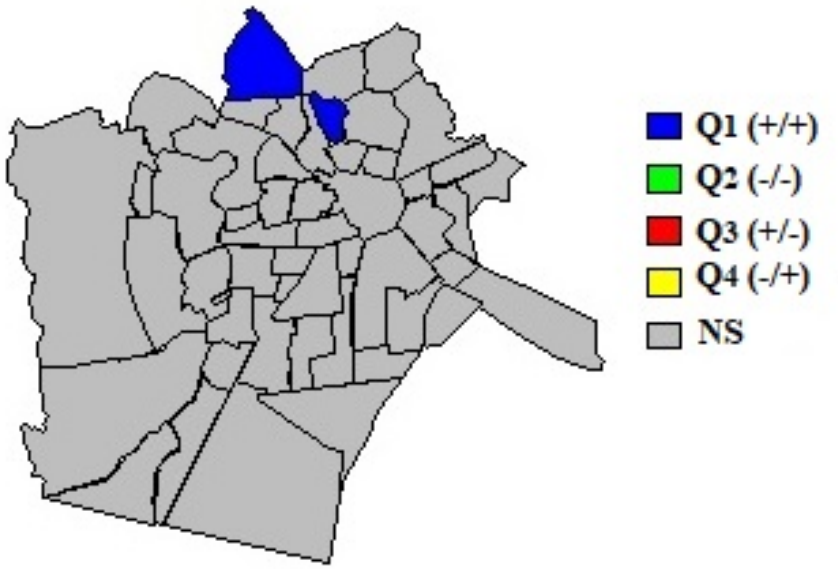


A última etapa dessa análise é construir o mapa do índice local de autocorrelação espacial, ou seja, o mapa LISA. Para construir o mapa, primeiramente tem-se que criar um vetor com os valores-p do índice de Moran local e atribuir esses valores aos critérios definidos em cada quadrante do diagrama de dispersão de Moran. Em seguida se estipula o valor da significância que deseja-se construir o mapa, nesse caso 5%. Por fim são definidas as cores que cada quadrante representará no mapa. As linhas de comando para construir o mapa LISA estão descritas a seguir:

```
> Q <- vector(mode = "numeric", length = nrow(ShapePB.mloc))
> Q[(Dengue_SD>0  & Dengue_W > 0)] <- 1
> Q[(Dengue_SD<0  & Dengue_W < 0)] <- 2
> Q[(Dengue_SD>=0 & Dengue_W < 0)] <- 3
> Q[(Dengue_SD<0  & Dengue_W >= 0)]<- 4
> signif=0.05
> Q[ShapePB.mloc[,5]>signif]<-5
> CORES.5 <- c("blue", "green" , "red", "yellow", "gray",
+              rgb(0.95,0.95,0.95))
> plot(ShapeCG, col=CORES.5[Q])
> legend("bottomright", c("Q1(+/+)", "Q2(-/-)",
+       "Q3(+/-)", "Q4(-/+)","NS"), fill=CORES.5)
```

No mapa LISA, Figura 5.5, pode-se observar que apenas dois bairros foram significativos para variável incidência de dengue e ambos pertencem ao quadrante Q1(+/+), indicando uma dependência espacial positiva. Isso significa que nesses bairros a incidência de casos de dengue é alta e suas vizinhanças também são bairros que possuem uma alta incidência de casos de dengue. Se voltarmos ao mapa de atributos da Figura 5.2 é possível observar a alta incidência de casos de dengue nessa região da cidade onde os bairros do mapa LISA foram significativos, reforçando a veracidade do resultado.

Figura 5.5 – Mapa LISA.



5.2 Análise da Estatística Scan

Um outro método utilizado neste trabalho é a Estatística Scan. Esse método busca identificar a presença de conglomerados (*clusters*) em áreas mais propensas a ocorrência do fenômeno em estudo.Neste trabalho, busca-se identificar os bairros que apresentam um risco significativamente alto de casos de dengue.

As análises pelo método da Estatística Scan podem ser realizadas de duas formas:a primeira pode ser realizada por meio do programa SaTScan (KULLDORFF, 2016); a segunda pode ser realizada pelo programa R (R Core Team, 2017), por meio dos pacotes ***smerc*** e ***rsatscan***.

- O pacote ***smerc*** foi desenvolvido por Joshua French no ano de 2015 e fornece métodos estatísticos para a análise de dados de área, com foco na detecção de *cluster*. O teste da Estatística Scan é calculado pela função *scan.test*, que realiza uma varredura espacial de forma circular, computando o número de casos nos locais de ocorrência do evento. A estatística do teste baseia-se na distribuição de Poisson, que é frequentemente utilizada para modelar o número de ocorrências de um evento por um certo período de tempo em uma determinada área.

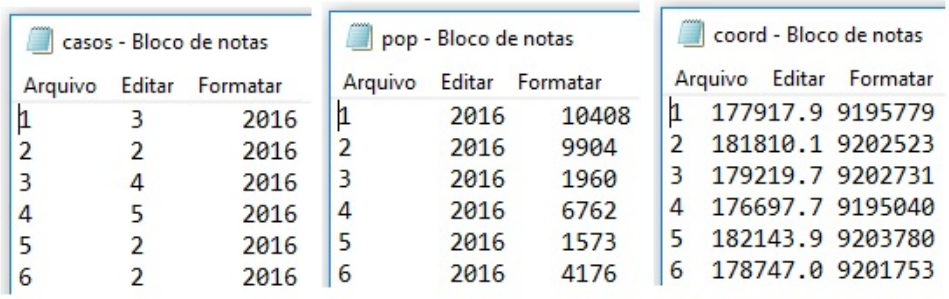
- O pacote ***rsatscan*** foi desenvolvido por Ken Kleinman no ano de 2016. Esse pacote faz a interação entre os programas R e SaTScan, possibilitando o usuário usar todas as ferramentas do SaTScan por meio de comandos no R. O pacote ***rsatscan*** não será utilizado neste trabalho e fica a critério do leitor procurar mais informações sobre ele.

5.2.1 Análise no programa SaTScan

O SaTScan é um programa livre que analisa dados espaciais, temporais e espaço-temporais, utilizando estatísticas discretas ou contínuas em suas análises. Esse programa foi desenvolvido inicialmente com o objetivo de analisar dados epidemiológicos, mas também pode analisar dados nas áreas de arqueologia, astronomia, botânica, criminologia, ecologia, economia, genética, geologia, dentre outras áreas do conhecimento. Neste trabalho nos prenderemos a analisar dados espaciais de epidemiologia utilizando estatística discreta.

