TÉCNICAS DE ANÁLISE DE DADOS DISTRIBUÍDOS EM ÁREAS

Jane Maiara Bertolla

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Fevereiro - 2015

TÉCNICAS DE ANÁLISE DE DADOS DISTRIBUÍDOS EM ÁREAS

Jane Maiara Bertolla

Orientador: Prof. Dr. **José Silvio Govone**

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Fevereiro - 2015

Dedicatória

Aos meus pais pelo amor incondicional.

Agradecimentos

Primeiramente, agradeço a Deus por nunca ter deixado eu perder a fé, a força e o foco.

Aos meus pais Carlos e Conceição, que sempre me apoiaram, incentivaram e compreenderam as minhas escolhas.

Aos meus irmãos Aguinaldo e Mara pela amizade e amor.

Ao meu melhor amigo Pedro que sempre esteve presente nos bons e maus momentos da minha vida.

Ao meu orientador professor Dr. José Silvio Govone pelo auxilio e incentivo no desenvolvimento desta dissertação.

À professora Dra. Maria da Conceição F. F. Tandel que foi minha orientadora durante a graduação e dispertou me o interesse pela Estatística.

Aos colegas de mestrado: Edimar, Luiz, Maria Cecília, Rafael, Sophia e Thomas pelo convivio e troca de experiências.

Aos colegas: Airton, Elaine, Fabiano, Maelson e Raimundo que compreenderam as minhas dificuldades e cederam moradia em Botucatu.

À minha colega Jaqueline que me recebeu, de volta, em Rio Claro.

Aos professores e funcionários do Programa de Pós Graduação em Biometria - UNESP - Botucatu pela disponibilidade e amizade.

Ao técnico do laboratório do DEMAC - UNESP - Rio Claro,Jorge Gustavo Falcão, pelo auxilio na parte computacional deste trabalho.

À CAPES pelo apoio financeiro para o desenvolvimento desta dissertação.

Sumário

Lista de Figuras

Página

Lista de Tabelas

Página

TÉCNICAS DE ANÁLISE DE DADOS DISTRIBUÍDOS EM ÁREAS

Autora: JANE MAIARA BERTOLLA
Orientador: Prof. Dr. JOSÉ SILVIO GOVONE

RESUMO

O objetivo deste trabalho é estudar técnicas de análise espacial para compreender os padrões associados a dados por área, testar se o padrão observado é aleatório ou se o evento se distribui por aglomerado, obter mapas mais suaves que o mapa observado e procurar estimativas melhores de estruturas adjacentes.

Utilizou-se um conjunto de dados referente a 1656 casos positivos de dengue na cidade de Rio Claro - SP registrados no primeiro semestre de 2011. Neste conjunto de dados, e com o auxílio do software Terra View 4.2.2, foram construídas as estimativas de Kernel para dois tipos de função de estimação: Kernel Normal e Kernel Quártico. Para a Função de Estimação Kernel Normal foram usados os seguintes raios de influência: 100m, 150m, 200m e 500m. Já para a Função de Estimação Kernel Quártico foram usados os seguintes raios de influência: 250m, 375m, 500m e 625m. Ainda no mesmo software foram construídos mapas para uma análise exploratória com três critérios diferentes: Intervalos iguais, Intervalos por

quantil e Intervalos por desvios padrão. Em seguida, foram construídos os respectivos mapas de suavização através da Média Móvel Espacial.

Dependendo do critério utilizado (quantil, intervalos iguais ou desvio padrão), notou-se que há diferenças nos resultados de ocorrências de dengue, tanto para os dados originais, quanto para os transformados pela média móvel.

Observou-se que o comportamento do kernel quártico é similar ao do kernel normal, porém com diferentes raios de influência. Este resultado corrobora as observações de Bailey e Gatrell(1995), de que a função de ajuste não é de grande importância, já que o controle pode ser feito através do raio de influência para a estimativa em cada ponto.

Através do teste de permutação aleatório verificou-se que há dependência espacial entre os valores observados, onde a estatística I é igual a 0,389828 e o p-valor igual a 0,01.

Kawamoto(2012) aplicou a estimação de kernel para o mesmo conjunto de dados, porém considerou os dados como dados por ponto e não como dados por área, isto é, não particionou a região de estudo. Utilizou a estimação de Kernel normal e quártico.

Compararam-se os resultados obtidos por Kawamoto (2012) com os obtidos neste trabalho e notou-se uma semelhança entre eles quando consideram-se raios maiores. As diferenças devem-se ao fato que Kawamoto(2012) trabalhou com processos pontuais, e quando aplica-se a estimação de Kernel para dados por ponto, o ponto da região de estudo a ser estimado é um ponto genérico enquanto que, para dados por área, o ponto a ser estimado é o centroide de cada bairro.

Portanto, a técnica de ajuste aplicada a dados espacialmente distribuídos em área pode ser usada, com sucesso, para mapear ocorrências de dados espacialmente distribuídos, onde a região de ocorrência da variável é subdividida em áreas, por questões físicas, geográficas, econômicas, etc., e o interesse do pesquisador é verificar o comportamento da variável nas diferentes áreas.

Palavras-chave: Análise Espacial; Dados por área, Estimação Kernel,

ANALYSIS TECHNIQUES OF DISTRIBUTED DATA OVER AREAS

Author: JANE MAIARA BERTOLLA
Adviser: Prof. Dr. JOSÉ SILVIO GOVONE

SUMMARY

The goal of this piece is to study spacial analysis techniques to understand the patterns associated to data over areas, test if the observed pattern is random or if the event distributes itself by agglomeration, get smoother maps than the observed maps and look for better estimates of adjacent structures.

A database concerning of 1656 positive dengue occurrences in the city of Rio Claro-SP during the first semester of 2011 was used. With this database, using the software Terra View 4.2.2, it were constructed the kernel estimates for two kind of estimate functions: Normal Kernel and Quartic Kernel. For the Normal Kernel Function estimate it were used the following influence radius: 100m, 150m, 200m and 500m. On the other hand, for the Quartic Kernel Function estimate it were used the following influence radius: 250m, 375m, 500m and 625m. Yet using the same software, maps were constructed for a visual exploratory analysis with three different criterions: equal intervals, quintiles intervals and standard deviation intervals. In the

sequence, the respective smoothing maps were constructed using the Spacial Moving Averages.

Depending on the used criterion (quintiles, equal intervals or standard deviation) it was observed differences between the dengue occurrences, considering the original data or the ones transformed by the moving average.

It was observed that the quartic kernel behavior is similar to the normal's kernel, but with different influence radius. This result corroborates Bailey and Gatrell's observations that the adjustment function is not of considerable importance, considering that the control can be made through the influence radius for the estimate in each point.

Through the random permutation test it was verified that there is special dependence between the observed values, where the stats equals 0,389828 and the p-value equals 0,01.

Kawamoto (2012) applied the kernel estimate for the same database, but considering the data as punctual instead of over area data, ie, Kawamoto didn't partitionated the studied region; the normal and quartic kernel estimate was used.

The results obtained by Kawamoto (2012) and the ones obtained in this piece was compared, and a similarity was noticed between them when bigger radius were used. The differences are due to the fact that Kawamoto (2012) worked with punctual processes, and when the Kernel estimate are applied to punctual data, the studied region point to be estimated is a generic one, as long that for area data the point to be estimated is the centroid of each neighborhood.

Thus, the adjusting technique applied to spacial distributed data can be used, successfully, to map special distributed data occurrence, where the variable occurrence region is subdivided in areas, because of physical, geographical, economic or other reasons, and the researcher's interest is to verify the variable behavior in different areas.

Keywords: Spatial Analysis; Area data; Kernel Estimative; Spatial Moving Average

1 INTRODUÇÃO

A Estatística Espacial é um ramo da Estatística que permite analisar a localização espacial de ocorrências de variáveis espacialmente distribuídas. Utiliza-se de métodos estatísticos para coletar, descrever, visualizar e analisar dados de eventos espacialmente distribuídos(Assunção, 2001).

É aplicada a variáveis espacialmente distribuídas, como climatológicas, ecológicas, sociais, geológicas, agronômicas e nas áreas de epidemiologia e saúde pública.

Para visualizar a distribuição espacial de um evento utiliza-se mapas, geralmente bidimensionais, e, além da perspectiva visual da distribuição espacial do evento, é interessante identificar padrões existentes, com considerações objetivas e mensuráveis, como por exemplo: A distribuição espacial dos casos de uma doença forma um padrão do espaço? Existe associação com alguma fonte de poluição? Há evidências de contágio? Há variação nos dados ao longo do tempo? (Câmara et al, 2004).

A ênfase da Análise Espacial é mensurar propriedades e relacionamentos, levando em conta a localização espacial do evento em estudo de forma explícita. Ou seja, a característica fundamental de uma técnica de análise espacial é que a referência geográfica é utilizada explicitamente no modelo(Assunção,2001).

A Análise Espacial é composta por um conjunto de procedimentos encadeados cuja finalidade é a escolha de um modelo inferencial que considere explicitamente o relacionamento espacial presente no fenômeno. Em geral, o processo de modelagem é precedido de uma fase de análise exploratória, associada à apresentação visual dos dados sob a forma de gráficos e mapas e a identificação de padrões de

dependência espacial do fenômeno em estudo(Câmara et al, 2004).

1.1 Tipologia dos Dados em Análise Espacial

Os tipos de dados mais utilizados para caracterizar os problemas de análise espacial podem ser analisados conceitualmente em três categorias: dados pontuais, dados contínuos e dados por áreas.

1.1.1 Dados Pontuais

Os fenômenos que geram dados pontuais estão intimamente ligados à localização espacial. Os dados estão na forma de um conjunto de pontos, distribuídos dentro de uma região de estudo, determinados por estes fenômenos. Por exemplo: a localização de árvores de certa espécie ou de ninhos de pássaros em uma floresta.

Um exemplo clássico para dados pontuais é o caso de ocorrências de cólera, na cidade de Londres, em 1854. O médico inglês Jonh Snow reconheceu, através dos mapas de Londres, que os casos de cólera ocorriam em certas localizações. Ele localizou os casos encontrados no registro de óbitos, e através dos endereços das residências chegou aos poços de provisão de água contaminada nas ruas correspondentes. Se esse fato ocorresse hoje, o médico publicaria os resultados de seu estudo e fecharia o poço de Broad Street, o qual era um poço altamente contaminado, se pudesse comprovar estatisticamente que esses poços eram responsáveis pela propagação da epidemia de cólera. Essa é uma situação em que a relação espacial entre os dados contribuiu significativamente para o avanço na compreensão do fenômeno(Câmara et al, 2004).

Em dados pontuais os eventos são representados por coordenadas geográficas, sendo o interesse a localização espacial de cada evento. É de grande importância conhecer o mecanismo de geração das ocorrências, sendo que as mesmas podem se dar de forma aleatória, por aglomerados ou de maneira uniforme (Bailey & Gatrell, 1995).

Tem como finalidade estudar a própria localização espacial do evento

em análise, como a distribuição espacial dos pontos, e testar hipóteses sobre o padrão observado: se é aleatório, ou ao contrario, apresenta-se em forma de aglomerado ou regularmente distribuído.

Essa distribuição pode ser analisada e descrita pelos chamados métodos de análise de processos pontuais.

1.1.2 Dados Contínuos

Diferentemente de eventos discretos que estão associados a ocorrências pontuais, os estudos sobre dados de eventos contínuos recorrem a amostras da superfície, isto é, a valores representativos do fenômeno na área de estudo.

O objetivo é predizer valores da variável em localizações não mensuradas, bem como gerar toda uma superfície a partir das observações coletadas em alguns pontos da mesma (Bailey & Gatrell, 1995).

Fenômenos naturais, tais como temperatura, pressão atmosférica, características de solo, etc., em princípio podem ser observados e medidos em qualquer ponto da área em estudo. Esses tipos de dados são os mais utilizados em variáveis ambientais.

A geoestatística consiste num conjunto de técnicas que descrevem a continuidade espacial das variáveis, e possui ferramentas para analisar as suas variações.

As técnicas mais usadas para análise espacial desse tipo de dados são voltadas para a predição espacial, ou seja, deseja se estimar valores da variável em locais onde não houve medida a partir das mensurações feitas. Uma das técnicas mais usadas para a predição é chamada de Krigagem.

1.1.3 Dados por Área

Dados por área são aqueles cuja localização está associada a áreas delimitadas por polígonos, onde não necessariamente se tem a localização exata do evento, mas um valor que represente toda a área. A divisão geográfica pode se dar

por meio de divisões administrativas ou por meio de divisões naturais como relevo, rios, zonas climáticas, etc.

O valor da variável pode ser, por exemplo, taxas de natalidade e mortalidade de determinada região, proporção de analfabetos nos municípios de um estado, etc.

Cada valor não representa uma posição específica dentro de uma área, mas sim representa toda a área. Muitas vezes necessita-se de um ponto de referência dentro da área para localizar o valor. Geralmente escolhe-se o centroide da área.