Para que a análise possa ser realizada, primeiramente é necessário a criação dos aquivos de dados. Esses arquivos de dados de entrada devem estar no formato ASCII e podem ser criados no programa Bloco de Notas do *Windows*. Para o modelo Poisson precisa-se criar três arquivos de dados de entrada distintos: o primeiro deve conter o número de casos, o segundo a população residente e o terceiro as coordenadas dos centroides de cada bairro da cidade em estudo. Ainda deve ser criado um aquivo de saída em branco, onde será salvo o resultado final da análise. Todos esses arquivos de entrada e saída devem estar salvos em uma única pasta em seu computador. Neste estudo denominamos os arquivos de dados de entrada, respectivamente, da seguinte forma: **casos**, **pop** e **coord**. O arquivo de saída foi denominado de **Resultado.** Os dados de entrada devem ser inseridos no Bloco de Notas conforme mostra a Figura 5.6

Figura 5.6 – Dados de entrada.



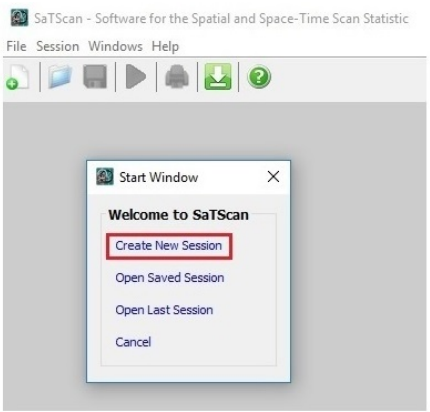
Observa-se na Figura 5.6 que cada arquivo de dados é dividido em três colunas, não possuem cabeçalho e que possuem na primeira coluna a mesma variável. Essa primeira coluna é o **Id**, de identificação de cada bairro e será utilizado pelo SaTScan para identificação dos *clusters*. As colunas em cada arquivo de dados são separadas pela tecla **Tab**. Para um melhor entendimento dos arquivos de dados, serão descritos o significado de cada coluna na Tabela 5.2.

Tabela 5.2 – Arquivos de dados.

Arquivo	**1ª coluna**	**2ª coluna**	**3ª coluna**
casos	Id	Casos	Ano
pop	Id	Ano	População
coord	Id	Latitude	Longitude

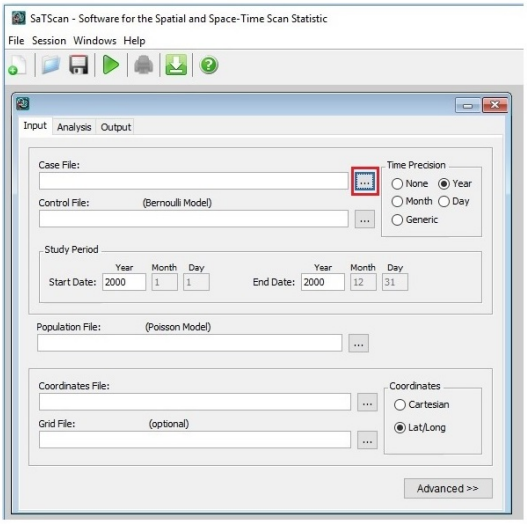
Para o arquivo de entrada **coord**, suas coordenadas podem ser do tipo Lat/Long (Latitude/Longitude) ou UTM (Universal Transverse Mercator). As coordenadas utilizadas neste trabalho são do tipo UTM. Após concluída a criação dos arquivos de dados, deve-se partir para o procedimento de análise no programa SaTScan. Ao clicar no ícone do programa abrirá a seguinte janela, como mostra a Figura 5.7.

Figura 5.7 – Tela inicial SaTScan.



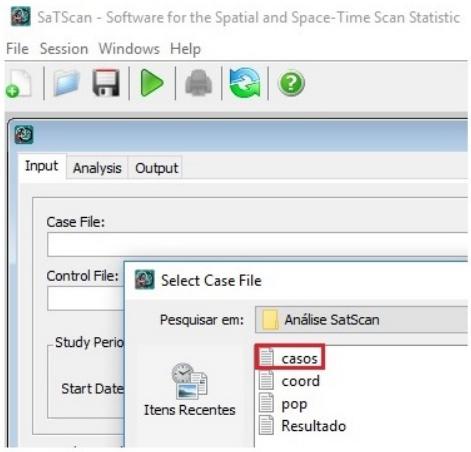
Na tela inicial do programa deve-se clicar em *Create New Session*, para iniciar uma nova sessão. Ao clicar em *Create New Session*, abrirá a seguinte tela, como mostra a Figura 5.8.

Figura 5.8 – Segunda tela SaTScan.



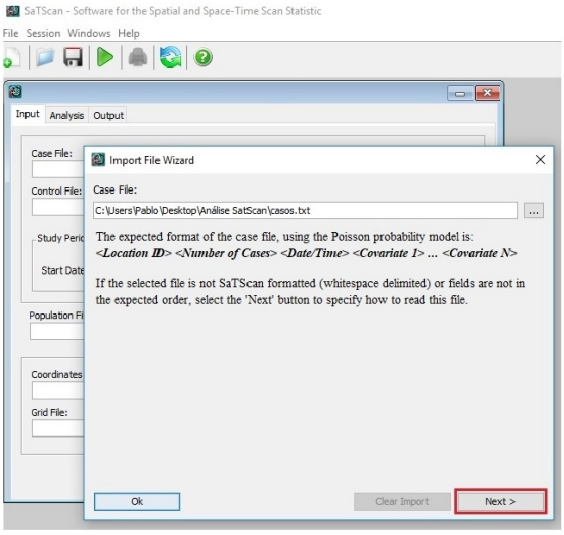
Na aba *Input*, deve-se clicar em *Case File* para selecionar o arquivo dos casos de dengue salvo em uma pasta anteriormente. É importante que a pasta contendo os arquivos de dados seja salva em um local de acesso fácil em seu computador. Neste caso a pasta foi salva na área de trabalho com o nome **Análise SaTScan**, como mostra a Figura 5.9.

Figura 5.9 – Selecionando o arquivo casos.



Ao selecionar o arquivo **casos**, o programa abrirá uma tela para definir as configurações e importação dos arquivos de dados. A tela de configurações pode ser visualizada na Figura 5.10.

Figura 5.10 – Tela de configurações do arquivo casos.

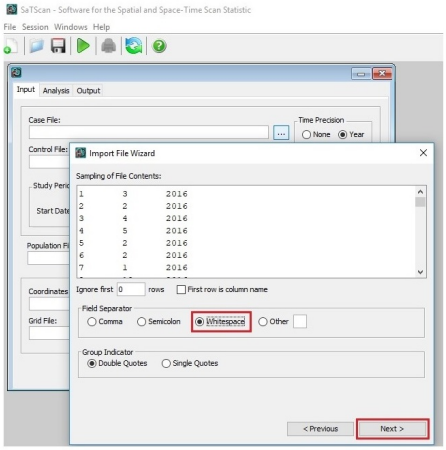


Essas configurações são necessárias para que o programa possa identificar corretamente as variáveis em estudo e possa transformar os arquivos de dados que estão no formato de extensão *.txt* para os formatos exigido pelo SaTScan. Para que a análise possa ser realizada, o SaTScan exige que os arquivos de entrada tenham as seguintes extensões:

- O arquivo contendo o número de casos tenha a extensão *.cas*;
- O arquivo contendo a população das áreas tenha a extensão *.pop*;
- O arquivo contendo a localização dos centroides tenha a extensão *.geo*.