Estuda-se a aleatoriedade da distribuição espacial da variável, bem como, compara-se as frequências de ocorrências, nas diversas áreas de uma região.

Desse modo, as análises sobre esses dados são imprescindíveis nas áreas de saúde, em estudos demográficos, em atividades políticas e outras afins.

O objetivo deste trabalho é estudar técnicas de análise espacial para compreender os padrões associados a dados de área, testar se o padrão observado é aleatório ou o evento se distribui por aglomerados, obter mapas mais suaves que o mapa observado e procurar estimativas melhores de estruturas adjacentes.

2 REVISÃO DE LITERATURA

2.1 Conceitos Básicos em Análise Espacial

2.1.1 Dependência Espacial

Um conceito importante na compreensão e na análise das variáveis distribuídas espacialmente é a dependência espacial. Segundo Cressie (1993), dados que estão mais próximos no espaço são mais parecidos que aqueles que estão mais distantes. Ou seja, os valores observados da variável aleatória mais próximos tendem a ser mais correlacionados e não podem ser tratados como independentes.

Basicamente, tais variáveis produzem valores espacialmente correlacionados, de acordo com a proximidade entre eles, de maneira que, quanto menor a distância entre dois pontos, maior a correlação entre os valores nestes pontos.

No caso de ocorrer dependência espacial,as inferências estatísticas não serão eficientes quanto nos casos de amostras independentes. Este fato ocasiona em variâncias maiores para as estimativas e um ajuste pior para os modelos estimados (Câmara et al, 2004).

Neste caso, considera se os dados espaciais como processo estocástico e não como amostras independentes. Assim, todos os eventos serão utilizados para descrever o padrão espacial do evento estudado.

2.1.2 Autocorrelação Espacial

Autocorrelação é uma medida da estrutura de Dependência Espacial. Como o próprio nome sugere, autocorrelação mede a correlação da própria variável no espaço.

Há diversos indicadores para medir a autocorrelação, porém todas tem o mesmo conceito: averiguar como varia a dependência espacial, a partir da comparação entre os valores de uma amostra e de seus vizinhos.

De acordo com Câmara et al (2004), os indicadores de autocorrelação espacial são casos particulares de uma estatística de produtos cruzados do tipo:

$$\Gamma(d) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(d)\xi_{ij}$$

Em que, d é distância entre os valores observados y_i e y_j; w_{ij} matriz que fornece uma medida de proximidade espacial entre as valores observados y_i e y_j e ξ_{ij} matriz que fornece a correlação entre essas valores observados.

A autocorrelação espacial de uma variável aleatória no espaço tem valor compreendido entre -1 e 1. Sendo assim, a autocorrelação de uma variável aleatória medida no mesmo local sempre será 1. Se a variável aleatória for espacialmente independente então a autocorrelação espacial será próximo de zero. Quando as variáveis aleatórias forem similares aos seus vizinhos próximos então a autocorrelação espacial será positiva e próximo de 1. No entanto, se as variáveis aleatórias forem dissimilares aos seus vizinhos próximos, a autocorrelação espacial será negativa e próxima de -1.

Uma representação gráfica da função autocorrelação é dada pelo correlograma. No correlograma é representado a autocorrelação em função da distância. Vale ressaltar que quanto maior a distância entre as variáveis aleatórias no espaço, mais próxima do valor zero tenderá a autocorrelação.

2.1.3 Estacionariedade e Isotropia

O principal conceito estatístico que define a estrutura espacial dos dados é a Estacionariedade. Um processo é dito estacionário quando não há tendência nos dados, estatisticamente quer dizer que a média e a variância são constantes em toda região estudada (Câmara et al, 2004).

Segundo Bailey & Gatrell(1995), um processo é estacionário se a cova-

riância entre valores em quaisquer dois pontos depende da distância e direção entre ambos e não de suas localizações na região de estudo (Figura 1).

Um aspecto exclusivo da estatística espacial é a ocorrência ou não ocorrência da Isotropia. Um processo é considerado isotrópico se for estacionário e a covariância depender somente da distância entre os eventos e não da direção entre eles(Bailey & Gatrell, 1995).

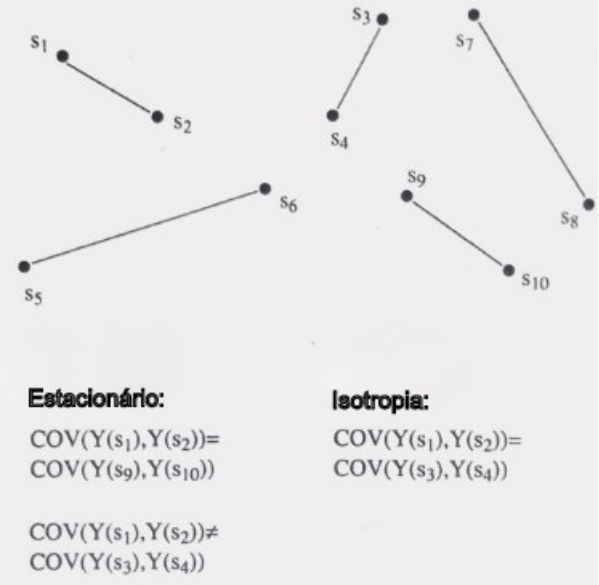


Figura 1: Ilustração de um processo estacionário e um processo isotrópico(Bailey & Gatrell, 1995)

Quando a estrutura de covariância, além de variar com a distância, varia simultaneamente em função da direção, é dito ser um processo Anisotrópico (Câmara et al, 2004). Um exemplo de processo anisotrópico é a densidade populacional do Brasil. A diminuição de densidade na direção Leste-Oeste, ou em direção ao interior do país, é mais intensa do que quando se caminha na direção do Sul ao Norte. A força da dependência espacial decai mais rapidamente nesse eixo do que acompanhando o litoral (Brasil, Ministério da Saúde, 2007).

2.2 Análise Espacial em Áreas

Dados por área representam uma agregação de valores que se encontram dispersos dentro de cada área. Para efeitos de análise, associa-se um valor como representativo de toda a área.

Matematicamente, considere $\{Y(A_i), A_i \in A_1, ..., A_n\}$ um conjunto de variáveis aleatórias indexadas por sub-regiões fixas $(A_1, ..., A_n)$ na região R, em que $A_1 \bigcup ... \bigcup A_n = R, \bigcap_{i=1}^{n} A_i = \emptyset$. Bailey & Gatrell (1995) generaliza as variáveis aleatórias $Y(A_i)$ como Y_i e denota como valores observados y_i.

Na maioria das vezes, as sub regiões (áreas) representam divisões administrativas de caráter político ou administrativo ou divisões com características geomorfológicas.

Contagem, proporção, media e mediana são algumas das variáveis aleatórias usadas para esse tipo de dados.

Dados agregados por área são observados, geralmente, em mapas em que o espaço é particionado em áreas e cada área é colorida de acordo com a variável quantitativa ou qualitativa. Esses mapas são denominados como mapas coropléticos. Cada área do mapa é classificada de acordo com o valor da variável, onde é feita uma divisão em classes, em que o número de classes se baseia em diferentes critérios: intervalos iguais, percentis ou desvios padrão. Esses critérios podem levar à diferente impressão visual de tendência e de padrão espacial para uma mesma região estudada (Câmara et al, 2004).

Intervalos iguais não são indicados para variáveis que tem distribuição assimétrica, pois quando se divide em classes de amplitudes iguais as variáveis de valor menor ou maior ficam reduzidas a poucas áreas, sendo assim haverá muitas áreas restritas a uma ou duas cores. Agrupando os dados por quantil significa que são geradas as mesmas quantidades de áreas em cada classe. Sendo assim, a amplitude das classes não é constante. O que prevalece é o número de áreas constante em cada classe. Essa forma de agrupamento ajuda a visualizar como um dado é percen-

tualmente distribuído. Por fim, o modo de agrupamento por desvio padrão produz classes em função do valor de desvio padrão em torno da média e o número de classes depende dos dados. Essa forma de agrupamento ajuda a analisar o comportamento dos dados em torno da média.

Portanto, para uma análise exploratória é importante experimentar diferentes critérios de agrupamento para a visualização do mapa coroplético.

É interessante ressaltar a escala em que os dados por área são coletados e analisados. Tratando-se de análise espacial, quando usa-se o termo escala refere-se ao tamanho da área de estudo (Atkinson, 2000) . Geralmente, quanto maior a área geográfica, mais heterogênea é a população onde acontece o processo em estudo. Entretanto, quanto menor a área, mais raros são os eventos observáveis, o que origina excessiva flutuação aleatória dos indicadores e problemas surgem.

Dados de áreas são os mais analisados na área da saúde, embora tenham sua importância frequentemente subestimada por receio da denominada falácia ecológica, que pode ser definida como uma tentativa de estimar associações entre indivíduos a partir de dados agregados. Por exemplo, ao analisar o problema de atropelamento, observa-se que a taxa de atropelamento em municípios ricos é maior. Porém não pode-se concluir que indivíduos ricos são os mais atropelados, mas pode-se concluir que municípios ricos tem um número maior de carros, e estes atropelam pedestres (Brasil, Ministerio da Saúde, 2007).

O problema de "unidade de área modificável" consiste em dois tipos de problemas: escala e zona (Openshaw, 1984). Problema referente a escala tem-se quando há muitas áreas e deseja-se agrega-las, porém surge a dúvida de como agrega-las. Já problemas de zona também consistem na escolha de como agregar as áreas, contudo, envolve na escolha entre zonas diferentes. As tabelas 1 e 2 ilustram uma escala (modificação do número de áreas) e zonas diferentes (modificação na forma da área).

Tabela 1: Escala: modificação no número de áreas

27	25	24	22
25	23	22	20
23	20	18	16
20	22	20	18

100	88
85	72

Tabela 2: Zonas diferentes: modificação na forma das áreas

185	160	127	126
127	131	148	176

188	148
157	105
131	133
127	191

2.3 Métodos de Suavização Espacial

2.3.1 Autocorrelação Espacial para Dados em Área

Para mostrar como os valores observados, $y_i, i = 1, ..., n$, estão correlacionados no espaço mede-se a dependência espacial. Para isso, utiliza-se os indicadores de autocorrelação: Índice de Moran I e Índice de Geary C. Através desses indicadores é possível medir quanto o valor observado na i-ésima área é dependente dos valores observados da mesma variável nas áreas vizinhas.

O índice de Moran I é indicado para dados que apresentam processos estacionários, pois calculando para dados não estacionários o índice perde a sua validade devido ao índice utilizar em seu cálculo a diferença entre cada valor observado y_i e a média global $\bar{y}$. Quando há não estacionariedade, ou seja, há tendência, o índice de Geary C é o indicado para calcular a autocorrelação espacial pois tal índice

utiliza em seu calculo a diferença entre os pares y_i e y_j (Bailey & Gatrell,1995).

Antes de apresentar os indicadores de autocorrelação, Índice de Moran I e Índice de Geary C, define-se uma matriz de proximidade espacial W a qual fornece as medidas de proximidade entre as áreas A_i e A_j da região de estudo.

Dado um conjunto de n áreas $\{A_i, ..., A_n\}$, seja a matriz de proximidade espacial $W_{(n,n)}$, em que cada elemento w_{ij} representa uma medida de proximidade espacial entre as áreas A_i e A_j. A escolha dos elementos w_{ij} é arbitrário e deve se levar em consideração o problema especifico sob análise. Esta medida de proximidade espacial pode ser calculada a partir dos seguintes critérios, (Bailey & Gatrell,1995):

- $w_{ij} = 1$, se o centroide de A_i está a uma determinada distância do centroide de A_j; caso contrário $w_{ij} = 0$
- $w_{ij} = 1$, se o centroide de A_i é um dos k centroides mais próximos de A_j; caso contrário $w_{ij} = 0$
- $w_{ij} = 1$, se A_i compartilha um lado comum com A_j; caso contrário $w_{ij} = 0$
- $w_{ij} = d_{ij}^{\gamma}$, em que d_{ij} é a distância entre os centroides de A_i e A_j, se $d_{ij} < \delta$, sendo $\delta > 0, \gamma < 0$; caso contrário $w_{ij} = 0$
- $w_{ij} = \frac{l_{ij}}{l_i}$, em que l_{ij} é o comprimento da fronteira entre A_i e l_i é o perímetro de A_i

Há casos que padroniza-se as linhas da matriz $W_{(n,n)}$, para que a soma das medidas de proximidade espacial w_{ij} de cada linha seja igual a 1. Isto facilita os cálculos.

Pode-se generalizar a matriz de proximidade espacial para vizinhos de maior ordem, isto é, vizinho do vizinho. Vale ressaltar que os critérios utilizados para a matriz de vizinhança de primeira ordem $W^{(1)}$ também se aplicam para as matriz $W^{(2)}, ..., W^{(n)}$.