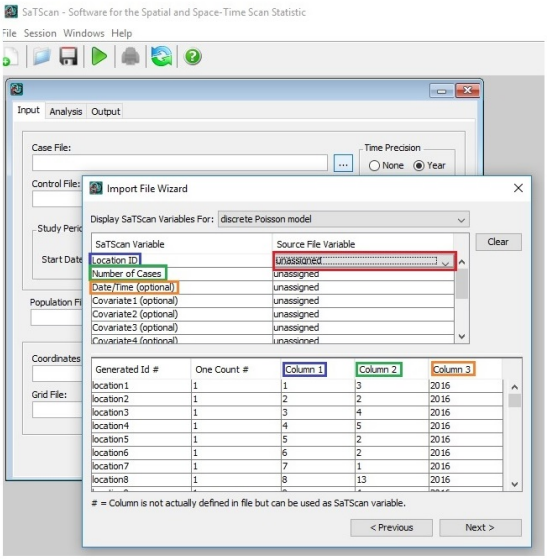
O passo-a-passo para as configurações e importação dos arquivos com as extensões exigidas pelo SaTScan serão mostradas a seguir. Ao clicar em *Next* na tela anterior o programa abrirá a seguinte tela de configuração, como mostra a Figura 5.11.

Figura 5.11 – Tela de configuração.



Na Figura 5.11, deve-se informar ao SaTScan como os dados foram separados na tabulação, ou seja, o programa precisa entender que critério foi utilizado para a separação das variáveis em colunas. Neste caso, o critério utilizado foi a tecla **Tab** e portando deve-se selecionar a opção *Whitespace* e avançar para próxima tela clicando em *Next*, como mostra a Figura 5.12.

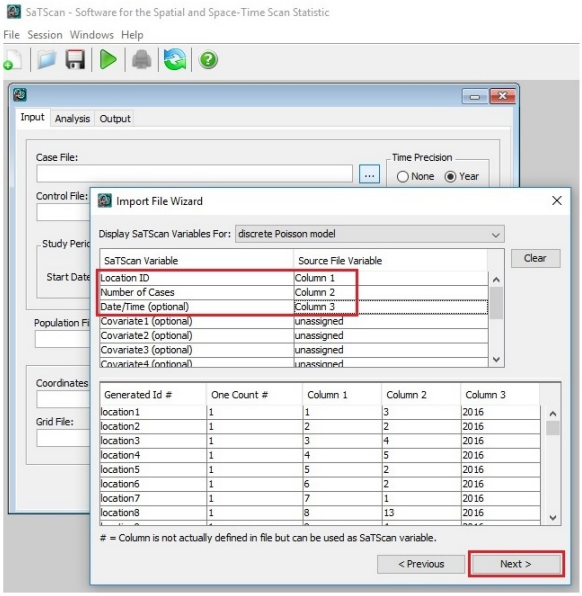
Figura 5.12 – Tela de configuração.



Observa-se na Figura 5.12 que o SaTScan criou uma tabela contendo três colunas, em que cada coluna representa uma variável. A coluna 1 representa a variável localização dos Ids, a coluna 2 representa a variável números de casos e a coluna 3 representa a variável data/hora. Nota-se que as variáveis ainda não foram associadas com suas colunas. Logo, deve-se selecionar a variável de seu arquivo de origem clicando em *unassigned*, fazendo

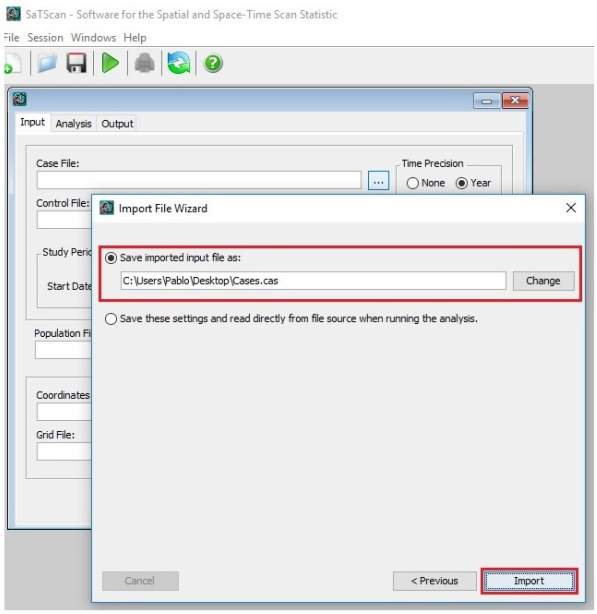
assim a associação de cada coluna com sua respectiva variável. Ao fazer a associação corretamente entre variáveis e colunas, a tela do SaTScan ficará como mostra a Figura 5.13.

Figura 5.13 – Tela de configuração.



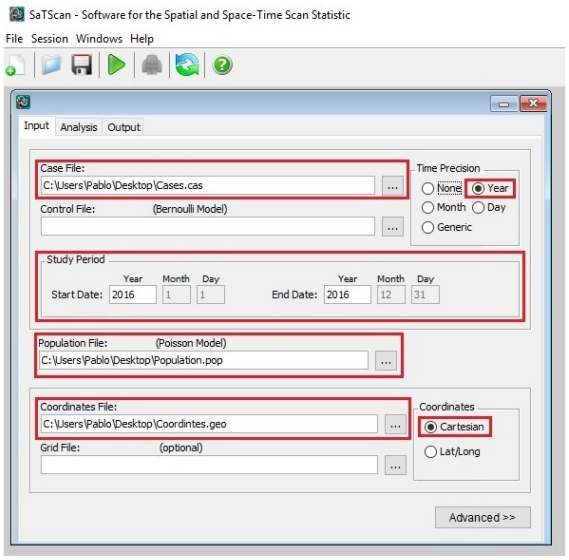
Ao clicar em *Next* na Figura 5.13, aparecerá a última tela de importação dos dados. Nela é selecionado o local do computador onde o arquivo dos casos com extensão *.cas* será salvo. Nesse estudo, os dados foram salvos na área de trabalho do computador, como mostra a Figura 5.14.

Figura 5.14 – Tela de configuração.