Índice de Moran I

O índice de Moran I (1950) mede o grau de correlação entre pares de valores observados y_i e y_j ponderado pela medida de proximidade w_{ij}, (Bailey & Gatrell,1995):

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{(\sum_{i=1}^{n}(y_i - \bar{y})^2)(\sum_{i \neq j} w_{ij})}$$

Em que: n é o número de áreas, y_i valor observado da variável Y_i da i-ésima área, $\bar{y}$ é a média dos valores observados na região de estudo e w_{ij} os elementos da matriz de proximidade espacial W.

A equação acima, é calculada para matriz de proximidade espacial de primeira ordem. Para matriz de proximidade de maior ordem (k), o índice de Moran I é dado por:

$$I^{(k)} = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^{(k)} (y_i - \bar{y}) (y_j - \bar{y})}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right) \left(\sum_{i \neq j} w_{ij}^{(k)}\right)}$$

Em que $w_{ij}^{(k)}$ são os elementos da matriz de proximidade espacial $W^{(k)}$ de ordem k.

Se os valores y_i e y_j forem espacialmente independentes, o valor de I é igual a zero. Quando há similaridades entre as áreas próximas, I tende há um valor positivo. No entanto, quando há dissimilaridade entre as áreas próximas, I tende há um valor negativo.

Observe que, o índice de Moran I tem a mesma forma do coeficiente de correlação usual. O numerador mostra a média de produtos dos desvios em relação à média e o denominador uma medida de variabilidade dos desvios.

Índice de Geary C

O índice de Geary C (1954) mede o grau de correlação entre pares de valores observados y_i e y_j ponderado pela medida de proximidade w_{ij},(Bailey & Gatrell,1995):

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(y_i - y_j)^2}{2\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)\left(\sum_{i\neq j} w_{ij}\right)}$$

Em que: n é o número de áreas, y_i valor observado em A_i, $\bar{y}$ é a média dos valores observados na região de estudo e w_{ij} os elementos da matriz de proximidade espacial W.

Valores pequenos de C indicam autocorrelação positiva, enquanto valores grandes de C indicam autocorrelação negativa (Lloyd, 2001). O índice de Geary C é usado com menor frequência que o índice de Moran I (Assunção, 2004).

2.3.2 Média Móvel Espacial

Para explorar a variação da tendência espacial dos dados calcula-se a média dos valores observados das áreas vizinhas. A média móvel espacial $\hat{\mu}_i$ produz uma nova superfície com menor flutuação aleatória que os dados originais, ou seja, a variabilidade espacial dos dados diminui. A média móvel $\hat{\mu}_i$ associada ao valor observado y_i, referente à i-ésima área, é calculada a partir dos elementos w_{ij} da matriz de proximidade espacial $W^{(1)}$.

Bailey & Gatrell (1995) definem a média móvel $\hat{\mu}_i$ como sendo:

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij} y_j}{\sum_{j=1}^{n} w_{ij}}$$

Note-se que, normalizando a matriz de proximidade espacial $W^{(1)}$ obtém-se o denominador igual a 1.

2.4 Testes de autocorrelação espacial

Considera-se a hipótese nula como a da independência espacial entre os valores observados y_i e y_j, isto é, $H_0 : I = 0$. Há duas abordagens para testar a significância do índice de Moran I: teste de distribuição amostral aproximado e teste de permutação aleatória.

O teste de distribuição amostral aproximado é um teste de dependência espacial e faz a suposição de que as variáveis aleatórias $Y_i, i = 1, ..., n$, são espacialmente independentes e que seguem uma distribuição normal.

O teste de permutação aleatória investiga a forma de como as variáveis aleatórias,Y_i, são organizadas espacialmente.

Teste de distribuição amostral aproximado

Segundo Bailey & Gatrell, um teste para autocorrelação espacial pode ser construído quando há um número n suficientemente grande de áreas.

Sejam os valores observados $y_i, i = 1, ..., n$, da variável Y_i espacialmente independentes e cuja a distribuição é normal. Então o índice de Moran I segue uma distribuição amostral que é, aproximadamente, normal com valor esperado e variância dados por:

$$E(I) = \frac{-1}{(n-1)}$$

$$VAR(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2}$$

Em que:

$$S_0 = \sum_{i \neq j} w_{ij}$$

$$S_1 = \frac{1}{2} \sum_{i \neq j} (w_{ij} + w_{ji})^2$$

$$S_2 = \sum_k \left(\sum_j w_{kj} + \sum_i w_{ik} \right)^2$$

Um valor "extremo" da estatística I indica autocorrelação espacial significativa.

Teste de permutação aleatória

Uma abordagem alternativa é um teste de permutação aleatória. A abordagem baseia-se na ideia de que, se existem n áreas sobre uma região R, então são geradas $n!$ permutações dos valores observados $y_i, i = 1, ..., n$, associados a cada área A_i; cada permutação produz um novo arranjo espacial, onde os valores observados são redistribuídas entre as áreas. Como apenas um dos arranjos corresponde à situação original, pode se construir uma distribuição empírica de I. O valor de I pode ser obtido por qualquer uma das permutações. Se o valor do índice I medido originalmente corresponder a um "extremo" da distribuição simulada, então trata-se de valor com significância estatística(Bailey & Gatrell,1995).

Na prática, geralmente não é possível obter as $n!$ permutações e o método de Monte Carlo pode aproximar da distribuição de permutação. Um número razoável de valores pode ser retirado aleatoriamente entre as $n!$ permutações(Bailey & Gatrell,1995).

Se assumir que o processo de redistribuir os valores observados em cada área é aleatório, onde o padrão observado é uma das muitas permutações possíveis, então a média e a variância de I são dadas por (Lloyd, 2011).

$$E(I) = \frac{-1}{(n-1)}$$

$$VAR(I) = \frac{nS_4 - S_3 S_5}{(n-1)(n-2)(n-3) S_0^2}$$

Em que:

$$S_3 = \frac{n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^4}{(n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2)^2}$$

$$S_4 = (n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2$$

$$S_5 = S_1 - 2nS_1 + 6S_0^2$$

2.5 Estimação de indicadores

Métodos de estimação bayesiana para taxas permitem a correção de efeitos associados a pequenas populações. De acordo com Câmara et al (1995), vários estudos têm mostrados que divisões políticas como bairros e municípios apresentam relações inversas de área e população, isto é, os maiores bairros em população tendem a ter menores áreas, e vice- versa. Consequentemente, em um mapa coroplético haverá taxas de valores extremos devido ao número reduzidíssimo de observações, sendo assim menos confiável.

O estimador Kernel é uma técnica de interpolação, não paramétrica, que gera uma superfície de acordo com a distribuição espacial das variáveis em estudo. A geração de superfícies é uma maneira eficiente para identificar visualmente os padrões espaciais.

O estimador bayesiano é utilizado para um modelo de variação espacial discreta, enquanto o estimador Kernel para um modelo de variação contínuo.

2.5.1 Estimador Bayesiano Empírico

Considere-se um processo estocástico $Y_i, i = 1, ..., n$, em que Y_i é a realização de um processo espacial na i-ésima área. Seja uma região R particionada em n áreas. Sejam θ_i uma taxa real associada a i-ésima área e y_i um valor observado na i-ésima área com distribuição de Poisson. Na formulação não bayesiana, o estimador de máxima verossimilhança de θ_i é $r_i = y_i/n_i$ em que n_i é o tamanho da população na i-ésima área(Marshall, 1991).

Suponha-se que a distribuição de probabilidade a priori de θ_i tenha média γ_i e variância ϕ_i . O melhor estimador bayesiano de θ_i é dado pela combinação linear entre a taxa observada r_i e a média γ_i (Bailey & Gatrell,1995):

$$\hat{\theta}_i = \omega_i r_i + (1 - \omega_i)\gamma_i$$

Em que:

$$\omega_i = \frac{\phi_i}{\phi_i + \gamma_i / n_i}$$

Em áreas em que o tamanho da população é muito grande o fator peso ω_i terá valor próximo de 1 e consequentemente o estimador bayesiano $\hat{\theta}_i$ terá um valor próximo a taxa observada r_i, pois o fator peso ω_i da taxa observada r_i é maior do que o ω_i do modelo a priori (ou seja, γ_i). Entretanto, se a população for muito pequena, ω_i terá um valor próximo de zero e $\hat{\theta}_i$ terá um valor próximo a γ_i.

Áreas com populações muito baixas terão uma correção maior, e áreas com populações grandes terão pouca alteração em suas taxas. Logo θ_i será estimado, quando n_i for pequeno, com maior peso da média.

Através do estimador bayesiano empírico obtém se a estimação dos parâmetros γ_i e ϕ_i de cada área. O método de bayesiano empírico recebe esse nome porque a média e a variância são estimadas a partir dos dados.

Suponha que a variável aleatória θ_i tem a mesma distribuição para cada área; sendo assim todas as médias e variâncias são iguais. Logo, γ_i e ϕ_i são estimados respectivamente por:

$$\hat{\gamma} = \frac{\sum y_i}{\sum n_i}$$

$$\hat{\phi} = \frac{\sum n_i (r_i - \hat{\gamma})^2}{\sum n_i} - \frac{\hat{\gamma}}{\bar{n}}$$

Em que $\bar{n}$ é a média do tamanho da população em todas as áreas.

Segundo Bailey & Gatrell (1995) o fator peso ω_i é estimado por:

$$\hat{\omega}_i = \frac{\hat{\phi}}{(\hat{\phi} + \hat{\gamma}/n_i)}$$

E portanto o estimador bayesiano empírico é estimado por:

$$\hat{\theta}_i = \hat{\gamma} + \frac{\hat{\phi}(r_i - \hat{\gamma})}{(\hat{\phi} + \hat{\gamma}/n_i)}$$

2.5.2 Estimador Kernel

Considera-se um modelo de variação espacial contínua, que supõe um processo estocástico $\{Y(x), x \in A, A \subset R^2\}$, cujos valores podem ser conhecidos em todos os pontos da área de estudo.

Para analisar a distribuição dos pontos da área de estudo, o primeiro passo é utilizar a técnica de Kernel. A técnica de Kernel é uma técnica de interpolação, não paramétrica, em que uma distribuição de pontos é transformada em uma superfície de densidade para a identificação visual da ocorrência da concentração do evento, indicando locais de aglomeração, caso existam, bem como a forma como os dados se distribuem na região (Bailey & Gatrell,1995).

A estimativa Kernel depende de dois parâmetros: raio de influência (τ) e a função de estimação Kernel $k(.)$. O raio de influência define uma vizinhança de pontos utilizada para estimar o valor em um ponto s, sendo s uma localização genérica em uma região R, a ser interpolado. Já a função de estimação Kernel tem propriedade de suavizar o fenômeno.

Sejam s uma localização genérica em uma região R e $s_i, i = 1, 2, ..., n$, as localidades de n eventos em R. Então a intensidade de $\lambda_\tau(s)$ em s, é estimada por:

$$\hat{\lambda}_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right), s - s_i \leq \tau$$

sendo o parâmetro τ o raio de influência que define a vizinhança do ponto s a ser interpolado, o qual controla o alisamento da superfície gerada e $k(.)$ uma função de estimação Kernel com propriedades de suavização do fenômeno.

As funções de estimação kernel normal ou quártico são as mais comumente utilizadas, porém saber qual a função de estimação Kernel que será utilizada não é um ponto crítico e sim, a escolha do raio de influência (τ) é crucial pois pode alterar as estimativas finais (Bailey & Gatrell, 1995).

A função de estimação kernel normal e quártico são dadas, respectivamente, por:

$$k(h) = \frac{1}{2\pi\tau} e^{\left(-\frac{h^2}{2\tau^2}\right)}$$

$$k(h) = \frac{3}{\pi}(1 - h^2)^2$$

em que h é a distancia entre s e s_i.

Segundo Bailey & Gatrell (1995), para uma análise da distribuição espacial de dados em áreas, o real interesse não é estimar o número de observações por unidade de área, $\lambda(s)$, mas estimar o valor médio, $\rho(s)$, de um atributo cujos valores tenham sido coletados nas localizações s_i da região R.

Uma forma intuitiva para introduzir os valores de atributos na estimativa de kernel é considerando,(Bailey & Gatrell,1995):

$$\sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right) y_i, s - s_i \leq \tau$$

em que s é uma localização genérica a ser estimado, s_i localização dos eventos dentro do raio de influência τ e y_i é o valor observado em s_i

Se a estimativa Kernel inicial representa "o número de observações por unidade de área", então esta extensão em certo sentido, representa "a soma total do valor do atributo por unidade de área". Consequentemente, para transformar a expressão acima em uma estimativa dos valores médios do atributo, terá que dividi-lo pelo "número de observações por unidade de área". Isto sugere que uma estimativa kernel apropriada para $\rho(s)$ dada por,(Bailey & Gatrell,1995):

$$\hat{\rho}_\tau(s) = \frac{\sum_{i=1}^{n} k\left(\frac{(s-s_i)}{\tau}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{(s-s_i)}{\tau}\right)}, s - s_i \leq \tau$$

onde τ^{-2} do numerador e do denominador foram simplificados. Nas localizações s_i em que o denominador é zero, o numerador também deve ser zero e por convenção $\hat{\rho}_\tau(s)$ é definido como zero nestes pontos.