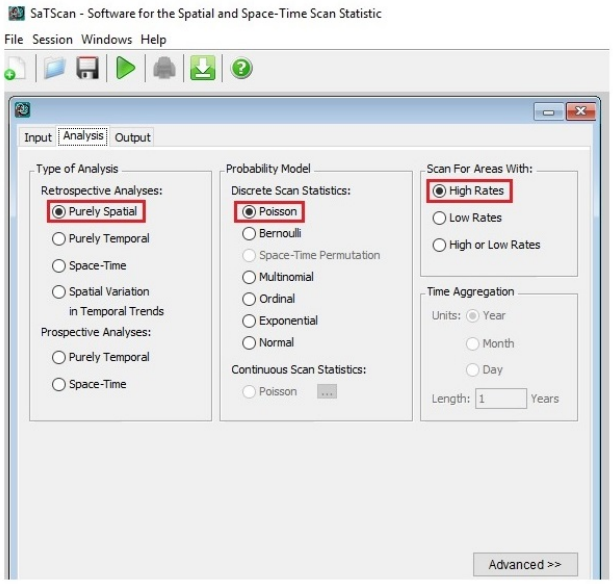
Ao clicar em *Import* na Figura 5.14, o SaTScan criará um novo arquivo de dados chamado **Cases**, com a extensão *.cas*. Esse novo arquivo conterá as mesmas informações do arquivo de dados **casos**, que possui extensão *.txt*. De forma análoga, o processo deverá ser repetido para os arquivos de dados da população **pop** e o de coordenadas **coor**. Ao fim do processo, o SaTScan deverá ter importado os seguintes arquivos de dados: *Cases.cas*, *Population.pop* e *Coordintes.geo*. Os arquivos deverão ser salvos na mesma pasta destino do computador, pois caso não estejam, acarretará em erro no hora de rodar o programa. Para concluir as configurações das entradas de dados na aba *Input*, deve-se selecionar a precisão do tempo, o período do estudo e o tipo de coordenadas. Neste estudo, a unidade do tempo será em anos, o período será o ano de 2016 e as coordenadas serão do tipo cartesianas. Ao finalizar todo esse processo, a aba *Input* deverá estar configurada da seguinte maneira, como mostra a Figura 5.15.

Figura 5.15 – Aba *Input* configurada.



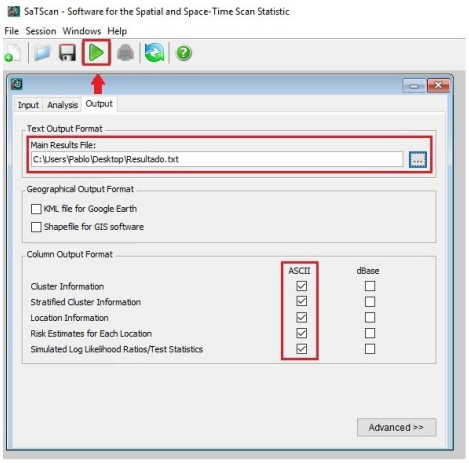
Após finalizar a configuração da aba *Input*, deve-se configurar a aba *Analysis*. Nessa aba deverão ser selecionadas três opções: o tipo de análise, a distribuição do estudo e o tipo de varredura espacial. Neste estudo estamos utilizando uma análise puramente espacial, logo deve-se marcar a opção *Purely Spatial*. O modelo de probabilidade é o Poisson, logo deve-se marcar a opção *Poisson*. A varredura espacial será do tipo *High Rates*, que significa varredura espacial em locais onde ocorrem altas taxas de incidência da doença. Ao finalizar as configurações da aba *Analysis*, ela deverá estar da seguinte maneira, como mostra a Figura 5.16.

Figura 5.16 – Aba *Analysis* configurada.



Na aba *Output*, deve-se selecionar o arquivo de saída de dados em branco que foi criado anteriormente, denominado de **Resultado**. Para que não ocorra erro é importante que o arquivo **Resultado** esteja salvo no mesmo local onde os arquivos *Cases.cas*, *Population.pop* e *Coordintes.geo* foram salvos no computador. Ainda na aba *Output*, deve-se selecionar quais informações irão sair no resultado do teste e em que formato deverão sair os resultados. O SaTScan disponibiliza dois tipos de formatos de saída de resultado, o do tipo ASCII e o do tipo dBase. Para esse estudo, escolhemos apenas a saída do tipo ASCII. Ao finalizar essas configurações, deve-se clicar no botão *Play* indicado pela seta vermelha, como mostra a Figura 5.17.

Figura 5.17 – Aba *Output* configurada.



Ao clicar no botão *Play*, a análise da estatística Scan será realizada e seu resultado poderá ser visto na forma descrita a seguir:

```
Purely Spatial analysis
scanning for clusters with high rates
using the Discrete Poisson model.
------------------------------------------------------------------------

SUMMARY OF DATA

Study period.....................: 2016/1/1 to 2016/12/31
Number of locations..............: 51
Total population.................: 407757
Total number of cases............: 430
Annual cases / 100000............: 105.2
------------------------------------------------------------------------
CLUSTERS DETECTED

1.Location IDs included.: 23, 2, 9, 26, 45, 5, 13, 33, 12, 28, 31, 25,
20, 35, 37, 24, 30, 46, 18, 32, 10, 7, 15, 3, 51, 41
Overlap with clusters.: 2, 3, 4, 5, 6
Coordinates / radius..: (9.20238e+006,182855) / 3953.60
Gini Cluster.........: No
Population...........: 166656
Number of cases.......: 277
Expected cases........: 175.75
Annual cases / 100000.: 165.9
Observed / expected...: 1.58
Relative risk.........: 2.62
Log likelihood ratio..: 48.319769
P-value..............: < 0.00000000000000001
```

A interpretação e discussão dos resultados da Estatística Scan realizado no programa SaTScan (KULLDORFF, 2016) será feita após a descrição das análises no programa R (R Core Team, 2017).

5.2.2 Análise no programa R

Para a análise no programa R, primeiramente deve-se requerer o pacote *smerc*. A estatística Scan é calculada pela função *scan.test*, com os argumentos *coords*, *cases*, *pop*, *nsim*, *alpha* e *lonlat*. As linhas de comando para realização do teste Estatística Scan estão descritas a seguir:

```
> require(smerc)
> EstScan = scan.test(coords = coordinates(ShapeCG), cases = floor
+                     (ShapeCG$Casos), pop = ShapeCG$Pop, nsim = 999,
+                     alpha = 0.05, lonlat = FALSE)
> EstScan

$clusters
$clusters[[1]]
$clusters[[1]]$locids
[1] 23  2  9 26 45  5 13 33 12 28 31 25 20 35 37 24 30 46 18
[20] 32 10  7 15  3 51 41

$clusters[[1]]$coords
[,1]    [,2]
23 182854.9 9202380

$clusters[[1]]$r
[1] 3953.161

$clusters[[1]]$pop
[1] 166656

$clusters[[1]]$cases
[1] 277

$clusters[[1]]$expected
[1] 175.747

$clusters[[1]]$smr
[1] 1.576129

$clusters[[1]]$rr
[1] 2.619186

$clusters[[1]]$loglikrat
[1] 48.31977

$clusters[[1]]$pvalue
[1] 0.001
```

O significado dos argumentos da função *scan.test* são descritos a seguir:

- *coords*: esse argumento é utilizado para informar as coordenadas dos centroides das áreas no *shapefile*, no R. Seleciona-se as coordenadas de um *shapefile* pela função *coordinates()*;
- *cases*: esse argumento define o número de casos. Esses casos podem ser filtrados do arquivo *shapefile* pela função *ShapeCG$Casos*;
- *pop*: esse argumento indica a população das áreas do *shapefile*. Essa população pode ser filtrada pela função *ShapeCG$Pop*;
- *nsim*: Define o número de simulações feitas no cálculo valor-p da estatística;
- *alpha*: representa o nível de significância para determinar se um *cluster* é significativo. Por padrão é utilizado o valor 5% de significância;
- *lonlat*: esse argumento define o tipo de coordenadas que o teste usará. Caso seja atribuído o valor *TRUE*, o teste utilizará as coordenadas latitude/longitude. Caso seja atribuído o valor *FALSE*, o teste utilizará coordenadas cartesianas.

Os resultados utilizando o programa R (R Core Team, 2017), conforme esperado, são os mesmos de quando foi utilizado o programa SaTScan (KULLDORFF, 2016).

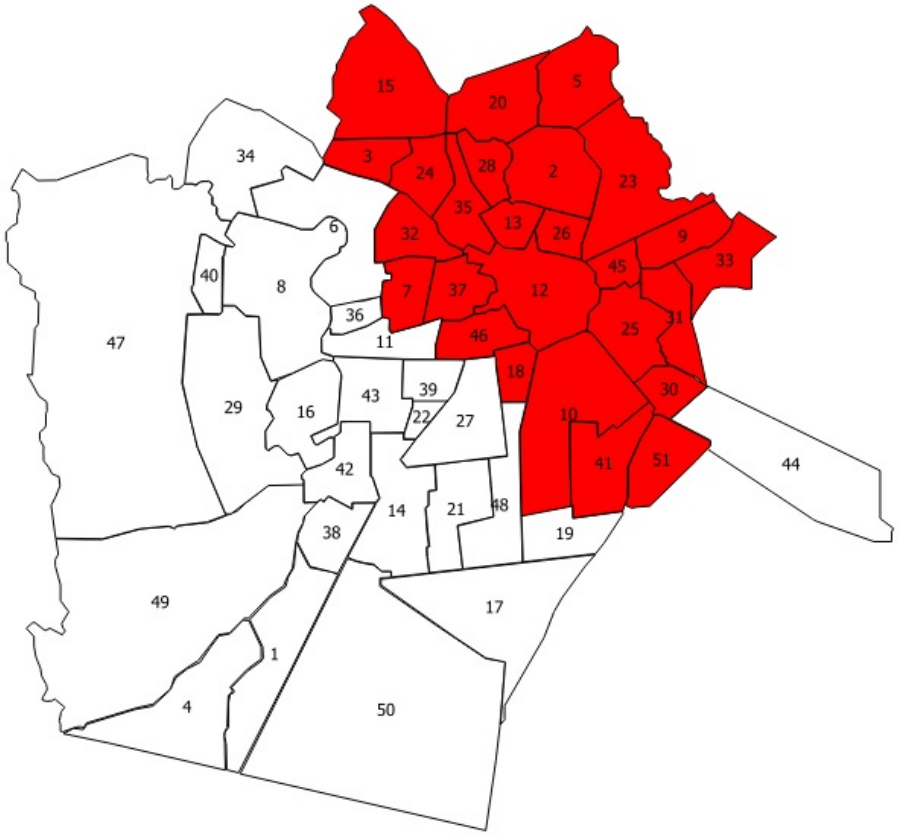
5.2.3 Conclusão do teste Estatística Scan

O teste Estatística Scan apresentou como resultado um total de 26 bairros com epidemia de dengue na cidade de Campina Grande - PB. Esses bairros foram identificados por cada programa pelos seguintes Ids:

Programa	**Bairros pertencentes ao cluster observado**
SaTScan	23 2 9 26 45 5 13 33 12 28 31 25 20 35 37 24 30 46 18 32 10 7 15 3 51 41
R	23 2 9 26 45 5 13 33 12 28 31 25 20 35 37 24 30 46 18 32 10 7 15 3 51 41

Esse resultado demostra que o teste *scan.teste*, do pacote *smerc*, desenvolvido por Joshua French no programa R, se mostrou bastante eficiente em comparação ao programa SaTScan, já que o SaTScan foi o pioneiro e é o programa mais utilizado na literatura nesse tipo de análise. Para uma melhor compreensão do estudo, mostraremos como se comporta a distribuição espacial do *cluster* observado em forma de mapa, como mostra a Figura 5.18.

Figura 5.18 – Distribuição espacial do *cluster* na cidade de Campina Grande - PB.



Destaca-se em vermelho os bairros com epidemias de dengue na cidade de Campina Grande - PB no ano de 2016. As taxas de incidência de casos de dengue nos bairros estão disponíveis no Anexo B deste trabalho. Pode-se concluir pelo teste Estatística Scan que os bairros com maior risco de dengue na cidade de Campina Grande - PB em 2016 foram identificadas nas zonas norte e leste da cidade, que são descritas na tabela a seguir:

Id	Bairro	Id	Bairro
23	Jardim Tavares	35	Palmeira
2	Alto Branco	37	Prata
9	Castelo Branco	24	Jeremias
26	Lauretzen	30	Mirante
45	Santo Antônio	46	São José
5	Bairro das Nações	18	Estação Velha
13	Conceição	32	Monte Santo
33	Nova Brasília	10	Catolé
12	Centro	7	Bela Vista
28	Louzeiro	15	Cuité
31	Monte Castelo	3	Araxá
25	José Pinheiro	51	Vila Cabral
20	Jardim Continental	41	Sandra Cavalcante

6 CONCLUSÃO

Neste trabalho utilizou-se os conceitos relacionados à análise de dados de área e estatística Scan com intuito de fazer um material didático que possa ser útil para pesquisadores, professores e estudantes de áreas aplicadas que necessitam realizar uma análise básica. Para isso, foi disponibilizado o passo-a-passo de como proceder com essas análises utilizando os programas R e SatScan, sendo explicado o significado de todas das funções utilizadas. Além disso, foram feitas as interpretações e discussões dos resultados.