Quando o valor observado y_i da i-ésima área representar contagens, tais como obtidas pelo censo, o estimador de kernel apresentado acima não é aplicável. Um valor "médio" de um atributo como "número de domicílios precários" não faria sentido, e deve-se pensar em termos de "número de domicílios precários por unidade de área". Uma estimativa para estimar o "total por unidade de área", sugerido por Bailey & Gatrell (1995)é o numerador da equação acima:

$$\hat{\lambda}_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right) y_i, s - s_i \leq \tau$$

em que s é o centro da i-ésima área a ser estimada, s_i localização dos eventos dentro do raio de influência τ e y_i contagem populacional da i-ésima área.

3 MATERIAL E MÉTODOS

Para aplicação dos métodos descritos no capítulo anterior, exceto o Estimador Bayesiano Empírico, foi utilizado um conjunto de dados referentes aos casos positivos de dengue na cidade de Rio Claro - SP. Nestes dados foi realizada uma análise exploratória através de mapas de diferentes escalas: bairros e setores censitários, e foram construídas as estimativas de Kernel. Não se aplicou o Estimador Bayesiano Empírico devido à dificuldade em obter o número total de moradores de Rio Claro por bairro e por setor censitário.

3.1 Conjunto de dados

Neste estudo foi utilizado um conjunto de dados referente a 1656 casos positivos de dengue na cidade de Rio Claro - SP, registrados no primeiro semestre de 2011. Os dados foram fornecidos pela Defesa Civil, órgão subordinado à Secretaria de Segurança Pública do Município de Rio Claro - SP.

Trata - se de 1656 endereços das residências de moradores que foram diagnosticados com dengue, porém não foi utilizado todos os endereços devido ao problema de georreferenciamento. Utilizou-se 1276 endereços para o mapa de 127 bairros (Figura 2) e 1353 endereços para o mapa de 263 setores censitários (Figura 3) e através do software gvSIG 1.12 contabilizou-se o número de casos positivos da dengue por bairro e por setor censitário.

A dengue é uma doença viral e é transmitida, no Brasil, através do mosquito Aedes aegypti. Atualmente, a dengue é considerada um dos principais problemas de saúde pública de todo o mundo.

Vale ressaltar que nem todos os moradores contraíram o vírus causador da dengue em sua residência, porém foi considerado o endereço residencial dos pacientes como sendo o local onde se contraiu o vírus.

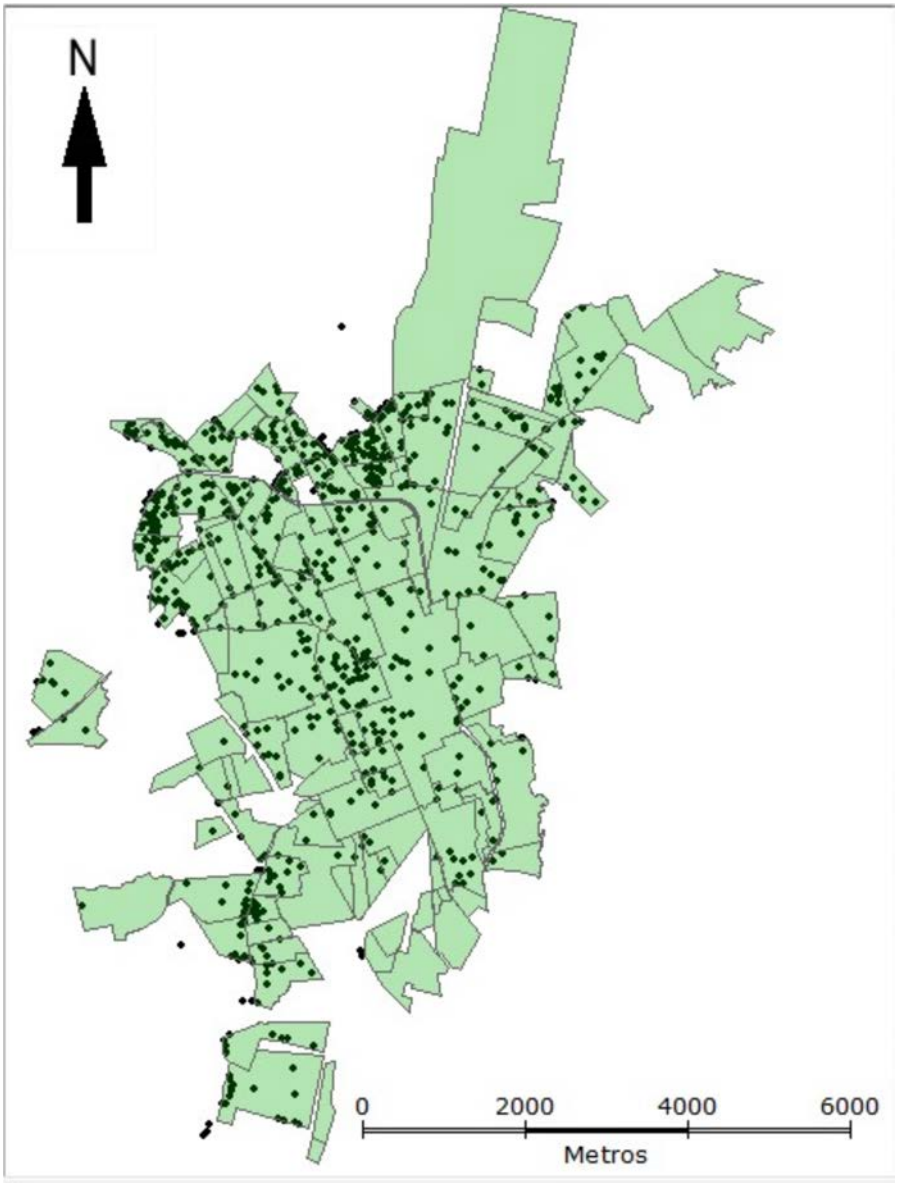


Figura 2: Localização dos casos positivos de dengue em 127 bairros no município de Rio Claro - SP

Figura 3: Localização dos casos positivos de dengue em 263 setores censitários no município de Rio Claro - SP

Observa-se que também é cabível outra abordagem, em que não se faz a contabilização dos casos, mas apenas uma análise pontual dos dados, obtendo assim uma escala menor; tal abordagem já foi objeto de estudo no trabalho *Análise de Técnicas de Distribuição Espacial com Padrões Pontuais e Aplicação a Dados de Acidentes de Trânsito e a Dados de Dengue de Rio Claro-SP* (Kawamoto, 2012).

Uma comparação entre os resultados dos dois métodos é feita no capítulo Resultados e Discussões.

Com o auxílio do software Terra View 4.2.2 foram construídos mapas para uma análise exploratória com três agrupamentos diferentes: Intervalos iguais, Intervalo por quantil e Intervalo por desvio padrão e, em seguida, construído os respectivos mapas de suavização através das Médias Móveis Espacial. Calculou-se o índice de Moran I e para testar a significância da estatística I utilizou-se o teste de permutação aleatória, considerando 99 permutações. Vale ressaltar que, o software Terra View 4.2.2 sugere duas opções de permutação: 99 e 999.

Ushizima(2005) em seu estudo, de mapeamento da dengue na área urbana de Rio Claro nos períodos de 2001 à 2003 e relacionando com variáveis sócio-econômicas, utilizou mapas do setor censitário de Rio Claro considerando agrupamento de intervalos iguais para as taxas de incidência da dengue de cada ano. Uma discussão entre os mapas do presente estudo e os mapas de Ushizima(2005) será apresentado no capítulo a seguir.

Ainda no software Terra View 4.2.2 foram construídas as estimativas de Kernel, detalhadas no capítulo anterior, para dois tipos de função de estimação: Kernel Normal e Kernel Quártico. Para a Função de Estimação Kernel Normal foram usados os seguintes raios de influência: 100m, 150m, 200m e 500m. Já para a Função de Estimação Kernel Quártico foram usados os raios de influência: 250m, 375m, 500m e 625m. As funções de estimação de Kernel: Normal e Quártico e seus respectivos raios de influências são as mesmas utilizados por Kawamoto(2012).

4 RESULTADOS E DISCUSSÕES

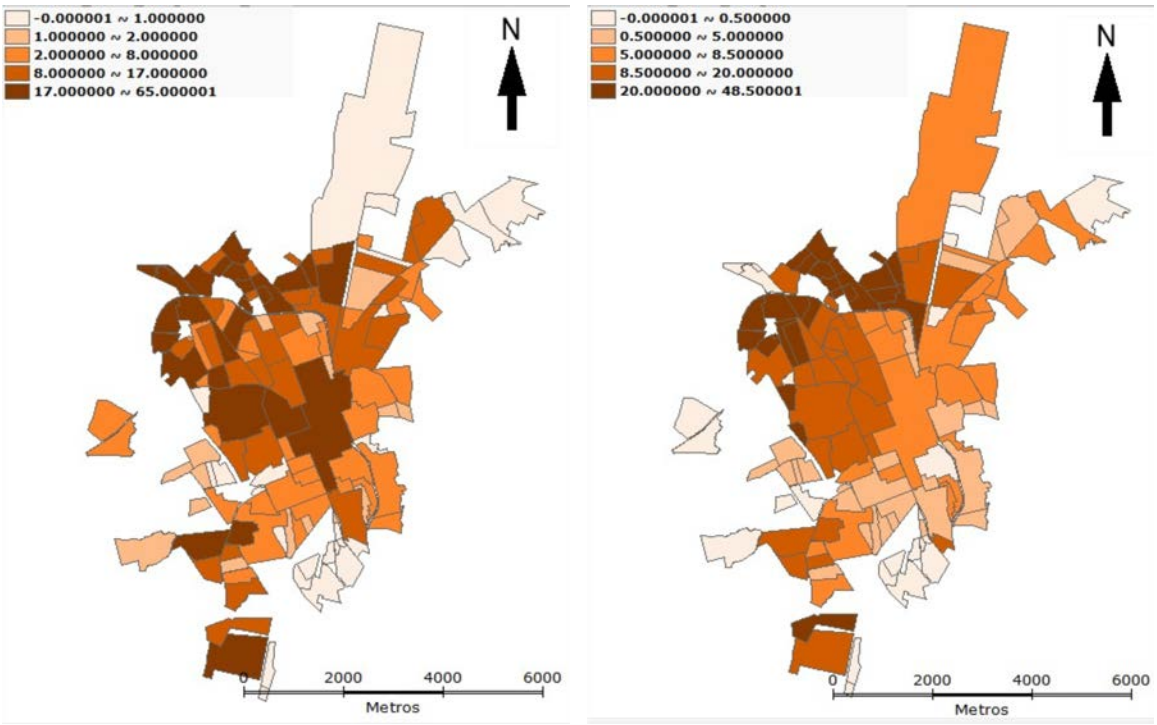
4.1 Análise Exploratória

As figuras 4, 5 e 6 apresentam o número de ocorrências de dengue, em 127 bairros, da cidade de Rio Claro-SP, com intervalos diferentes.

Na figura 4(a), o intervalo foi construido por quintis, ou seja, com o número de ocorrências de dengue, de cada bairro, ordenados de forma crescente divide-se em cinco partes iguais, representando 20% dos 127 bairros.

Visualmente, é dificil identificar algum padrão para as ocorrências de dengue, logo tem-se uma flutuação aleatória. Ademais pode-se observar que 20% dos bairros tem ocorrências de dengue entre 17 e 65. Entretanto, 20% dos bairros possui ocorrências baixas entre 0 e 1 e os 60% restantes tem ocorrências entre 1 e 17.