No estudo de análise de dados de área pode-se concluir que houve a presença de autocorrelação global sobre a taxa de incidência de dengue na cidade de Campina Grande-PB pelos índices I de Moran, estatística c de Geary e a estatística EBI. Esses índices foram significativos tanto pelo teste por normalidade como por permutação. A estatística G Getis-Ord apresentou ausência de autocorrelação global na área de estudo. Em relação as estatísticas locais, o índice de autocorrelação local (LISA) detectou dependência espacial positiva nos bairros do Cuité e Louzeiro, localizados na zona norte da cidade. Portanto, deve-se ter uma atenção especial por parte das secretárias de saúde e órgãos competentes ao combate do mosquito da dengue nesse bairros e nos bairros que compartilham fronteira física com eles.

Em relação a estatística Scan, o teste conseguiu identificar um total de 26 áreas de risco na cidade de Campina Grande - PB. Comparado ao LISA, a estatística Scan mostrou-se, para esse estudo, uma ferramenta mais sensível na identificação de potenciais áreas de risco de contaminação de dengue. Para trabalhos futuros, pode-se aprofundar no estudo de suas metodologias, com o objetivo de explicar suas divergências na captação de áreas de risco.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, E. **Econometria Espacial Aplicada**. 1. ed. SP: Editora Alívia, 2012. v. 1.

ANSELIN, L. Local indicators of spatial association–lisa. **Geographical analysis**, v. 27, n. 2, p. 93–115, 1995.

ASSUNÇÃO, R. M.; COSTA, M. A. Uma análise de desempenho dos métodos scan e besag&newell na detecção de clusters espaciais. **Revista Brasileira de Estatística**, 2004.

ASSUNÇÃO, R. M.; REIS, E. A. A new proposal to adjust moran's i for population density. **Statistics in Medicine**, v. 18, p. 2147–2162, 1999.

BAILEY, T. C.; GATRELL, A. C. **Interactive Spatial Data Analysis**. 1. ed. New York: Routledge, 1995.

BALIEIRO, A. A. da S. **Detecção de conglomerados dos alertas de desmatamentono estado do amazonas usando estatística de varredura espaço-temporal**. Dissertação (Mestrado) — Universidade Federal de Viçosa, 2008.

BIVAND, R. spdep spatial dependence: Weighting schemes, statistics and models. **R package version 0.6-15**, 2017.

BIVAND, R.; LEWIN-KOH, N. maptools: Tools for reading and handling spatial objects. **R package version 0.9-2**, 2017.

CÂMARA, G. et al. **Análise Espacial de Dados Geográficos**. Brasília: EMBRAPA, 2004.

CARVALHO, M. S. et al. **Introdução à Estatística Espacial para a Saúde Pública**. Brasília-DF: Ministério da Saúde e Fundação Oswaldo Cruz, 2007. v. 3.

CRESSIE, N. A. C. **Statistics for Spatial Data**. Revised edition. Iowa State University: A Wiley-Interscience Publication, 1993.

DIGGLE, P. J. **Statistical Analysis of Spatial and Spatio-Temporal Point Patterns**. Third. Lancaster University England, UK: A CHAPMAN & HALL BOOK, 2013.

FISCHER, M. M.; WANG, J. **Spatial data analysis: models, methods and techniques**. London: Springer Science & Business Media, 2011.

FOTHERINGHAM, A. S.; ROGERSON, P. A. **The SAGE handbook of spatial analysis**. New York: Sage, 2008.

FRENCH, J. **smerc: Statistical Methods for Regional Counts**. EUA, 2015. R package version 0.2.2. Disponível em: <https://CRAN.R-project.org/package=smerc>.

GEARY, R. C. The contiguity ratio and statistical mapping. **The incorporated statistician**, v. 5, n. 3, p. 115–146, 1954.

GETIS, A.; ORD, J. K. The analysis of spatial association by use of distance statistics. **Geographical Analysis**, v. 24, p. 189–206, 1992.

GETIS, A.; ORD, J. K. Local spatial autocorrelation statistics: Distributional issues and an application. **Geographical Analysis**, v. 27, p. 286–306, 1995.

HAINING, R. Spatial autocorrelation problems. **Geography and the Urban Environment**, v. 3, p. 1–43, 1980.

HAINING, R. **Spatial Data Analysis: Theory and Practice**. New York: Cambridge University Press, 2003.

JÚNIOR, I. S. **A influência da urbanização no clima da cidade de Campina Grande-pb**. Dissertação (Mestrado) — Universidade Federal de Campina Grande, 2006.

KULLDORFF, M. A spatial scan statistic. **Communication in Statistics - Theory and Methods**, v. 26, p. 1481–1496, 1997.

KULLDORFF, M. **SaTScanTM Manual do Usuário, versão 9.4**. EUA, 2016. Disponível em: <https://www.satscan.org/SaTScan_TM_Manual_do_Usu%C3%A1rio_v9.4_Portugues.pdf>.

KULLDORFF, M. et al. An elliptic spatial scan statistic. **Statistics in Medicine**, v. 25, p. 3929–3943, 2006.

KULLDORFF, M.; NAGARWALLA, N. Spatial disease clusters: detection and inference. **Statistics in Medicine**, v. 16, p. 779–810, 1995.

LIMA, M. S. de. **Métodos Adaptativos para Detecção de Clusters no Espaço-tempo**. 104 p. Tese (Doutorado em Estatística) — Universidade Federal de Minas Gerais-UFMG, 2011.

MANUEL, L. **Modelos de regressão Linear com efeitos espaciais na Análise da mortalidade infantil**. 82 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) — Universiadade Federal de Lavras, 2011.

MARCONATO, M. et al. Análise espacial da taxa de pobreza e da população rural da região sul do país. **Textos de Economia**, v. 18, n. 2, p. 16–40, 2017.

MINISTÉRIO DA SAÚDE. **Boletim epidemiológico de 1 a 52**. Brasília, 2017. Disponível em: <http://combateaedes.saude.gov.br/pt/situacao-epidemiologica>.

MORAN, P. A. P. The interpretation of statistical maps. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 10, n. 2, p. 243–251, 1948. Disponível em: <http://www.jstor.org/stable/2983777>.

ODEN, N. Adjusting moran's i for population density. **Statistics in Medicine**, v. 14, p. 17–26, 1995.

OLIVEIRA, D. R. X. de. **O Problema de Detecção de Clusters Espaciais Irregulares: Uma Nova Abordagem Multiobjetivo**. Dissertação (Mestrado) — Universidade Federal de Ouro Preto, 2017.

PEBESMA, E. J.; BIVAND, R. S. Classes and methods for spatial data in R. **R News**, v. 5, n. 2, p. 9–13, November 2005. Disponível em: <https://CRAN.R-project.org/doc/Rnews/>.