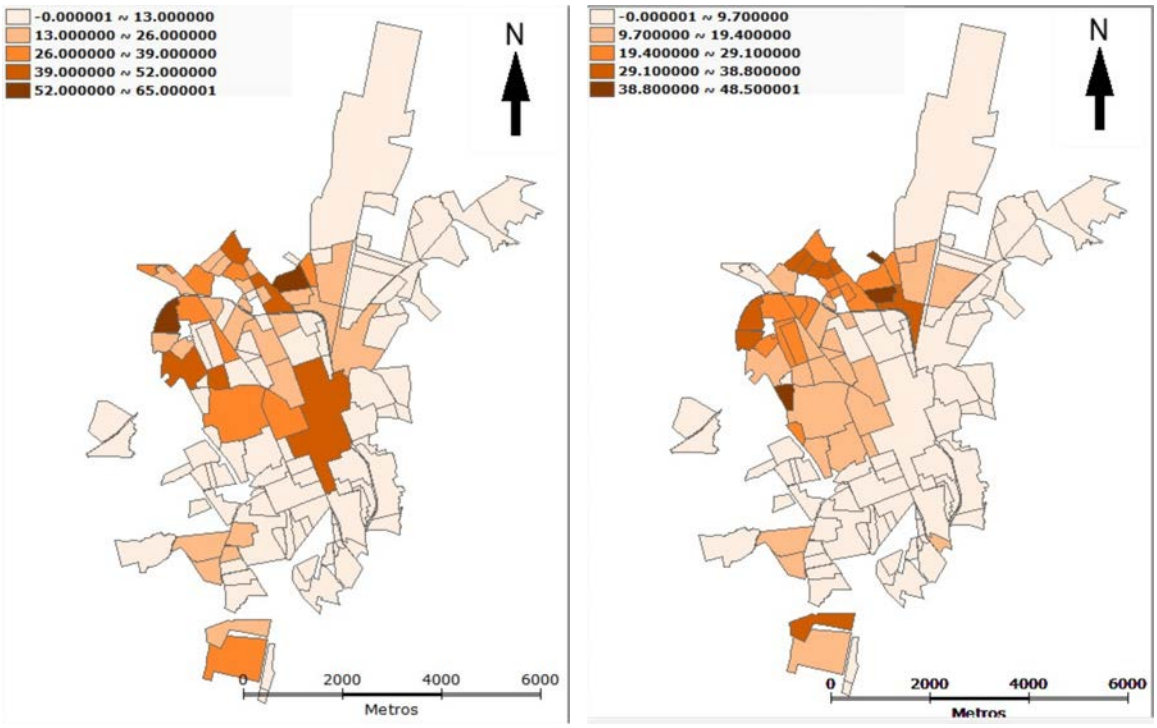
Ora, aplicando as Médias Móveis na Figura 2(a) obtem-se um mapa com a redução da flutuação aleatória(Figura 4(b)).



(a) Quintis (b) Média Móvel

Figura 4: Casos positivos de dengue do município de Rio Claro em 127 bairros: (a)Intervalos por quintis, (b)Média Movel para intervalos por quintis

Na figura 5(a), utilizou-se intervalos iguais para visualizar as ocorrências de dengue, com 5 classes de amplitude igual a 13.



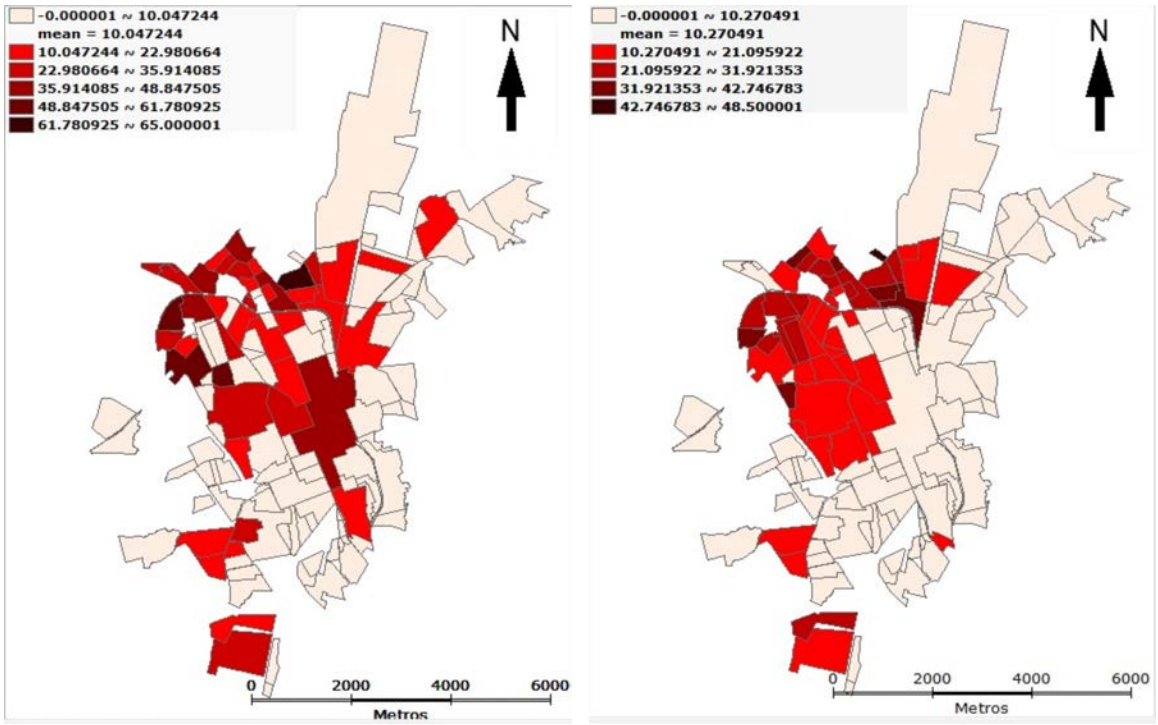
(a) Intervalo Igual (b) Média Móvel

Figura 5: Casos positivos de dengue do município de Rio Claro em 127 bairros: (a)Intervalo Igual, (b)Média Movel para intervalos iguais

Observe-se que, o maior índice de casos de dengue acontece em dois bairros que se localizam na região norte da cidade com ocorrências entre 52 e 65. Um número grande de bairros apresentam índices entre 0 e 13.

Na Figura 5(b) nota-se um mapa mais suave em comparação com a Figura 5(a), ou seja, é perceptível uma aglomeração de indices altos de dengue na região norte da cidade.

A figura 6(a), os dados foram agrupados por desvios padrão mostrando como o número de ocorrências da dengue comporta em torno da média de ocorrências, 10,047244.



(a) Desvio padrão (b) Média Móvel

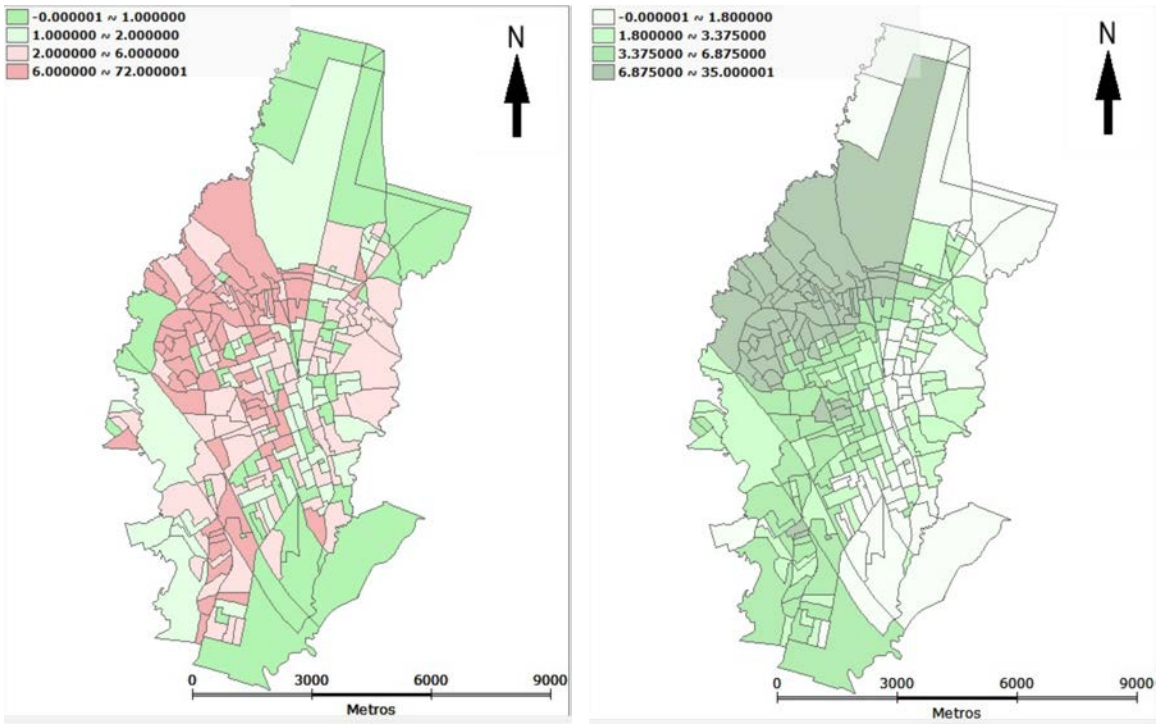
Figura 6: Casos positivos de dengue do município de Rio Claro em 127 bairros: (a)Intervalo: desvio padrão, (b)Média Móvel para desvio padrão

Nota-se que, há mais bairros com ocorrências de dengue abaixo da média do que bairros com ocorrências maiores. Sendo assim, os dados possuem uma distribuição assimetria positiva. Bairros que possuem ocorrências acima da média se localizam na região norte da cidade. A Figura 4(b), visualmente, mostra uma aglomeração de bairros com ocorrências acima da média em uma mesma região da cidade.

Comparando as figuras 4, 5 e 6, nota-se que há diferença nos resultados em relação aos critérios: quintis, intervalos iguais e desvio padrão, e também os transformados por média móvel.

As figuras 7, 8 e 9 também apresentam o número de ocorrências da dengue porém para 263 setores censitários, da cidade de Rio Claro - SP, e agrupados por: intervalos iguais, quartil e desvio padrão.

Na figura 7(a), os dados estão agrupados por quartil, isto é, o número de ocorrência da dengue, de cada setor censitário, ordenados crescentemente e divididos em quatro partes iguais, representando 25% dos 263 setores.



(a) Quartil (b) Média Móvel

Figura 7: Casos positivos de dengue do município de Rio Claro em 263 setores censitários: (a)Intervalo: Quartil, (b)Média Móvel para Quartil

Não se observa nenhum padrão para as ocorrências da dengue mas quando calcula-se a média móvel pode-se notificar uma aglomeração na região norte (Figura 7(b)).

A figura 8(a) mostra os dados agrupados por intervalos iguais com 5 classes de amplitude igual a 14,4.

(a) Intervalos Iguais (b) Média Móvel

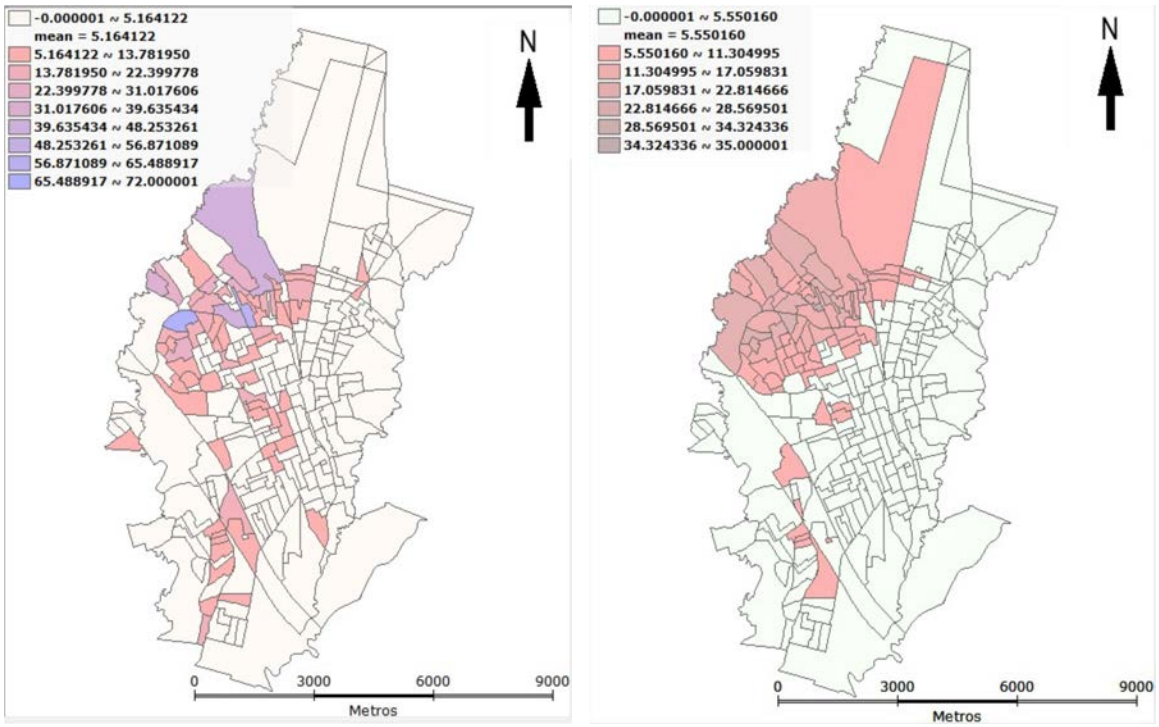
Figura 8: Casos positivos de dengue do município de Rio Claro em 263 setores censitários: (a)Intervalo Iguais, (b)Média Móvel para intervalos iguais

Nota-se que o maior número de casos da dengue apareceu em dois setores censitários localizados na região norte da cidade com ocorrências entre 58 e 72. Um número grande de setores apresentam índices entre 0 e 14.

Assim como na figura 5(a), independente da escala, obteve-se o mesmo resultado: dois bairros da região norte com maior número de casos da dengue.

A figura 8(b) mostra uma evidência aglomeração dos índices altos na região norte da cidade.

Na figura 9(a) apresenta os dados agrupados por desvios padrão e na figura 9(b) sua respectiva média móvel. Comparando com a figura 6, tem-se o mesmo resultado: os dados possuem uma distribuição assimétrica positiva porém a média é dada por 5,164122.



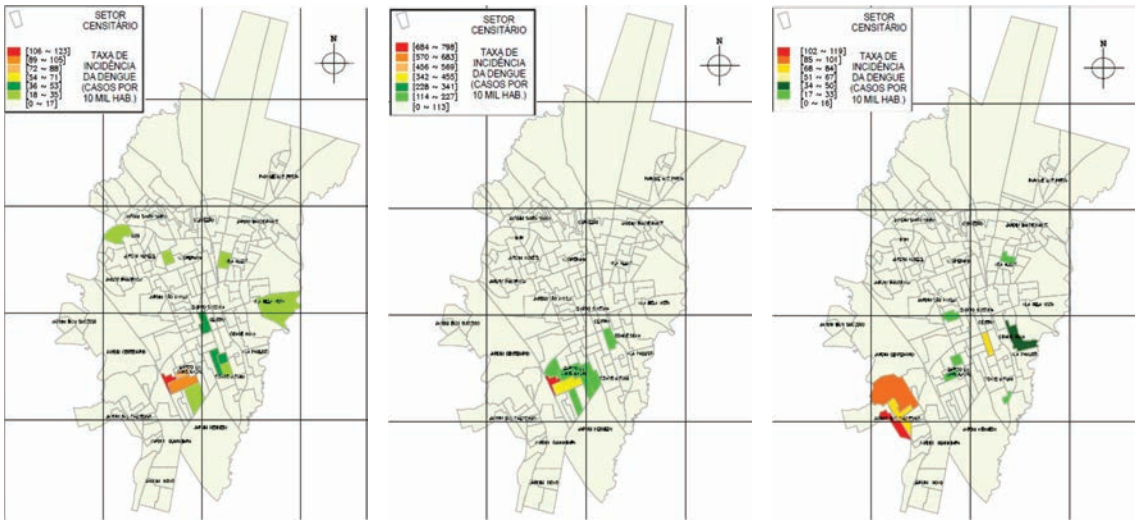
(a) Desvio padrão (b) Média Móvel

Figura 9: Casos positivos de dengue do município de Rio Claro em 263 setores censitários: (a)Intervalo: desvio padrão, (b)Média Móvel para desvio padrão

Aplicando o teste de permutação aleatório tem se que há dependencia espacial entre os valores observados, onde a estatistica I é igual a 0,389828 e o p-valor igual a 0,01.

De forma geral, independente da escala utilizada: bairros ou setor censitário, a análise exploratória com os mapas obteve-se resultados semelhantes, em que os maiores casos de dengue, no primeiro semestre de 2011, encontra-se na região norte da cidade de Rio Claro -SP. Cabe ao pesquisador escolher qual a melhor escala que deve-se trabalhar.

Na figura 10, Ushizima(2005) utilizou intervalos iguais para agrupar a incidência da dengue nos anos 2001, 2002 e 2003 na cidade de Rio Claro-SP.



(a) 2001 (b) 2002 (c) 2003

Figura 10: Mapa da incidência da dengue para os anos 2001, 2002 e 2003(Ushizima, 2005)

Observa-se que, neste estudo, as maiores ocorrências de dengue encontra-se localizada na região sul, para os anos 2001 e 2002, e na região sudoeste, para o anos 2003, da cidade.

A diferença entre os trabalhos de Ushizima(2005) e o presente trabalho deve-se, possilvemente, ao fato de que no período de 2001 à 2003 o cemitério, que está localizado na região sudoeste, ser um foco de criação do mosquito da dengue. Em 2011, devido à conscientização da população por parte da Vigilância Sanitária, o cemitério deixou de ser um foco de criação do referido mosquito, contribuindo para a diminuição do mosquito na área e, consequentemente, da ocorrência de dengue.

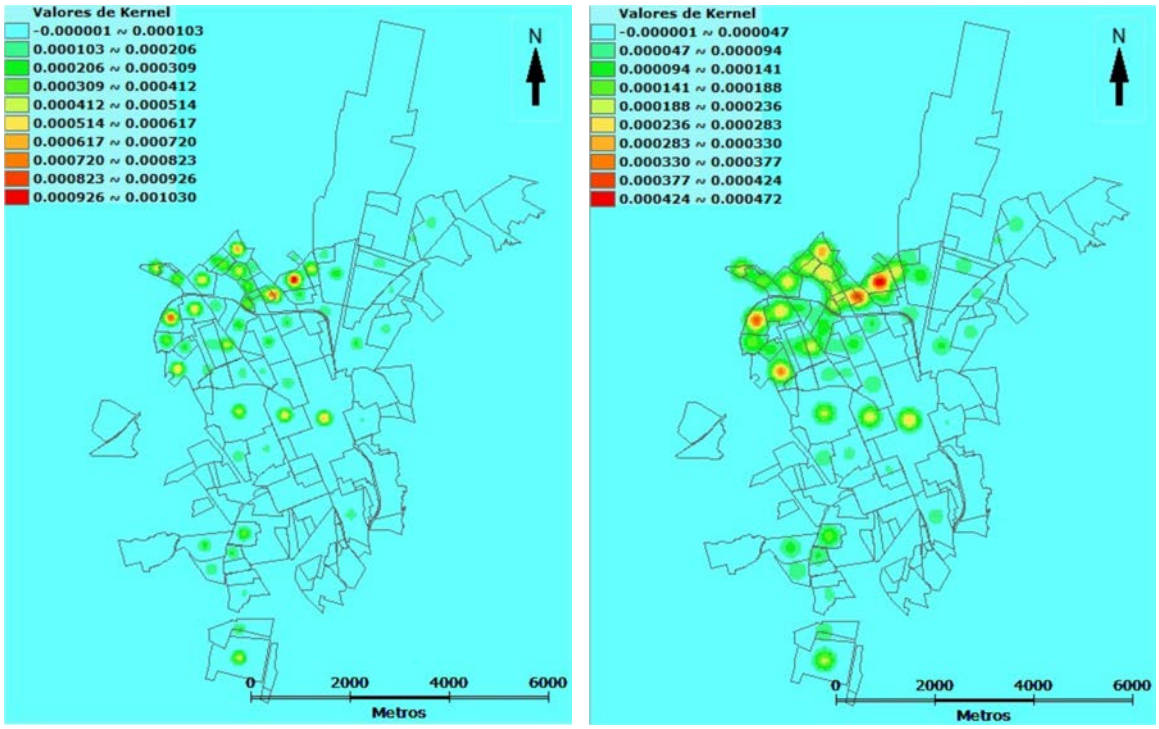
No ano de 2011, possíveis criadouros do mosquito aparecem em área da região norte da cidade, contribuindo para o aumento do número de ocorrências.

4.2 Estimador Kernel

Para a aplicação do Estimador Kernel utilizou mapas de escalas diferentes: bairros e setor censitário. As figuras 11, 12, 13 e 14 foi utilizado mapa de

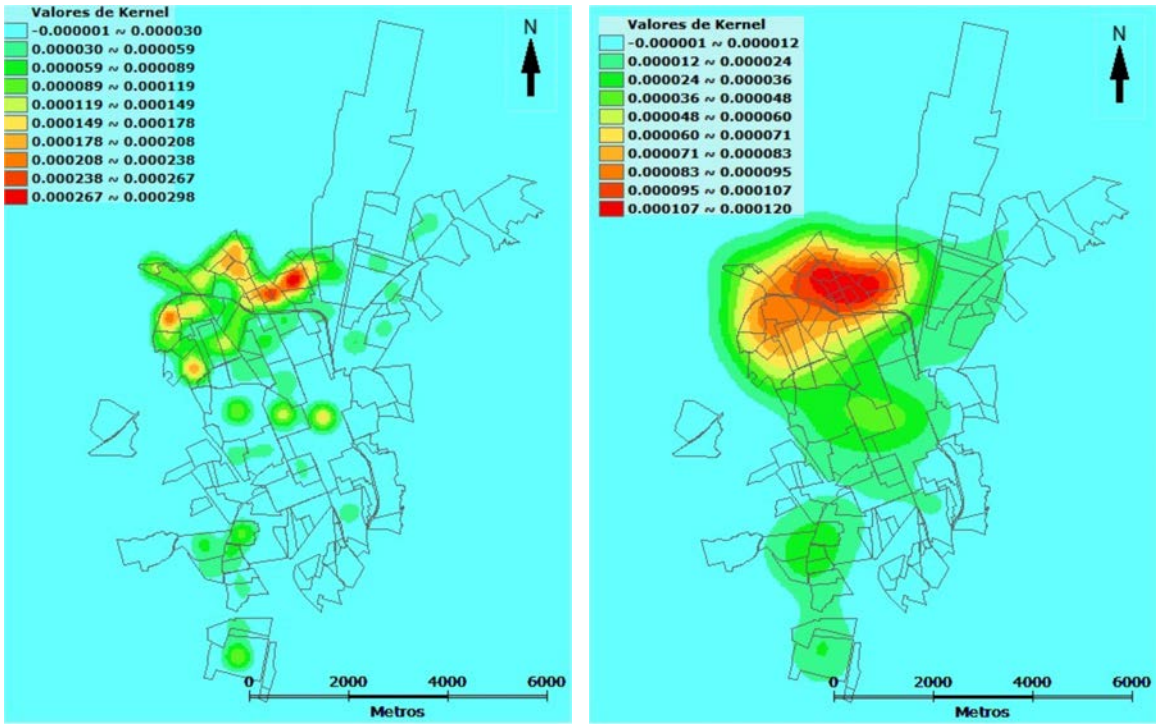
escala bairro, enquanto as figuras 17, 18, 19 e 20 mapa de escala setor censitário.

As figuras 11 e 12 apresentam os resultados da aplicação do Estimador Kernel para a função de estimação kernel normal para os respectivos raios de influência: 100m, 150m, 200m e 500m.



(a) $\tau = 100$m (b) $\tau = 150$m

Figura 11: Estimação kernel normal aplicado aos dados de dengue: (a)raio de influência = 100m, (b)raio de influência = 150m

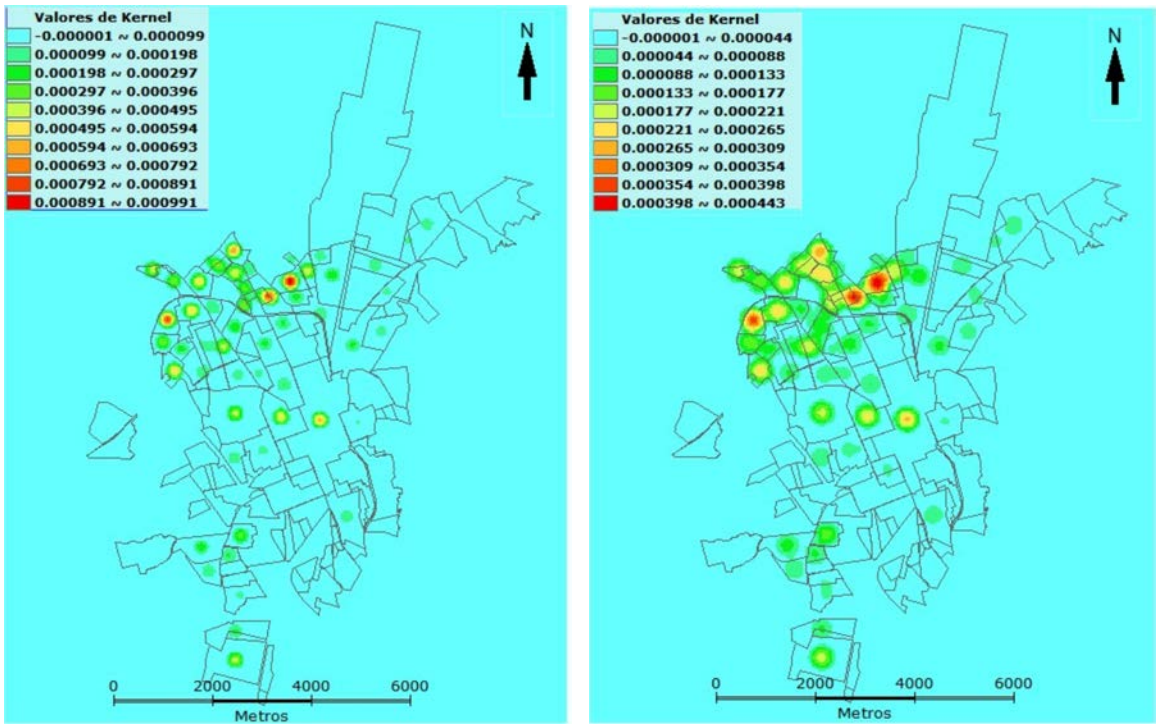


(a) $\tau = 200$m (b) $\tau = 500$m

Figura 12: Estimação kernel normal aplicado aos dados de dengue: (a)raio de influência = 200m, (b)raio de influência = 500m