PFEIFFER, D. U. et al. **Spatial Analysis in Epidemiology**. New York: Oxford University Press, 2008.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>.

RIBEIRO, E. C. A.; ALMEIDA, E. S. Convergência local para municípios brasileiros. **IX Encontro Nacional da Associação Brasileira de Estudos Regionais e Urbanos (ENABER)**, 2011.

ROGERSON, P.; YAMADA, I. **Statistica Detection And Surveillance of Geographic Clusters**. Boca Raton: CHAPMAN & HALL, 2008.

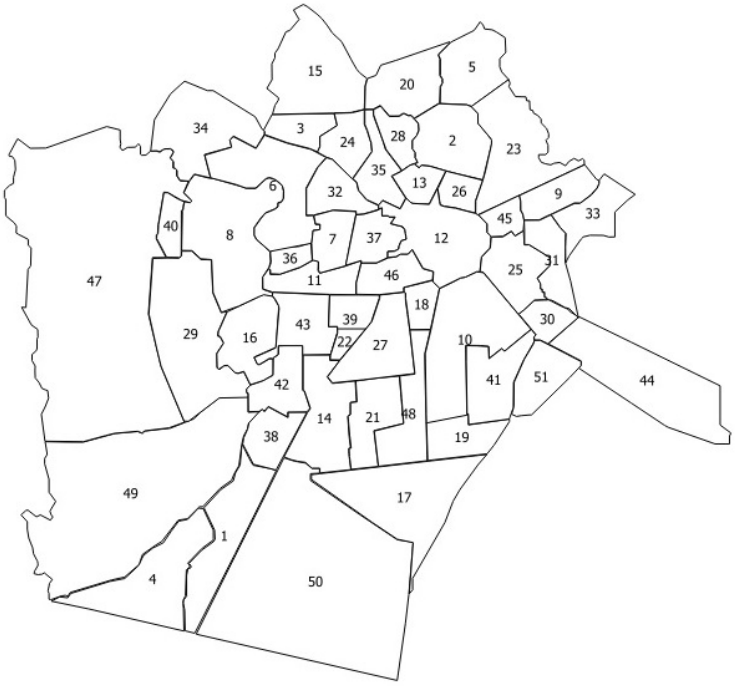
ROMERO, J. A. R. Analise espacial da pobreza municipal no estado de minas gerais - 1991 - 2000. **Anais do XV Encontro Nacional de Estudos Populacionais**, 2006.

SILVA, N. C. N. da. **Análise de dados de área aplicada a dois indicadores Econômicos de mesorregião do estado de Minas Gerais**. Dissertação (Mestrado) — Univerdidade Federal de Lavras, 2010.

UPTON, G. J. G.; FINGLETON, B. **Spatial Data Analysis by Example: Point Pattern and Quantitative Data**. New York: John Wiley & Sons, 1985. v. 1. (Wiley series in probability and statistics, v. 1).

WALLER, L. A.; GOTWAY, C. A. **Applied Spatial Statistics for Public Health Data**. New Jersey: John Wiley & Sons, 2004.

ANEXO A – Identificacão dos bairros da cidade de Campina Grande-PB



ID	Bairro	ID	Bairro
1	Acácio Figueiredo	27	Liberdade
2	Alto Branco	28	Louzeiro
3	Araxá	29	Malvinas
4	Bairro das Cidades	30	Mirante
5	Bairro das Nações	31	Monte Castelo
6	Bairro Universitário	32	Monte Santo
7	Bela Vista	33	Nova Brasília
8	Bodocongó	34	Novo Bodocongó
9	Castelo Branco	35	Palmeira
10	Catolé	36	Pedregal
11	Centenário	37	Prata
12	Centro	38	Presidente Médice
13	Conceição	39	Quarenta
14	Cruzeiro	40	Ramadinha
15	Cuité	41	Sandra Cavalcante
16	Dinamérica	42	Santa Cruz
17	Distrito Industrial	43	Santa Rosa
18	Estação Velha	44	Santa Terezinha
19	Itararé	45	Santo Antônio
20	Jardim Continental	46	São José
21	Jardim Paulistano	47	Serrotão
22	Jardim Quarenta	48	Tambor
23	Jardim Tavares	49	Três Irmãs
24	Jeremias	50	Velame
25	José Pinheiro	51	Vila Cabral
26	Lauritzen		

ANEXO B – Banco de Dados

Tabela 2 – Banco de dados dos casos de dengue Campina Grande-PB ano de 2016.

Id	Bairros	Casos	População	Incidência/1000 hab.
1	Acácio Figueiredo	3	9814	0,306
2	Alto Branco	2	9339	0,214
3	Araxá	4	1848	2,165
4	Bairro das Cidades	5	6376	0,784
5	Bairro das Nações	2	1484	1,348
6	Bairro Universitário	2	3938	0,508
7	Bela Vista	1	6406	0,156
8	Bodocongó	13	14550	0,893
9	Castelo Branco	4	3055	1,309
10	Catolé	30	20635	1,454
11	Centenário	6	8760	0,685
12	Centro	41	7943	5,162
13	Conceição	3	3629	0,827
14	Cruzeiro	12	14796	0,811
15	Cuités	16	2030	7,882
16	Dinamérica	1	5782	0,173
17	Distrito Industrial	1	2658	0,376
18	Estação Velha	9	3496	2,574
19	Itararé	3	3302	0,909
20	Jardim Continental	5	2393	2,089
21	Jardim Paulistano	9	8471	1,062
22	Jardim Quarenta	1	2941	0,340
23	Jardim Tavares	2	3682	0,543
24	Jeremias	28	11217	2,496
25	José Pinheiro	38	17003	2,235
26	Lauritzen	5	2863	1,746
27	Liberdade	19	16711	1,137
28	Louzeiro	5	1388	3,602
29	Malvinas	20	43323	0,049
30	Mirante	17	1891	8,990
31	Monte Castelo	5	8883	0,563
32	Monte Santo	13	8020	1,621
33	Nova Brasília	5	9905	0,505
34	Novo Bodocongó	5	1618	3,090

Id	Bairros	Casos	População	Incidência/1000 hab.
35	Palmeira	11	6006	1,832
36	Pedregal	7	8913	0,785
37	Prata	4	3771	1,061
38	Presidente Médice	3	4536	0,661
39	Quarenta	5	8213	0,609
40	Ramadinha	3	2290	1,310
41	Sandra Cavalcante	13	6877	1,890
42	Santa Cruz	9	9935	0,906
43	Santa Rosa	6	11328	0,530
44	Santa Terezinha	4	6812	0,587
45	Santo Antônio	4	4149	0,964
46	São josé	3	4168	0,720
47	Serrotão	1	7293	0,137
48	Tambor	4	8207	0,487
49	Três Irmãs	7	12884	0,543
50	Velame	4	6370	0,628
51	Vila Cabral	7	5071	1,380