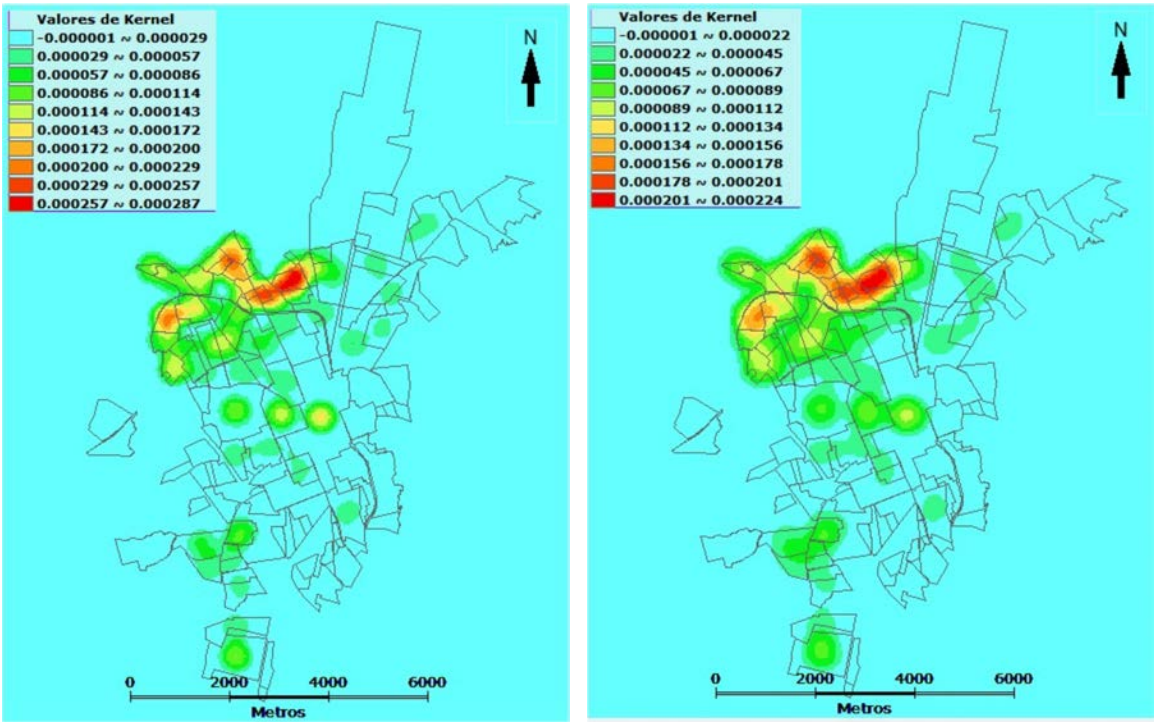
Das figuras 11 e 12, observa-se que há um aumento da suavização com o aumento do raio de influência. Para o raio de 100 m, são realçados somente os pontos extremamente críticos das ocorrências, enquanto que, para raio de 500m, toda uma região problemática é enfatizada. Cabe ao pesquisador decidir sobre o melhor raio para o seu problema.

As figuras 13 e 14 apresentam os resultados da aplicação do Estimador Kernel para a função densidade kernel quártico para os respectivos raios de influência: 250m, 375m, 500m e 625m.



(a) $\tau = 250$m (b) $\tau = 375$m

Figura 13: Estimação kernel quártico aplicado aos dados de dengue: (a)raio de influência = 250m, (b)raio de influência = 375m



(a) $\tau = 500$m (b) $\tau = 625$ metros

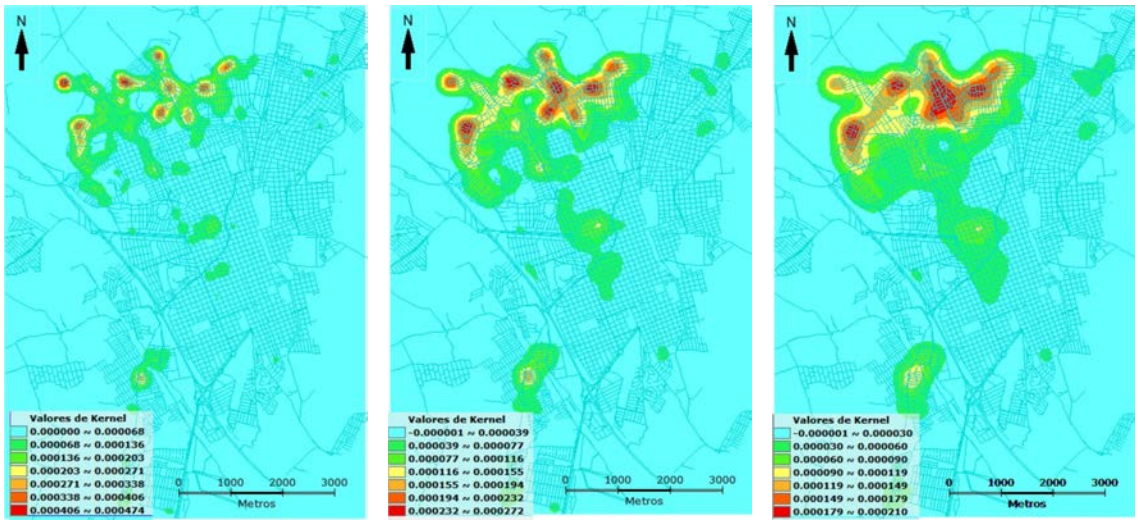
Figura 14: Estimação kernel quártico aplicado aos dados de dengue: (a)raio de influência = 500m, (b)raio de influência = 625m

Observa-se que o comportamento do kernel quártico é similar ao do kernel normal, porém com diferentes raios de influência. Por exemplo, os resultados da figura 11.a (kernel normal, com raio de influência igual a 100m) são semelhantes aos resultados da figura 13.a (kernel quártico, com raio de influência igual 250m). Este resultado corrobora as observações de Bailey e Gatrell(1995), de que a função de ajuste não é de grande importância, já que o controle pode ser feito através do raio de influência para a estimativa em cada ponto.

Kawamoto(2012) aplicou a estimação de kernel para o mesmo conjunto de dados, porém considerou os dados como dados por ponto e não como dados por área, isto é, não particionou a região de estudo.

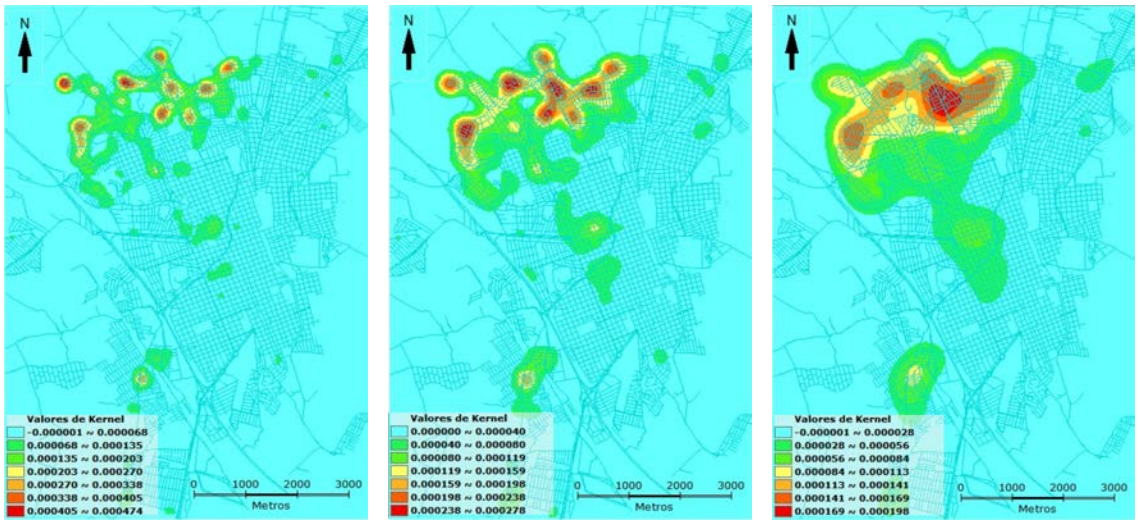
As figuras 15 e 16 apresentam os resultados que Kawamoto(2012) obteve aplicando a estimação de Kernel para a cidade de Rio Claro desconsiderando o

número de ocorrências de dengue por bairros. Na figura 15 apresenta a estimação de Kernel normal para os respectivos raios de influência: 100m, 150m e 200m. Enquanto a figura 16 apresenta a estimação de kernel quártico para raio de influência de 250m, 375m e 625m.



(a) $\tau = 100$m (b) $\tau = 150$m (c) $\tau = 200$m

Figura 15: Estimação kernel normal aplicado aos dados de dengue:(a)raio de influência = 100m, (b)raio de influência = 150m, (c)raio de influência = 200m (Kawamoto,2012)

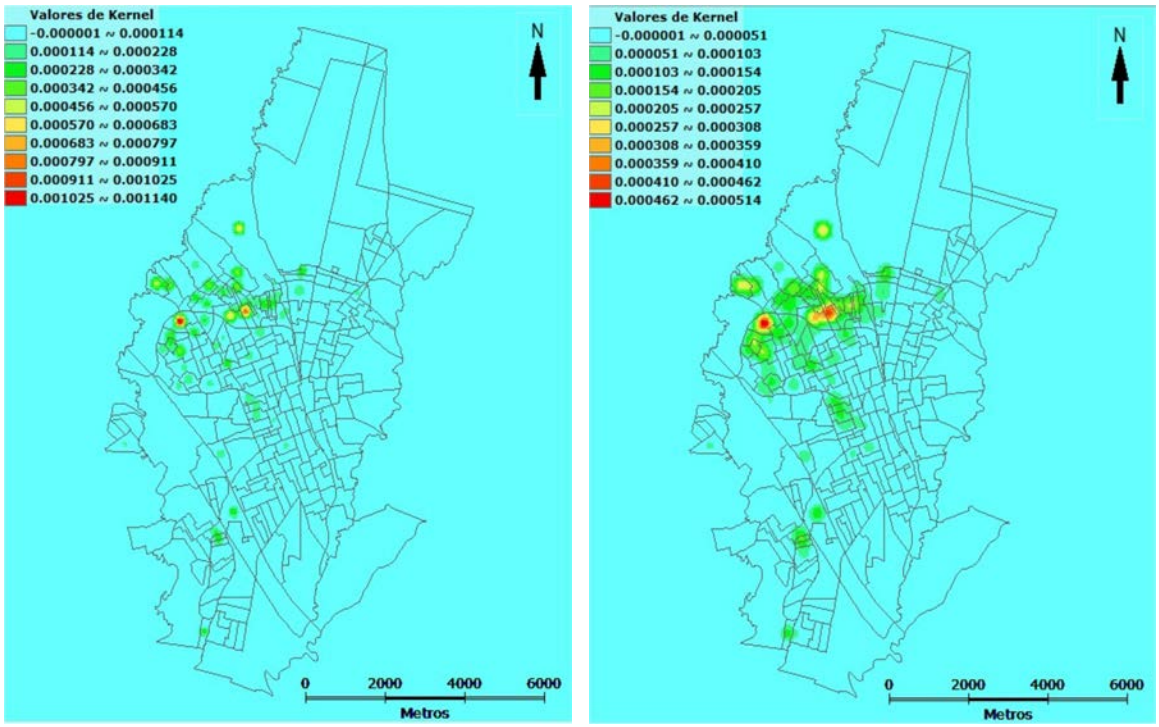


(a) $\tau = 250$m (b) $\tau = 375$m (c) $\tau = 625$m

Figura 16: Estimação kernel quártico aplicado aos dados de dengue:(a)raio de influência = 250m, (b)raio de influência = 375m, (c)raio de influência = 625m (Kawamoto,2012)

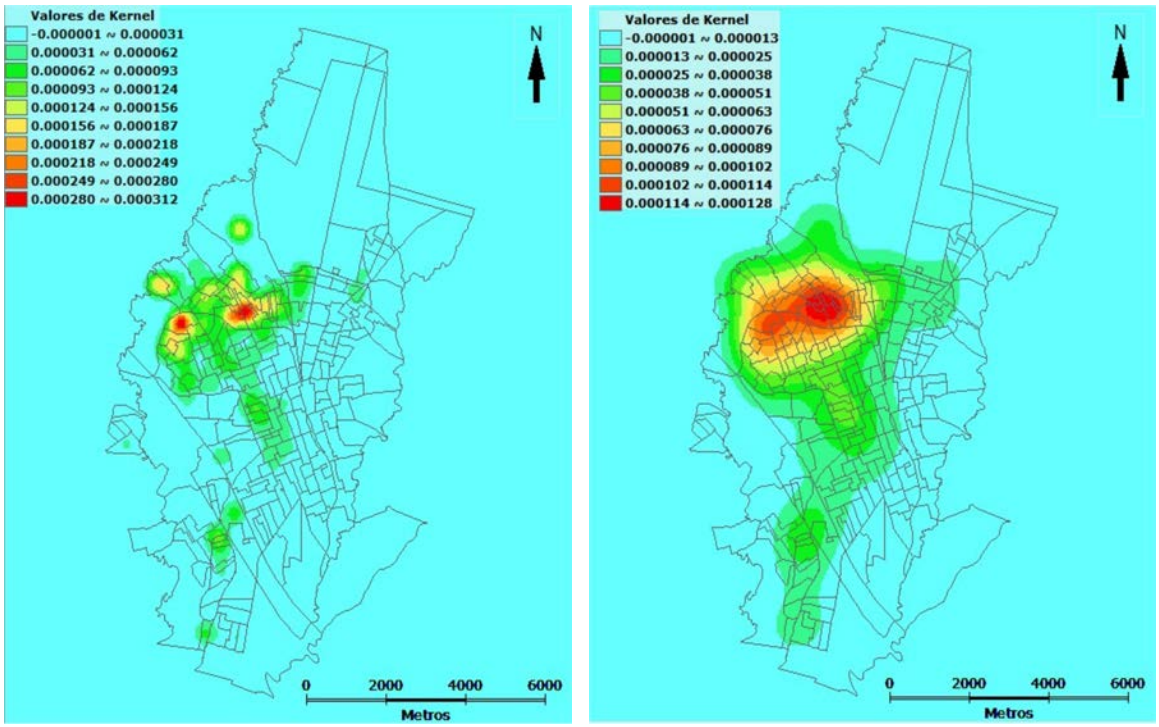
Comparando as figuras 11 e 15, observa-se uma semelhança entre os resultados obtidos por Kawamoto(2012) e obtidos no presente trabalho. As diferenças devem-se ao fato que Kawamoto(2012) trabalhou com processos pontuais e quando aplica-se a estimação de Kernel para dados por ponto, o ponto da região de estudo a ser estimado é genérico enquanto que, para dados por área o ponto a ser estimado é o centroide de cada bairro.

As figuras 17 e 18 apresentam os resultados da aplicação do Estimador Kernel para a função de estimação kernel normal para os respectivos raios de influência: 100m, 150m, 200m e 500m, para a escala setor censitário.



(a) $\tau = 100$m (b) $\tau = 150$m

Figura 17: Estimação kernel normal aplicado aos dados de dengue: (a)raio de influência = 100m, (b)raio de influência = 150m

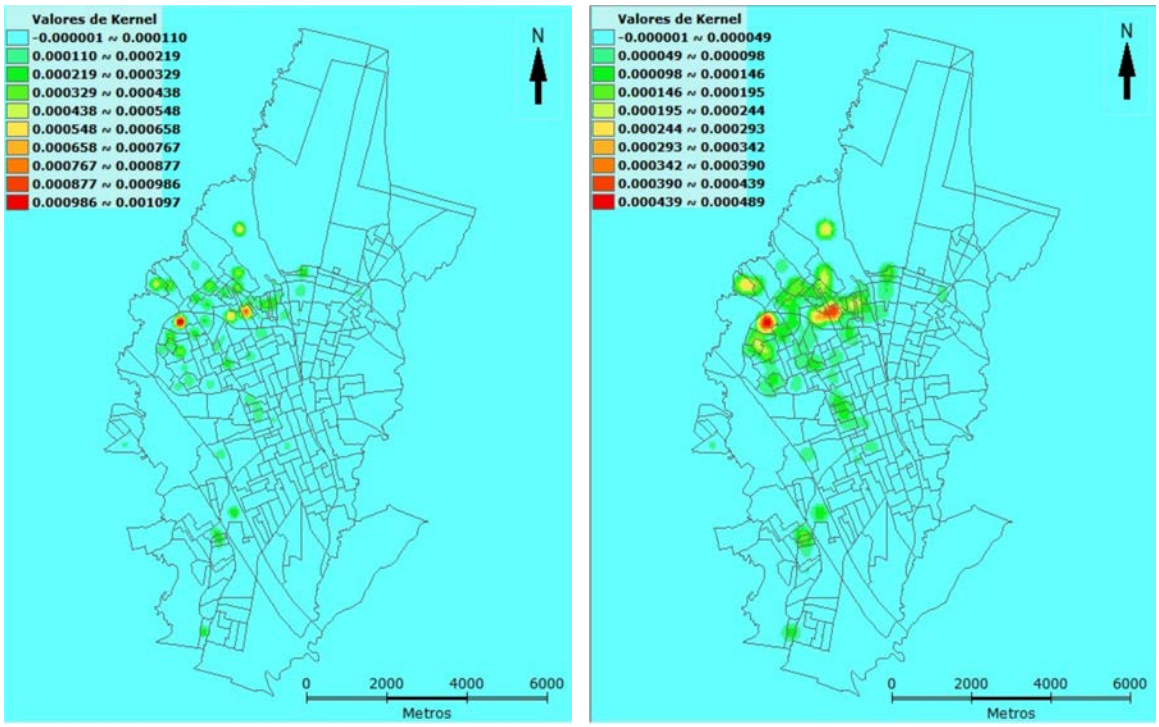


(a) $\tau = 200$m (b) $\tau = 500$ metros

Figura 18: Estimação kernel normal aplicado aos dados de dengue: (a)raio de influência = 200m, (b)raio de influência = 500m

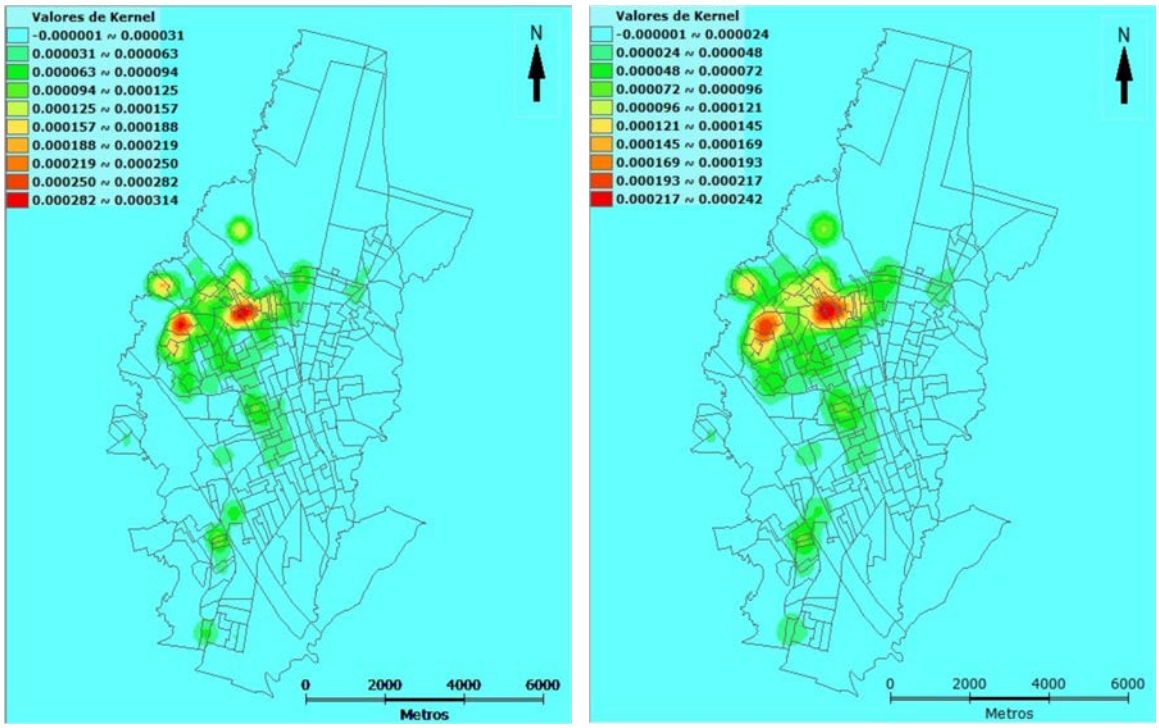
Observa-se as figuras 12 e 18, os resultados obtidos pela estimação kernel normal aplicados à diferentes escalas, há semelhança dos resultados conforme aumenta-se a medida do raio de influência.

As figuras 19 e 20 apresentam os resultados da aplicação do Estimador Kernel para a função densidade kernel quártico para os respectivos raios de influência: 250m, 375m, 500m e 625m, para a escala setor censitário.



(a) $\tau = 250$m (b) $\tau = 375$m

Figura 19: Estimação kernel quártico aplicado aos dados de dengue: (a)raio de influência = 250m, (b)raio de influência = 375m



(a) $\tau = 500$m (b) $\tau = 625$ metros

Figura 20: Estimação kernel quártico aplicado aos dados de dengue: (a)raio de influência = 500m, (b)raio de influência = 625m

Comparando as figuras 14 e 20, os resultados obtidos pela estimação kernel quártico aplicados à diferentes escalas, também assemelha-se para raios de de influência maiores. Cabe ao pesquisador decidir qual escala e raio melhor para o seu problema.

5 CONCLUSÕES

A técnica de ajuste a dados espacialmente distribuídos em área pode ser usada, com sucesso, para mapear ocorrências de dados espacialmente distribuídos, onde a região de ocorrência da variável é subdividida em áreas, por questões físicas, geográficas, econômicas, etc., e o interesse do pesquisador é verificar o comportamento da variável nas diferentes áreas.

Nota-se semelhança entre os resultados obtidos por processos pontuais e uma região geográfica particionada em áreas. Dependendo dos objetivos da pesquisa e da partição geográfica de uma área, o pesquisador deve optar por uma das técnicas a qual melhor refletirá os resultados.

REFERÊNCIAS BIBLIOGRÁFICAS

ASSUNÇÃO, R. M., Estatística Espacial com aplicações em Epidemiologia, Economia,Sociologia, 7^a Escola de Modelos de Regressâo, UFSCar, 14 a 18/02/2001, São Carlos, SP., 131 pags.

ATKINSON, P. M.; TATE, N. J., Spatial scale problems and geostatistical solutions: a review. *Professional Geographer*, **52**: 607 - 623, 2000.

BAILEY, T. C.; GATRELL, A. C., *Interactive Spatial Data Analysis*, London: Longman, 1995. 413pags.

BRASIL. MINISTERIO DA SAÚDE.*Introdução a Estatística Espacial para a Saúde Pública*. Brasília: Ministério da Saúde. 2007

CÂMARA, G.; CARVALHO, M. S., Análise Espacial de Eventos. In: *Análise Espacial de Dados Geográficos*, EMBRAPA, 2004.

CÂMARA, G.; MONTEIRO, A. M.; FUCKS, S. D.; CARVALHO, M. S.,*Análise Espacial e Geoprocessamento. In: Análise Espacial de Dados Geográficos*, EMBRAPA, 2004.

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

GVSIG, *http://gvsig.org*

KAWAMOTO, M.T. *Análise de Técnicas de Distribuição Espacial com Padrões Pontuais e Aplicação a Dados de Acidentes de Trânsito e a Dados de Dengue de Rio Claro- SP*. Dis. Mestrado- Biometria, IBB/UNESP. Botucatu-SP. 2012.

LLOYD, C. D.,*Local Models for Spatial Analysis*, 2ed,CRC Press, 2011.

MARSHALL, R., Mapping disease and mortality rates using empirical Bayes estimators.*Applied Statistics*, 40, 1991, 283 - 294

MiKTeX, *http://miktex.org*

OPENSHAW, O. *The Modifiable Areal Unit Problem*. Geobooks, Norwich, 1984. Concepts and techniques in Modern Geography 38.

OPENSHAW, O.; TAYLOR, P. J. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In N. Wrigley, editor, *Statistical Applications in the Spatial Sciences*, pages 127-144. Pion, London, 1979.

RIPLEY, B.D.,*Spatial Statistics*, London: Wiley, 2004, 252pags.

USHIZIMA, T.M. *Mapeamento da dengue na área urbana de Rio Claro(SP), no período de 2001-2003, e sua relação com condicionantes sócio-econômicas*. Dis. Mestrado - Geociências, IGCE/UNESP. Rio Claro - SP. 2005.

Terraview 4.2.2 São José dos Campos: INPE, 2010. Disponível em: www.dpi.inpe.br/terraview. Acesso em: 11/02/2014.