ANEXO C – Rotina de programação no R

```
## Rotina para Análise de Dados de Área

## Limpar a memória
rm(list=ls())

## Mudar diretório
setwd ("C:/Users/Pablo/Desktop/Analises")

## Verificar o diretório
getwd()

## Requerino pacotes
require(maptools)
gpclibPermit()
require(sp)
require(spdep)
require(classInt)
require(RColorBrewer)

par.ori <- par(no.readonly=TRUE)

## Importando mapa de Campina Grande - PB
ShapeCG<- readShapePoly("Mapa_Campina_Grande.shp",IDvar="Id")
plot(ShapeCG, axes=F)
title("Bairros da cidade de Campina Grande")
head(ShapeCG@data,51)

## Reconhecendo os nomes dos bairros como caracter
ShapeCG@data$Nome <- as.character(ShapeCG@data$Nome)
ShapeCG@data$Nome

## Importanto arquivo com dados dos casos
dengue <- read.table("Dengue2016.txt",header=TRUE)
str(dengue)

## Ordenando os dados corretamente
head(ShapeCG@data)
head(dengue)
ind <- match(ShapeCG@data$Id, dengue$Id)
ind
```

```
## Concatenando o shapeCG com os casos de dengue
dengue <- dengue[ind,]
row.names(dengue) <- ShapeCG$Id
ShapeCG <- spCbind(ShapeCG, dengue)
head(ShapeCG@data)

## Visualizando um mapa do atributo: Quantis
INT1 <- classIntervals(ShapeCG$Incid, n=4, style="quantile")
CORES.1 <- c(rev(brewer.pal(3, "Blues")), brewer.pal(3, "Reds"))[-1]
COL1 <- findColours(INT1, CORES.1)
plot(ShapeCG, col=COL1)
TB1 <- attr(COL1, "table")
legtext <- paste(names(TB1))
legend("bottomright", fill=attr(COL1, "palette"), legend=legtext,
bty="n",cex=0.8)

## Calculando a Matriz de Vizinhanças
ccods = coordinates(ShapeCG)
points = cbind(ccods[,1],ccods[,2])
dnb = dnearneigh(points,0,1807)
W.Bin= nb2mat(neighbours = dnb, style = "B")
W.Normal= nb2mat(neighbours = dnb, style = "W")

## Calculando o Índice de Moran Global

# Pelo teste de Normalidade
moran.test(ShapeCG$Incid,listw=nb2listw(dnb, style ="W"),
           randomisation= FALSE)

# Pelo teste de Permutação
moran.test(ShapeCG$Incid,listw=nb2listw(dnb, style ="W"),
           randomisation= TRUE)

# Por simulação de Monte-Carlo
moran.mc(ShapeCG$Incid, listw=nb2listw(dnb, style ="W"), nsim=999)

## Calculando Emperical Bayes Index - EBI

# Pelo teste de Permutação
EBImoran.mc(ShapeCG$Casos,ShapeCG$Pop, nb2listw(dnb, style="W"),
nsim=999)

# Por simulação de Monte-Carlo
shapeCG.p=ShapeCG$Casos/ShapeCG$Pop
moran.mc(shapeCG.p, nb2listw(dnb, style="W", zero.policy=TRUE),
nsim=999, zero.policy=TRUE)
```

```
## Calculando a Estatística c de Geary Global

# Pelo teste de Normalidade
geary.test(ShapeCG$Incid, listw=nb2listw(dnb, style = "W"),
           randomisation= FALSE)

# Pelo teste de Permutação
geary.test(ShapeCG$Incid, listw=nb2listw(dnb, style = "W"),
           randomisation=TRUE)

# Por simulação de Monte-Carlo
geary.mc(ShapeCG$Incid, listw=nb2listw(dnb, style = "W"),nsim=999)

## Calculando Índice de Getis e Ord Global
globalG.test(ShapeCG$Incid, nb2listw(dnb,style="B"))

## Getis e Ord Local
localG(ShapeCG$Incid, nb2listw(dnb,style="B"), zero.policy=NULL,
spChk=NULL, return_internals=FALSE)

## Moran Local
ShapePB.mloc <- localmoran(ShapeCG$Incid, listw=nb2listw
(dnb, style="W"))
head(ShapePB.mloc)

## Mapa das probabilidades (Significâncias do I de Moral Local)
INT4 <- classIntervals(ShapePB.mloc[,5], style="fixed",
fixedBreaks=c(0,0.01, 0.05, 0.10))
CORES.4 <- c(rev(brewer.pal(3, "Reds")), brewer.pal(3, "Blues"))
COL4 <- findColours(INT4, CORES.4)
plot(ShapeCG, col=COL4)
TB4 <- attr(COL4, "table")
legtext <- paste(names(TB4))
legend("bottomright", fill=attr(COL4, "palette"), legend=legtext,
bty="n", cex=0.7, y.inter=0.7)

## LISA Map (+/+) , (-,-), (+,-), (-,+)

## Montando matrix W de vizinhança
ShapeCG.nb1.mat <- nb2mat(dnb)

## Incidência de dengue Padronizado
Dengue_SD <- scale(ShapeCG$Incid)
```

```
## Média das incidência de dengue Padronizada nos visinhos
Dengue_W <- ShapeCG.nb1.mat %*% Dengue_SD

## Diagrama de espalhamento de Moran
plot(Dengue_SD, Dengue_W,xlab="Z",ylab="WZ")
abline(v=0, h=0)
title("Diagrama de Espalhamento de Moran")

## Montando o Mapa LISA
Q <- vector(mode = "numeric", length = nrow(ShapePB.mloc))
Q[(Dengue_SD>0  & Dengue_W > 0)] <- 1
Q[(Dengue_SD<0  & Dengue_W < 0)] <- 2
Q[(Dengue_SD>=0 & Dengue_W < 0)] <- 3
Q[(Dengue_SD<0  & Dengue_W >= 0)]<- 4
signif=0.05

Q[ShapePB.mloc[,5]>signif]<-5

CORES.5 <- c("blue", "green" , "red", "yellow", "gray",
           rgb(0.95,0.95,0.95))
plot(ShapeCG, col=CORES.5[Q])
title("Mapa LISA")
legend("bottomright", c("Q1(+/+)", "Q2(-/-)",
 "Q3(+/-)", "Q4(-/+)","NS"), fill=CORES.5)

## Rotina para Análise da Estatística Scan

require(smerc)

EstScan = scan.test(coords = coordinates(ShapeCG), cases = floor
                    (ShapeCG$Casos), pop = ShapeCG$Pop, nsim = 999,
                    alpha = 0.05, lonlat = FALSE)
EstScan
```